Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Collaborative User Clustering for Short Text Streams

Shangsong Liang[†], Zhaochun Ren[†], Emine Yilmaz[†], and Evangelos Kanoulas[‡]
[†]University College London, London, United Kingdom
[‡]University of Amsterdam, Amsterdam, The Netherlands
{shangsong.liang, zhaochun.ren, emine.yilmaz}@ucl.ac.uk, e.kanoulas@uva.nl

Abstract

In this paper, we study the problem of user clustering in the context of their published short text streams. Clustering users by short text streams is more challenging than in the case of long documents associated with them as it is difficult to track users' dynamic interests in streaming sparse data. To obtain better user clustering performance, we propose a user collaborative interest tracking model (UCIT) that aims at tracking changes of each user's dynamic topic distributions in collaboration with their followees', based both on the content of current short texts and the previously estimated distributions. We evaluate our proposed method via a benchmark dataset consisting of Twitter users and their tweets. Experimental results validate the effectiveness of our proposed UCIT model that integrates both users' and their collaborative interests for user clustering by short text streams.

Introduction

Popular microblogging platforms provide a light-weight, easy form of communication that enables users to broadcast and share information about their recent activities, opinions and status via short texts (Kwak et al. 2010). A good understanding and clustering of users' dynamic interests underlying their posts are critical for further design of applications that cater for users of such platforms, such as time-aware user recommendation (Arru, Gurini, and Gasparetti 2013) and personalized microblog search (Vosecky, Leung, and Ng 2014). In this paper, we study the problem of *collaborative user clustering in the context of short text streams*. Our goal is to infer users' and their collaborative topic distributions over time and dynamically cluster users that share interests in streams.

Most previous work (Chen et al. 2015; Xie et al. 2015) on user clustering uses collections of static, long documents, and hence makes the assumption that users' interests do not change over time. Recent work (Zhao et al. 2016) clusters users in the context of short documents streams, however it ignores any collaborative information, such as friends' messages. Our hypothesis is that accounting for this information is critical, especially for those users with limited activity, infrequent posts, and thus sparse information. In this work we dynamically cluster users in the context of short documents, by also utilizing each user's collaborative information, i.e. their friends' posts, from which we can infer users' collaborative interests for further improvement of the clustering.

Specifically, we propose a **U**ser **C**ollaborative **I**nterest **T**racking topic model, abbreviated as **UCIT**, for our collaborative user clustering. Our UCIT topic model is a dynamic multinomial Dirichlet mixture topic model that can infer and track each user's dynamic interests based not only on the user's posts but also his followees' posts for user clustering. Traditional topic models such as latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) and author topic model (Rosen-Zvi et al. 2004a) have been widely used to uncover topics of documents and users. These topic models ignore collaborative information, do not work well as they assume documents are long texts, or can not be directly applied in the context of short text streams as they assume the documents are in static collections.

In our UCIT topic model, to alleviate the sparsity problem in short texts, and by following previous work (Yan et al. 2013; 2015), we extract word-pairs in each short text, and form a word-pair set for each user to explicitly capture word co-occurrence patterns for the inference of users' topic distributions. To track users' dynamic interests, UCIT assumes that users' interests change over time and can be inferred by integrating the interests at previous time periods with newly observed data in the streams. To enhance the performance of dynamic user clustering in streams, UCIT infers not only a user's but also his followees' interests from the his own posts and also his followees' posts.

The contributions of the paper are threefold: (1) We propose a topic model that can collaboratively and dynamically track each user's and his followees' interests. (2) We propose a collapsed Gibbs sampling for the inference of our UCIT topic model. (3) We provide a thorough analysis of UCIT and of the impact of its key ingredients in user clustering, and demonstrate its effectiveness compared to the state-of-the-art algorithms.

Related Work

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 the Dynamic Topic Model (DTM) (Blei and Lafferty 2006), Dynamic Mixture Model (DMM) (Wei, Sun, and Wang 2007), Topic over Time (ToT) (Wang and McCallum 2006), Topic Tracking Model (TTM) (Iwata et al. 2009), and more recently, Generalized Linear Dynamic topic model (Caballero and Akella 2015), the dynamic User Clustering Topic model (UCT) (Zhao et al. 2016), Interaction Topic Model (Hua et al. 2016), Dynamic Clustering Topic model (DCT) (Liang, Yilmaz, and Kanoulas 2016) and scaling-up dynamic model (Bhadury et al. 2016). All of these models except DCT aim at inferring documents' dynamic topic distributions rather than user clustering. Except UCT and DCT that work in the context of short text streams, most of the the previous dynamic topic models works in the context of long text streams. To the best of our knowledge, none of existing dynamic topic models has considered the problem of clustering users with collaborative information, e.g., followees' interests, in the context of short text streams.

Problem Formulation

The problem we address is to track users' dynamic interests and cluster them over time in the context of short text streams such that users in the same cluster at a specific point in time share similar interests. The dynamic user clustering algorithm is essentially a function g that satisfies:

$$\mathbf{u}_t = \{u_1, u_2, \ldots, u_{|\mathbf{u}_t|}\} \xrightarrow{g} \mathbf{C}_t = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_Z\},$$

where $\mathbf{u}_t$ 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, while $\mathbf{C}_t$ is the resulting set of clusters of users with $\mathbf{c}_z$ being the z-th cluster in $\mathbf{C}_t$ and Z being the total number of clusters. We let $\mathbf{D}_t = \{\ldots, \mathbf{d}_{t-2}, \mathbf{d}_{t-1}, \mathbf{d}_t\}$ denote the *stream* of documents generated by users in $\mathbf{u}_t$ up to time t with $\mathbf{d}_t$ being the most recent set of short documents arriving at time period t. We assume that the length of a document d in $\mathbf{D}_t$ is no more than a predefined small length (for instance, 140 characters in the case of Twitter).

