Collaboratively Tracking Interests for User Clustering in Streams of Short Texts

Shangsong Liang, Emine Yilmaz, and Evangelos Kanoulas

Abstract—In this paper, we aim at tackling 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 two user collaborative interest tracking models that aim at tracking changes of each user's dynamic topic distributions in collaboration with their followees' dynamic topic distributions, based both on the content of current short texts and the previously estimated distributions. Our models can be either short-term or long-term dependency topic models. Short-term dependency model collaboratively tracks users' interests based on users' topic distributions at the previous time period only, whereas long-term dependency model collaboratively tracks users' interests based on users' topic distributions at multiple time periods in the past. We also propose two collapsed Gibbs sampling algorithms for collaboratively inferring users' dynamic interests for their clustering in our short-term and long-term dependency topic models, respectively. We evaluate our proposed models via a benchmark dataset consisting of Twitter users and their tweets. Experimental results validate the effectiveness of our proposed models that integrate both users' and their collaborative interests for user clustering by short text streams.

Index Terms—Clustering; Topic Models; Streaming Text; Twitter

1 Introduction

Popular microblogging platforms provide a light-weight form of communication that enables users to broadcast and share information about their recent activities, opinions and status via short texts [1]. 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 [2] and personalized microblog search [3]. In this paper, we study the problem of *collaborative user clustering in the context of streams of short texts*. Our goal is to infer users' and their collaborative topic distributions over time and dynamically cluster users that share interests in streams of short texts.

Most previous work [4, 5] on user clustering uses collections of static, long documents, and hence makes the assumption that users' interests do not change over time. Recent work [6] clusters users in the context of streams of short documents, 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 utilizing both the users' own posts and the users' collaborative information, i.e. their friends' posts, from which we can infer each user's collaborative interests for further improvement of the clustering.

- *S. Liang is the corresponding author and is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. E-mail: liangshangsong@gmail.com.*
- *E. Yilmaz is with the Department of Computer Science, University College London, and The Alan Turing Institute, London, UK. Email: emine.yilmaz@ucl.ac.uk.*
- *E. Kanoulas is with the Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands. E-mail: e.kanoulas@uva.nl.*

Manuscript received May 28, 2018.

Specifically, we propose two **U**ser **C**ollaborative **I**nterest **T**racking topic models for our collaborative user clustering at time t, including a short-term dependency one that collaboratively tracks users' interests based on users' topic distributions at the previous time period $t-1$ only, abbreviated as **UCIT**, and a long-term dependency one that tracks users' interests based on users' topic distributions at multiple time periods $(t-1), (t-2), \ldots, (t-L)$ in the past, abbreviated as **UCIT-L**. Here we let L indicate the length of the history we consider for the inference of the topic distributions at the current time t. When $L = 1$, UCIT-L reduces to the short-term dependency version, UCIT. Our topic models are dynamic multinomial Dirichlet mixture topic models that can infer and track each user's dynamic interests based not only on the user's posts but also her followees' posts for user clustering. Traditional topic models such as latent Dirichlet allocation (LDA) [7] and author topic model [8] 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 streams of short texts as they assume the documents are in static collections.

In our UCIT and UCIT-L topic models, to alleviate the sparsity problem in short texts, and by following previous work [9, 10], 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, the proposed two models, either short-term dependency or long-term dependency, assume 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, the models infer not only a user's but also her followees' interests from the her own posts and also her followees' posts.

In this paper, we extend the work in [11]. Different from [11],

here we propose the long-term dependency user collaborative interest tracking topic model, which tracks users' interests based on users' topic distributions at multiple time periods. We detail the inferences and derivations of the proposed Gibbs sampling algorithms for the proposed models, examine their performance, study and answer a larger variety of research questions. We also expand the motivation of our work in the introduction, our discussion of related work, and our analysis of experimental results. The contributions of the paper are fourfold: (1) We propose two user collaborative interest tracking topic models, short-term and long-term dependency models – UCIT and UCIT-L, that can collaboratively and dynamically track each user's and her followees' interests for user clustering in the context of streaming of short texts. (2) We propose two collapsed Gibbs sampling algorithms for the inference of our short-term and long-term dependency topic models – UCIT and UCIT-L. (3) Our proposed models can collaboratively cluster previous existing users, newly arriving ones, and those with very limited number of posts. (4) We provide a thorough analysis of the two models and of the impact of their key ingredients in user clustering, and demonstrate their effectiveness compared to the state-of-the-art algorithms.