Method

In this section, we describe our proposed User Collaborative Interest Tracking topic model, **UCIT**.

Overview

We use Twitter as our default setting of short text streams and provide an overview of our proposed UCIT model in Algorithm 1. Following (Liang, Ren, and de Rijke 2014; Zhao et al. 2016; Liang et al. 2016), we represent each user's interests by topics. Thus, the interests of each user $u \in \mathbf{u}_t$ at time period t are represented as a multinomial distribution $\boldsymbol{\theta}_{t,u} = \{\theta_{t,u,z}\}_{z=1}^{Z}$ over topics. Here Z is the total number of topics. The distribution $\boldsymbol{\theta}_{t,u}$ is inferred by the UCIT model. To alleviate the sparsity problem of short texts, and by following recent work on the topic (Yan et al. 2013; 2015), we construct and represent documents by their biterms, i.e. word pairs in them (step 1 in Algorithm 1). Next, we propose a dynamic Dirichlet multinomial mixture user collaborative interest tracking topic model to capture each user's dynamic interests $\boldsymbol{\theta}_{t,u} = \{\theta_{t,u,z}\}_{z=1}^{Z}$ and their collaborative interests $\boldsymbol{\psi}_{t,u} = \{\psi_{t,u,z}\}_{z=1}^{Z}$ inferred from their followees $\mathbf{f}_{t,u}$, at time t, in the context of short text streams (step 2 in Algorithm 1). Here $\mathbf{f}_{t,u}$ is user u's all followees at t.Based on each user's multinomial distributions $\boldsymbol{\theta}_{t,u}$ and $\boldsymbol{\psi}_{t,u}$, we cluster users using K-means clustering (Jain 2010) (step 3 in Algorithm 1). With the time period t moving forward, the clustering result changes dynamically.

Algorithm 1: Overview of the proposed UCIT model.

```
Input  : A set of users $\mathbf{u}_t$ along with their tweets $\mathbf{D}_t$
Output: Clusters of users $\mathbf{C}_t$
1 Construct a collection of word-pairs $\mathbf{b}_{t,u}$ for each user $u$
2 Use UCIT model to track each user's interests as $\boldsymbol{\theta}_{t,u}$ and their collaborative interest as $\boldsymbol{\psi}_{t,u}$
3 Cluster users based on each user's interest $\boldsymbol{\theta}_{t,u}$ and their collaborative interest $\boldsymbol{\psi}_{t,u}$
```

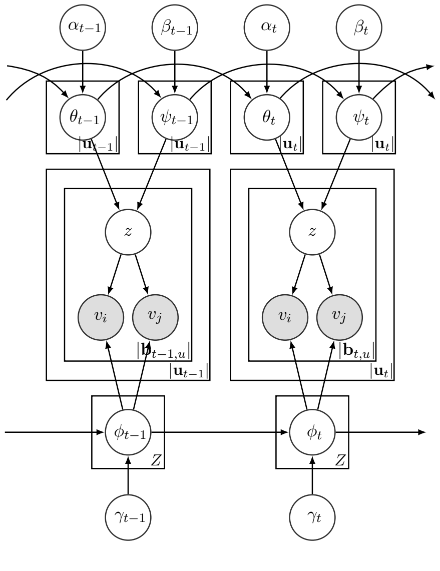


Figure 1: Graphical representation of our user interest tracking clustering topic model, UCIT. Shaded nodes represent observed variables.

User Collaborative Interest Tracking Model

Modeling Interests over Time. The goal of UCIT topic model is to infer the dynamical topic distribution of each user, $\boldsymbol{\theta}_{t,u} = \{\theta_{t,u,z}\}_{z=1}^{Z}$, and the user's collaborative topic distribution, $\boldsymbol{\psi}_{t,u} = \{\psi_{t,u,z}\}_{z=1}^{Z}$, in short text streams at a given time t, and dynamically cluster all users based on information of each user's $\boldsymbol{\theta}_{t,u}$ and $\boldsymbol{\psi}_{t,u}$ over time. Fig. 1 shows a graphical representation of our UCIT model.

Given a user u, to track the dynamics of their interests, we make the assumption that the mean of the user's current interests at time period t is the same as that at the previous time period $t-1$, unless otherwise newly arrived documents at the current time period are observed. In particular, following the work of past 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 $\boldsymbol{\theta}_{t-1,u}$:

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

where the precision value $\alpha_{t,z}$ represents users' topic persistency, that is how saliency topic z is at time t compared to that at time $t-1$ for the users. The distribution is a conjugate prior of the multinomial distribution, hence the inference can be performed by Gibbs sampling (Liu 1994). Similarly, to track the dynamic changes of a user u's collaborative interests, we assume a Dirichlet prior, in which the mean of the current distribution $\boldsymbol{\psi}_{t,u}$ evolves from the mean of the previous distribution $\boldsymbol{\psi}_{t-1,u}$ with a set of precision values $\boldsymbol{\beta}_t = \{\beta_{t,z}\}_{z=1}^{Z}$:

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

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 $\boldsymbol{\phi}_{t-1,z}$:

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

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 $\gamma_{t,v}$ representing the persistency of the words in topics at time t, a measure of how consistently the words belong to the topics at time t compared to that at the previous time $t-1$. We describe the inference for all users' and their collaborative distributions $\boldsymbol{\Theta}_t = \{\boldsymbol{\theta}_{t,u}\}_{u=1}^{|\mathbf{u}_t|}$ and $\boldsymbol{\Psi}_t = \{\boldsymbol{\psi}_{t,u}\}_{u=1}^{|\mathbf{u}_t|}$, the words' dynamic topic distribution $\boldsymbol{\Phi}_t = \{\boldsymbol{\phi}_{t,z}\}_{z=1}^{Z}$ and the update rules of the persistency values $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$ later in the section.

Assuming that we know all users' topic distribution at time $t-1$, $\boldsymbol{\Theta}_{t-1}$, their collaborative topic distribution at time $t-1$, $\boldsymbol{\Psi}_{t-1}$, and the words' topic distribution, $\boldsymbol{\Phi}_{t-1}$, the proposed user interest tracking model is a generative topic model that depends on $\boldsymbol{\Theta}_{t-1}$, $\boldsymbol{\Psi}_{t-1}$ and $\boldsymbol{\Phi}_{t-1}$. For initialization, we let $\theta_{0,u,z} = 1/Z$, $\psi_{0,u,z} = 1/Z$ and $\phi_{0,z,v} = 1/V$. The generative process (used by the Gibbs sampler for parameter estimation) of our model for documents in stream at time t, is as follows,

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

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 $\boldsymbol{\alpha}_t\boldsymbol{\theta_{t-1,u}}$ and $\boldsymbol{\beta}_t\boldsymbol{\psi}_{t-1,u}$, respectively; then for each biterm $b \in \mathbf{b}_{t,u}$:

   (a) Draw a topic $z_{t,u,b}$ based on multinomials $\boldsymbol{\theta}_{t,u}$ and $\boldsymbol{\psi}_{t,u}$;

   (b) Draw a word $w_i \in b$ from multinomial $\boldsymbol{\phi}_{t,z_{t,u,b}}$;

   (c) Draw another word $w_j \in b$ from multinomial $\boldsymbol{\phi}_{t,z_{t,u,b}}$.

Algorithm 2: Inference for the UCIT model at time t.

```
Input : Distributions $\boldsymbol{\Theta}_{t-1}$, $\boldsymbol{\Psi}_{t-1}$ and $\boldsymbol{\Phi}_{t-1}$ at $t-1$;
        Initialized $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$, $\boldsymbol{\gamma}_t$; 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 each biterm $b = (v_i, v_j) \in \mathbf{b}_{t,u}$ do
            draw $z_{t,u,b}$ from the conditional probability,
            i.e., (5)
            update $m_{t,u,z_{t,u,b}}$, $\{o_{t,u',z_{t,u,b}}\}_{u' \in \mathbf{f}_{t,u}}$,
            $n_{t,z_{t,u,b},v_i}$ and $n_{t,z_{t,u,b},v_j}$
    update $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$
Compute the posterior estimates $\boldsymbol{\Theta}_t$, $\boldsymbol{\Psi}_t$ and $\boldsymbol{\Phi}_t$.
```

Fig. 1 illustrates the graphical representation of our model, where shaded and unshaded nodes indicate observed and latent variables, respectively, and a dependency of two multinomials is assumed to exist between two adjacent time periods.

Inference. We employ a collapsed Gibbs sampler (Griffiths and Steyvers 2004) for an approximate inference of the distribution parameters of our model. As can be seen 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 $\boldsymbol{\phi}_t$. In this way, we enable sampling since we do not need to sample these multinomials.

Algorithm 2 shows an overview of our proposed collapsed Gibbs sampling algorithm for the inference, where $m_{t,u,z}$ and $n_{t,z,v}$ are the number of biterms assigned to topic z and the number of times word v is assigned to topic z for user u at time t, respectively; $o_{t,u',z}$ is the number of biterms assigned to topic z for user u' who is one of user u' followees.

In the Gibbs sampling procedure we need to calculate the conditional distribution $P(z_{t,u,b} \mid \mathbf{z}_{t,-b}$, $\mathbf{d}_t$, $\boldsymbol{\Theta}_{t-1}$, $\boldsymbol{\Psi}_{t-1}$, $\boldsymbol{\Phi}_{t-1}$, $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$, $\boldsymbol{\gamma}_t)$, at time t, where $\mathbf{z}_{t,-b}$ represents the topic assignments for all biterms in $\mathbf{d}_t$ except biterm b. We begin with the joint probability of the current document set, $P(\mathbf{z}_t$, $\mathbf{d}_t$ $|\boldsymbol{\Theta}_{t-1}$, $\boldsymbol{\Psi}_{t-1}$, $\boldsymbol{\Phi}_{t-1}$, $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$, $\boldsymbol{\gamma}_t)$:

$$\begin{aligned} & P(\mathbf{z}_t, \mathbf{d}_t|\boldsymbol{\Theta}_{t-1}, \boldsymbol{\Psi}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t) \qquad (4) \\ & = (1-\lambda)P(\mathbf{z}_t, \mathbf{d}_t|\boldsymbol{\Theta}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t) + \\ & \quad \lambda P(\mathbf{z}_t, \mathbf{d}_t|\boldsymbol{\Psi}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t) \\ & = (1-\lambda)\left(\prod_z \left(\frac{\mathit{\Gamma}(\sum_v(\varkappa_b))}{\prod_v \mathit{\Gamma}(\varkappa_b)} \frac{\prod_v \mathit{\Gamma}(\varkappa_a)}{\mathit{\Gamma}(\sum_v \varkappa_a)}\right)\right)^2 \times \end{aligned}$$