2 Related Work

There are two lines of related work, topic modeling and clustering. We only discuss the most related models and algorithms.

2.1 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 [7]. Since the well-known topic models, probabilistic latent semantic indexing [12] and LDA (Latent Dirichlet Allocation) [7], were proposed, topic models with dynamics have been widely studied. These include the Dynamic Topic Model (DTM) [13], Dynamic Mixture Model (DMM) [14], Topic over Time (ToT) [15], Topic Tracking Model (TTM) [16], infinite topic-cluster model [17], and more recently, generalized dynamic topic model [18], dynamic User Clustering Topic model (UCT) [6, 10], dynamic topic model for search diversification [19], Dynamic Clustering Topic model (DCT) [20] and scaling-up dynamic model [21]. 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 previous dynamic topic models work 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 streams of short texts.

2.2 Clustering

Clustering has been widely studied and applied into a number of applications, such as document clustering in data mining [22] and information retrieval [23, 24]. In this paper we focus on user clustering only. Previous user clustering algorithms are mainly designed to work for web user clustering [25–29]. These papers study users' access information from logged server data including query and click data and then uncovers clusters of these users that exhibit similar information needs. For instance, Elbamby et al. [29] study the problem of content-aware user clustering in the context of wireless small cell networks, where users are supposed to have different preferences over different content types. Buscher et al. [28] cluster users based on user interaction information, including clicks, scrolls and cursor movements for search queries on long text documents. Another line of work, which mostly focuses on content-based similarity, has grouped users by expertise [30, 31]; recent advances in distributed representation learning have given rise to new types of joint topic and entity representations [32]. But, so far, these have not been used for user clustering yet. Zhao et al. [6] and Liang et al. [10] propose user clustering algorithms in the context of streams of short texts. But they do not take both users' followees and their long-term interest distributions into account during tracking users' interests for clustering, and thus there is still some room to improve the performance. To the best of our knowledge, all existing content-based user clustering algorithms do not consider clustering users via fully utilizing information from each user and the user's followees, i.e., user's collaborative information, in the context of streams of short documents. In contrast, our UCIT topic models utilize users' own as well as their corresponding collaborative information for dynamic clustering in streams of short documents.

3 Problem Formulation

The problem we address is tracking users' dynamic interests and clustering 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 at time t, 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. Each cluster $\mathbf{c}$ in $\mathbf{C}_t$ is a set of users who share similar interests at time t. In addition, 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 longer than a predefined small length (for instance, 140 characters in the case of Twitter).

4 Method

In this section, we describe our short-term and long-term dependency User Collaborative Interest Tracking topic models, **UCIT** and **UCIT-L**, aiming at tracking users' and their followees' interests, and dynamically clustering them in the streams.

4.1 Overview

We use Twitter as our default setting of streams of short texts and provide an overview of our proposed UCIT/UCIT-L models in Algorithm 1. Following [6, 33, 34], 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/UCIT-L models. To alleviate the sparsity problem of short texts, and by following recent work on the topic [9, 10], we construct and represent documents by

Algorithm 1: Overview of the proposed models.

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 either UCIT or UCIT-L 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 $\boldsymbol{\psi}_{t,u}$.

TABLE 1
Main notation used in the proposed UCIT and UCIT-L models.