$$\prod_u \frac{\varGamma(\sum_z(\varkappa_2))}{\prod_z \varGamma(\varkappa_2)} \frac{\prod_z \varGamma(\varkappa_1)}{\varGamma(\sum_z \varkappa_1)} +$$
$$\lambda \left(\prod_z \left(\frac{\varGamma(\sum_v(\varkappa_d))}{\prod_v \varGamma(\varkappa_d)} \frac{\prod_v \varGamma(\varkappa_c)}{\varGamma(\sum_v \varkappa_c)} \right) \right)^2 \times$$
$$\prod_u \frac{\varGamma(\sum_z(\varkappa_4))}{\prod_z \varGamma(\varkappa_4)} \frac{\prod_z \varGamma(\varkappa_3)}{\varGamma(\sum_z \varkappa_3)},$$

where $\varGamma(\cdot)$ is a gamma function, λ is a free parameter that governs the linear mixture of a user's own interests and their followees' interests, and parameters $\varkappa$ are defined as the following:

$$\begin{aligned}
\varkappa_1 &= m_{t,u,z} + \alpha_{t,z}\theta_{t-1,u,z} - 1, & \varkappa_2 &= \alpha_{t,z}\theta_{t-1,u,z},\\
\varkappa_3 &= o_{t,u,z} + \beta_{t,z}\psi_{t-1,u,z} - 1, & \varkappa_4 &= \beta_{t,z}\psi_{t-1,u,z},\\
\varkappa_a &= n_{t,z,v} + \gamma_{t,v}\phi_{t-1,z,v} - 1, & \varkappa_b &= \gamma_{t,v}\phi_{t-1,z,v},\\
\varkappa_c &= o_{t,z,v} + \gamma_{t,v}\phi_{t-1,z,v} - 1, & \varkappa_d &= \gamma_{t,v}\phi_{t-1,z,v}.
\end{aligned}$$

Based on the above joint probability and using the chain rule, we can obtain the following conditional probability conveniently:

$$\begin{aligned}
&P(z_{t,u,b} = z | \mathbf{z}_{t,-b}, \mathbf{d}_t, \mathbf{\Theta}_{t-1}, \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t)\\
&\propto (1-\lambda) \frac{m_{t,u,z} + \alpha_{t,z}\theta_{t-1,u,z} - 1}{\sum_{z'=1}^{Z} m_{t,u,z'} + \alpha_{t,z'}\theta_{t-1,u,z'} - 1} \times \qquad (5)\\
&\quad \prod_{v \in b} \frac{n_{t,z,v} + \gamma_{t,v}\phi_{t-1,z,v} - 1}{\sum_{v'=1}^{V}(n_{t,z,v'} + \gamma_{t,v'}\phi_{t-1,z,v'}) - 1} +\\
&\quad \lambda \frac{o_{t,u,z} + \beta_{t,z}\psi_{t-1,u,z} - 1}{\sum_{z'=1}^{Z} o_{t,u,z'} + \beta_{t,z'}\psi_{t-1,u,z'} - 1} \times\\
&\quad \prod_{v \in b} \frac{o_{t,z,v} + \gamma_{t,v}\phi_{t-1,z,v} - 1}{\sum_{v'=1}^{V}(o_{t,z,v'} + \gamma_{t,v'}\phi_{t-1,z,v'}) - 1},
\end{aligned}$$

for the proposed Gibbs sampling (step 5 in Algorithm 2). At each iteration during the sampling, we estimate the precision parameters $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$ by maximizing the joint distribution $P(\mathbf{z}_t,\, \mathbf{d}_t\, |\mathbf{\Theta}_{t-1},\, \mathbf{\Psi}_{t-1},\, \mathbf{\Phi}_{t-1},\, \boldsymbol{\alpha}_t,\, \boldsymbol{\beta}_t,\, \boldsymbol{\gamma}_t)$. We apply fixed-point iterations to obtain the optimal $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$. The following update rule of $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$ for maximizing the joint distribution in our fixed-point iteration is derived by applying two bounds in (Minka 2000):

$$\begin{aligned}
\alpha_{t,z} &\leftarrow \frac{(1-\lambda)\alpha_{t,z} \sum_u (\varDelta(\varkappa_1) - \varDelta(\varkappa_2))}{\sum_u (\varDelta(\sum_z \varkappa_1) - \varDelta(\sum_z \varkappa_2))},\\
\beta_{t,z} &\leftarrow \frac{\lambda\beta_{t,z} \sum_u (\varDelta(\varkappa_3) - \varDelta(\varkappa_4))}{\sum_u (\varDelta(\sum_z \varkappa_3) - \varDelta(\sum_z \varkappa_4))}, \qquad (6)\\
\gamma_{t,v} &\leftarrow \frac{(1-\lambda)\gamma_{t,v} \sum_z (\varDelta(\varkappa_a) - \varDelta(\varkappa_b))}{\sum_z (\varDelta(\sum_v \varkappa_a) - \varDelta(\sum_v \varkappa_b))} +\\
&\qquad \frac{\lambda\gamma_{t,v} \sum_z (\varDelta(\varkappa_c) - \varDelta(\varkappa_d))}{\sum_z (\varDelta(\sum_v \varkappa_c) - \varDelta(\sum_v \varkappa_d))},
\end{aligned}$$

where $\varDelta(x) = \frac{\partial \log \varGamma(x)}{x}$ is a Digamma function.

Once the Gibbs sampling procedure has been done, with the fact that Dirichlet distribution is conjugate to multinomial distribution, we can conveniently infer each user's, their collaborative and the words' topic distributions, $\boldsymbol{\theta}_{t,u}$, $\boldsymbol{\psi}_{t,u}$, and $\boldsymbol{\phi}_{t,z}$, as follows, respectively:

$$\begin{aligned}
\theta_{t,u,z} &= \frac{m_{t,u,z} + \alpha_{t,z}\theta_{t,u,z}}{\sum_{z'=1}^{Z}(m_{t,u,z'} + \alpha_{t,z'}\theta_{t,u,z'}) - 1},\\
\psi_{t,u,z} &= \frac{o_{t,u,z} + \beta_{t,z}\psi_{t,u,z}}{\sum_{z'=1}^{Z}(o_{t,u,z'} + \beta_{t,z'}\psi_{t,u,z'}) - 1}, \qquad (7)\\
\phi_{t,z,v} &= \frac{n_{t,z,v} + \gamma_{t,z}\phi_{t,z,v}}{\sum_{v'=1}^{V}(n_{t,z,v'} + \gamma_{t,z'}\phi_{t,z,v'}) - 1}.
\end{aligned}$$

Clustering Users

After we obtain each user's and his collaborative topic distributions, $\boldsymbol{\theta}_{t,u}$ and $\boldsymbol{\psi}_{t,u}$ from (7), we use the following mixture distribution $\boldsymbol{\rho}_{t,u}$ to represent each user:

$$\boldsymbol{\rho}_{t,u} = (1-\lambda)\boldsymbol{\theta}_{t,u} + \lambda\boldsymbol{\psi}_{t,u}. \qquad (8)$$

Then, we can conveniently cluster users based on their interests $\boldsymbol{\rho}_{t,u}$ using the K-means algorithm (Jain 2010). Other traditional unsupervised clustering algorithms can be applied, but we found that the performance with other clustering algorithms is not significantly different from that with K-means. For previously unseen users, we can not directly utilize (7) for the clustering, as $\boldsymbol{\theta}_{t-1,u}$ and $\boldsymbol{\psi}_{t-1,u}$ are not defined at t. In this case, we use the distribution of topics for each biterm in the users' text according to the current assignment of topics to biterms.

Experiments and Results

In what follows, we detail our experimental setup, report and analyze the results.

Experimental Setup

Research Questions. The research questions that guide the remainder of the paper are:

RQ1 How does UCIT perform compared to state-of-the-art methods for user clustering?

RQ2 What is the impact of the length of the time intervals, $t_i - t_{i-1}$, in UCIT?

RQ3 What is the contribution of the collaborative information for user clustering?

RQ4 What is the quality of the topical representation inferred by UCIT?

RQ5 What is the generalization performance of UCIT compared to state-of-the-art topic models?

Dataset. In order to answer our research questions, we work with a dataset collected from Twitter (Zhao et al. 2016). The dataset contains 1,375 active users and their tweets spanning a time period that starts at each user's registration and ends on May 31, 2015. Most of the users are being followed by 2 to 50 followers. In total, there is 7.52 million tweets with timestamps including those from users' followees'. The average length of a tweet is 12 words. The dataset contains ground truth clusters for partitions of 5 different time intervals, a week (48 to 60 clusters), a month (43 to 52 clusters), a quarter (40 to 46 clusters), half a year (28 to 30 clusters) and a year (28 to 30 clusters).

Baselines. We compare our UCIT with the following baselines and state-of-the-art clustering algorithms:

- **K-means.** It represents users by TF-IDF vectors, and clusters them based on their cosine similarities.
- **GSDMM.** This model represents each short document through a single topic to alleviate sparsity (Yin and Wang 2014).
- **Latent Dirichlet Allocation (LDA).** This model infers topic distributions specific to each document via the LDA model.
- **Author Topic Model (AuthorT).** This model (Rosen-Zvi et al. 2004b) infers topic distributions specific to each user in a static dataset.
- **Dynamic topic model (DTM).** This model (Blei and Lafferty 2006) utilizes a Gaussian distribution for inferring topic distribution of long text documents in streams.
- **Topic over time model (ToT).** This model (Wang and McCallum 2006) normalizes timestamps of long documents in a collection and then infers topics distribution for each document.
- **Topic tracking model (TTM).** This model (Iwata et al. 2009) captures the dynamic topic distribution of long documents arriving at time t in streams based on the content of the documents and the previous estimated distributions.

For fair comparisons, the GSDMM, LDA, DTM, ToT and TTM baselines use both each user u's interests $\boldsymbol{\theta}_{t,u}$ and their collaborative interests for clustering. As these baselines can not directly infer collaborative interests, we use the average interests of the user's followees as the collaborative interests. Thus, we can use the mixture interests $\boldsymbol{\rho}_{t,u} = (1-\lambda)\boldsymbol{\theta}_{t,u} + \lambda\frac{1}{|\mathbf{f}_{t,u}|}\sum_{u'\in\mathbf{f}_{t,u}}\boldsymbol{\theta}_{t,u'}$ for each user in the user clustering, and then cluster users based on the similarities of their $\boldsymbol{\rho}_{t,u}$ distributions in these baselines. For static topic models, i.e., LDA and AuthorT, we set $\alpha = 0.1$ and $\beta = 0.01$. We set the number of topics $Z = 50$ and the number of clusters equal to the number of topics.

For further analysis of the contribution of collaborative interests $\boldsymbol{\psi}_{t,u}$ inferred by our model to the clustering, we use two additional baselines UCIT$_{\text{avg}}$ and UCIT$_{\text{avg+}\boldsymbol{\psi}}$, where $\boldsymbol{\rho}_{t,u}$ is set to be $(1-\lambda)\boldsymbol{\theta}_{t,u} + \lambda\frac{1}{|\mathbf{f}_{t,u}|}\sum_{u'\in\mathbf{f}_{t,u}}\boldsymbol{\theta}_{t,u'}$, and $(1-\lambda_1-\lambda_2)\boldsymbol{\theta}_{t,u} + \lambda_1\frac{1}{|\mathbf{f}_{t,u}|}\sum_{u'\in\mathbf{f}_{t,u}}\boldsymbol{\theta}_{t,u'} + \lambda_2\boldsymbol{\psi}_{t,u}$, respectively. Here $\boldsymbol{\theta}_{t,u}$ and $\boldsymbol{\psi}_{t,u}$ are generated by our proposed model. Note that we use UCIT$_{\boldsymbol{\psi}}$ to denote the model where $\boldsymbol{\rho}_{t,u} = (1-\lambda)\boldsymbol{\theta}_{t,u} + \lambda\boldsymbol{\psi}_{t,u}$. Note again that when $\lambda = 0$, both UCIT$_{\text{avg}}$ and UCIT$_{\boldsymbol{\psi}}$ will reduce to the state-of-the-art user clustering baseline, UCT (Zhao et al. 2016), where each user's friends' posts are not taken into account, and similarly, when both $\lambda_1 = 0$ and $\lambda_2 = 0$, UCIT$_{\text{avg+}\boldsymbol{\psi}}$ will reduce to UCT.

Evaluation Metrics. We use Precision, Purity, NMI (Normalized Mutual Information), and ARI (Adjusted Rank Index) to evaluate the performance of user clustering, all of which are widely used in the literature (Manning, Raghavan, and Schütze 2008). Higher Precision, Purity, NMI scores indicate better user clustering performance.

We further use H-score (Bordino et al. 2010) to evaluate the quality of topical representations of user clusters generated by UCIT and the baseline models. The intuition behind the H-score is that if the average inter-cluster distance is smaller compared to the average intra-cluster distance, the topical representation of the users in the clusters reaches better performance. A lower H-score indicates better topic representations of users in the output clusters.

In terms of evaluating the generalization performance of the model we adopt Perplexity. This metric, used by convention in many topic models (Blei, Ng, and Jordan 2003), is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower Perplexity score indicates better generalization performance.

Results and Analysis

In the following, we discuss and analyze our experimental results and answer the research questions **RQ1** to **RQ5**.

Effectiveness of UCIT. We begin by answering research question **RQ1**. Following (Gama et al. 2014), we split the dataset into two parts: half of the dataset for training, and the remaining for testing. Table 1 provides the evaluation performance of our UCIT model and the baseline models using time periods of a month in terms of clustering metrics, Precision, Purity, ARI and NMI, respectively.

We have the following findings from Table 1: (a) All the three versions of UCIT model, UCIT$_{\text{avg}}$, UCIT$_{\text{avg+}\boldsymbol{\psi}}$ and UCIT$_{\boldsymbol{\psi}}$, can statistically significantly outperform the baselines in terms of all the metrics, which demonstrates the effectiveness of our way of inferring users' interests and their collaborative interests for user clustering. (b) Both UCIT$_{\boldsymbol{\psi}}$ and UCIT$_{\text{avg+}\boldsymbol{\psi}}$ outperform UCIT$_{\text{avg}}$, which demonstrates that utilizing the inferred collaborative interests $\boldsymbol{\psi}$ can yield better performance compared to simply utilizing the average of followees' interests as collaborative information. (c) UCIT$_{\boldsymbol{\psi}}$ works better than UCIT$_{\text{avg+}\boldsymbol{\psi}}$, which demonstrates that the contribution of $\boldsymbol{\psi}$ is more critical for user clustering compared to that of the average of the interests for user clustering. The reason UCIT$_{\boldsymbol{\psi}}$ works better than UCIT$_{\text{avg+}\boldsymbol{\psi}}$ is, again, that using average interests as collaborative interests from followees is less effective than that explicitly inferred in the model.

Impact of Time Interval Length. We now turn to answer research question **RQ2**. To understand the influence on UCIT of the length of the time period used for evaluation, in Fig. 2 we compare the performance for different time periods: a week, a month, a quarter, half a year and a year, respectively.

According to Fig. 2, all the UCIT models, UCIT$_{\text{avg}}$, UCIT$_{\text{avg+}\boldsymbol{\psi}}$ and UCIT$_{\boldsymbol{\psi}}$, outperforms the baselines for time intervals of all lengths. This finding, again, confirms the fact that UCIT works better than the state-of-the-art algorithms for user clustering in short text streams regardless of interval length. When the interval length increases from a week to a month, the performance of the UCIT models and the

Table 1: Clustering performance of UCIT and the baselines using a time period of a month. Statistically significant differences between UCIT$_\psi$ and UCIT$_{\text{avg}+\psi}$, between UCIT$_\psi$ and UCIT$_{\text{avg}}$ are marked in the upper and lower right hand corner of UCIT$_\psi$'s score, respectively. The statistical significance is tested using a two-tailed paired t-test and is denoted using $^\blacktriangle$ for $\alpha = .01$, and $^\vartriangle$ for $\alpha = .05$.

	Precision	Purity	ARI	NMI
K-Means	.265	.512	.397	.414
LDA	.305	.551	.473	.464
AuthorT	.322	.571	.487	.488
DTM	.336	.579	.499	.473
TTM	.344	.587	.522	.521
ToT	.359	.605	.552	.582
GSDMM	.398	.632	.592	.561
UCIT$_{\text{avg}}$	.505	.714	.718	.818
UCIT$_{\text{avg}+\psi}$	.560	.736	.762	.861
UCIT$_\psi$	$.583^{\blacktriangle}_{\blacktriangle}$	$.746^{\vartriangle}_{\blacktriangle}$	$.776^{\vartriangle}_{\blacktriangle}$	$.883^{\blacktriangle}_{\blacktriangle}$

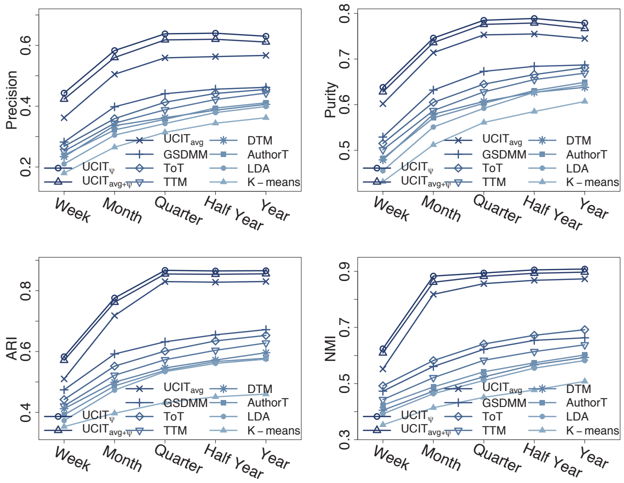


Figure 2: Precision, Purity, ARI and NMI Performance of our models UCIT and the baselines on time periods of a week, a month, a quarter, half a year, and a year, respectively.

baseline models improves significantly on all metrics, while performance reaches a plateau as the time intervals further increase. In all cases the UCIT models significantly outperform the baseline models. These findings demonstrate that the performance of UCIT is robust and is able to maintain significant improvements over the state-of-the-art.

Contribution of the Collaborative Interests. Next, we turn to answer research question **RQ3** to further analyze the contribution of the main ingredient, the collaborative information ψ inferred in our UCIT model. We vary λ and show the performance of our models, UCIT$_\psi$ and UCIT$_{\text{avg}}$, and the best baseline model, GSDMM in Fig. 3. The rest of the baselines yield similar or worse performance than GSDMM and they are not reported here. Also, we do not report the performance of UCIT$_{\text{avg}+\psi}$, as it obtains quantitively similar to UCIT$_{\text{avg}}$ performance. As λ increases from 0 to 0.6, giving more weight to the collaborative information in UCIT models and the average of followees' interests in GSDMM, respectively, the performance of all models improves, with UCIT$_\psi$ outperforming UCIT$_{\text{avg}}$ and GSDMM. This, again, confirms the fact that integrating collaborative interests into the model does make contribution to the improvement, and our models work better than the best baseline. Fig. 3 also shows that UCIT$_\psi$ that uses collaborative interests for clustering outperforms UCIT$_{\text{avg}}$ that simply uses the average of the followees' interests as collaborative interests, which again, demonstrates that the inferred collaborative interests in UCIT does help to further improve the performance compared to the average of the followees' interests. When $\lambda = 0$, both UCIT$_\psi$ and UCIT$_{\text{avg}}$ reduce to the state-of-the-art baseline model, UCT, that does not infer and utilize collaborative information for user clustering. It is clear from Fig. 3 that both UCIT$_\psi$ and UCIT$_{\text{avg}}$ outperform UCT.

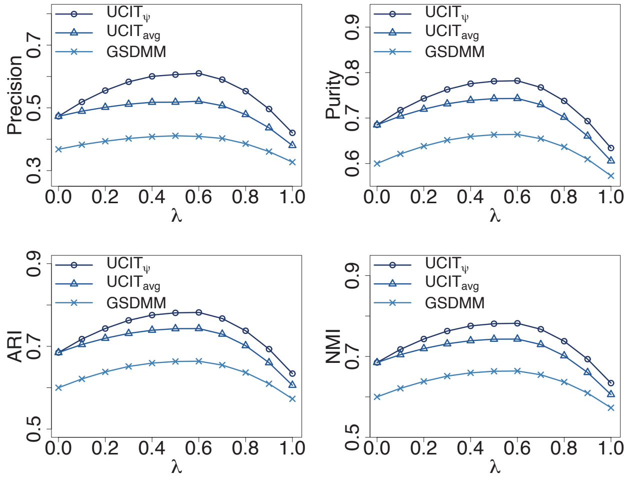
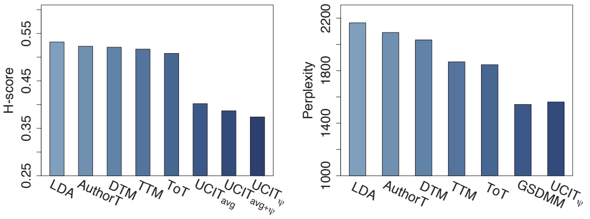


Figure 3: Precision, Purity, ARI and NMI Performance of our models UCIT and GSDMM on varying scores of λ, respectively.

Quality of Topic Representation and Perplexity Performance. Finally, we turn to research questions **RQ4** and **RQ5**. In order to answer **RQ4** and analyze the topical representation ability of UCIT and the baseline models, we use H-score for evaluation. A smaller H-score indicates that the topical representation of users is more similar to the manually labeled one. It is clear from Fig. 4(a) that the UCIT models outperform all other baselines. Note that the H-score cannot be computed for GSDMM, as it assigns one single topic to each short document and each user.

In order to answer **RQ5** and understand the generalization performance of UCIT and the baseline models, we use perplexity for the evaluation. Fig. 4(b) shows the result. A lower perplexity score indicates better generalization performance. As it can be observed, UCIT$_\psi$ performs better than all the baseline models except GSDMM. Note that the perplexity performance of UCIT$_{\text{avg}}$ and UCIT$_{\text{avg}+\psi}$ is the same as that of UCIT$_\psi$, and thus not reported in the figure.



(a) H-score (b) Perplexity

Figure 4: (a) Quality of topic representations evaluated by H-score and (b) generalization performance evaluated by Perplexity, for UCIT and the baselines using time periods of a quarter, respectively.

Conclusion

In this paper we studied the problem of dynamically clustering users in the context of short text streams. We have proposed a user collaborative interest tracking topic model (UCIT) that can infer and track each user and their followees' dynamic interests for user clustering. Our UCIT can effectively handle both the textual sparsity of short documents, and the dynamic nature of users' and their followees' interests over time. We evaluated the performance of UCIT in terms of clustering, topical representation and generalization effectiveness, and make comparisons with state-of-the-art models. Our experimental results demonstrated that UCIT can effectively cluster users in short text streams. As future work, we intent to incorporate other information such as the users' social network for user clustering. Like most previous work, it is challenging to obtain the ground-truth number of user clusters in our model. Thus, we leave this as future work. We also plan to consider other collaborative strategies for user clustering in streams.

Acknowledgments. We thank the anonymous reviewers for their helpful comments on this paper. This research was partially supported by the UCL Big Data Institute and the Elsevier company.

References

Arru, G.; Gurini, D. F.; and Gasparetti, F. 2013. Signal-based user recommendation on twitter. In *WWW*, 941–944.

Bhadury, A.; Chen, J.; Zhu, J.; and Liu, S. 2016. Scaling up dynamic topic models. In *WWW*, 381–390.

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.

Bordino, I.; Castillo, C.; Donato, D.; and Gionis, A. 2010. Query similarity by projecting the query-flow graph. In *SIGIR*, 515–522.

Caballero, K., and Akella, R. 2015. Dynamically modeling patients health state from electronic medical records: A time series approach. In *KDD*, 69–78.

Chen, W.; Wang, J.; Zhang, Y.; Yan, H.; and Li, X. 2015. User based aggregation for biterm topic model. In *ACL*, 489–494.

Gama, J.; Žliobaitė, I.; Bifet, A.; Pechenizkiy, M.; and Bouchachia, A. 2014. A survey on concept drift adaptation. *ACM Comput. Surv.* 46(4):44:1–44:37.

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.

Hua, T.; Ning, Y.; Chen, F.; Lu, C.-T.; and Ramakrishnan, N. 2016. Topical analysis of interactions between news and social media. In *AAAI*, 2964–2971.

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.

Jain, A. K. 2010. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters* 31(8):651–666.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *WWW*, 499–508.

Liang, S.; Cai, F.; Ren, Z.; and de Rijke, M. 2016. Efficient structured learning for personalized diversification. *IEEE Transactions on Knowledge and Data Engineering* 28(11):2958–2973.

Liang, S.; Ren, Z.; and de Rijke, M. 2014. Personalized search result diversification via structured learning. In *KDD*, 751–760.

Liang, S.; Yilmaz, E.; and Kanoulas, E. 2016. Dynamic clustering of streaming short documents. In *KDD*, 995–1004.

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.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to information retrieval*. Cambridge university press.

Minka, T. 2000. Estimating a dirichlet distribution.

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

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

Vosecky, J.; Leung, K. W.-T.; and Ng, W. 2014. Collaborative personalized twitter search with topic-language models. In *SIGIR*, 53–62.

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.

Xie, P.; Pei, Y.; Xie, Y.; and Xing, E. 2015. Mining user interests from personal photos. In *AAAI*, 1896–1902.

Yan, X.; Guo, J.; Lan, Y.; and Cheng, X. 2013. A biterm topic model for short texts. In *WWW*, 1445–1455.

Yan, X.; Guo, J.; Lan, Y.; Xu, J.; and Cheng, X. 2015. A probabilistic model for bursty topic discovery in microblogs. In *AAAI*, 353–359.

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

Zhao, Y.; Liang, S.; Ren, Z.; Ma, J.; Yilmaz, E.; and de Rijke, M. 2016. Explainable user clustering in short text streams. In *SIGIR*, 155–164.