Notation	Gloss	Notation	Gloss
t	Time	L	Length of time periods
u	User	$\mathbf{u}_t$	Set of users at t
$\mathbf{c}$	Cluster of users	$\mathbf{C}_t$	Cluster result
d	Document	$\mathbf{D}_t$	Text stream up to t
$\mathbf{b}_d$	Biterms of d	$\mathbf{B}_t$	Biterms of $\mathbf{D}_t$
z	Topic	Z	Number of topics
v	Word	$\mathbf{v}$	Vocabulary
$\mathbf{f}_{t,u}$	u's all followees at t	$\mathbf{b}_t$	Docs arriving at t
V	Size of the vocabulary	n	Number of words
$m_{t,u,z}$	Number of documents assigned to u on topic z at t	$o_{t,u,z}$	Number of documents assigned to u's followees on topic z at t
$\theta_{t,u,z}$	u's interests on topic z at t	$\alpha_{t,z}$	u' interest persistency at topic z
$\psi_{t,u,z}$	u's followees' interests on topic z at t	$\beta_{t,z}$	u's followees' interest persistency at topic z
$\phi_{t,z,v}$	Word v's distribution on topic z at t	$\gamma_{t,v}$	Word v's distribution persistency at t

their biterms, i.e. word pairs in them (step 1 in Algorithm 1). In the following Section 4.3 and Section 4.4, we propose two dynamic Dirichlet multinomial mixture user collaborative interest tracking topic models, UCIT and UCIT-L, respectively, 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 [22] (step 3 in Algorithm 1). With time moving forward, the clustering result changes dynamically.

4.2 Biterm Construction

In our proposed UCIT /UCIT-L models, we construct word pairs, also called "biterms" [9, 10], before we conduct topic inference. The motivations of constructing word pairs rather than directly using each single word for topic inference are: (1) topics are groups of correlated words, and the correlations are revealed by words' co-occurrence patterns in documents; (2) the underlying topic expressed by a single word is more ambiguous than that of a word pair. Following the previous work [9, 10], we construct biterms for each short document $d \in \mathbf{D}_t$ as:

$$\mathbf{b}_d = \{(v_i, v_j) | v_i, v_j \in d, i \neq j\},$$

where v_i and v_j are two distinct words in a biterm $b = (v_i, v_j)$, $(\cdot, \cdot)$ is unordered, $b \in \mathbf{b}_d$ and $\mathbf{b}_d$ is a collection of biterms extracted from document d. For instance, considered a document "Both Apple and Amazon are companies", we can construct three biterms, i.e., "apple amazon", "apple compan" and "amazon compan" after removing stop-words and stemming. Thus, in a short document stream $\mathbf{D}_t$, we can construct a set of biterms $\mathbf{B}_t$ such that $\mathbf{B}_t = \sum_{d \in \mathbf{D}_t} \mathbf{b}_d$. Let $\mathbf{b}_{t,u}$ be a collection of biterms generated from documents posted by user u at t. In the inference of our UCIT and UCIT-L models in the following, unlike most topic models where topic assignment is sampled to each independent word, we sample a topic assignment for each biterm.

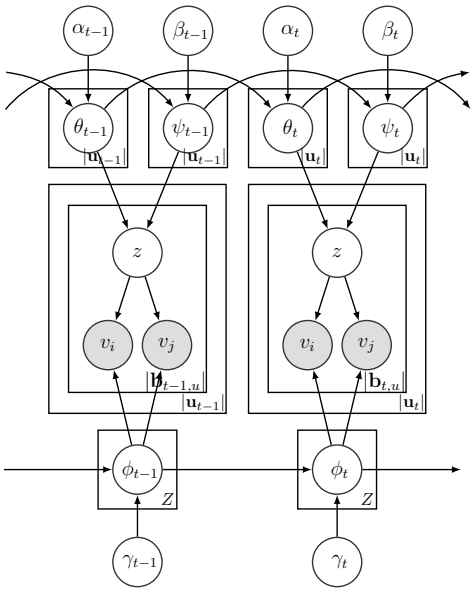

Fig. 1. Graphical representation of our short-term dependency user collaborative interest tracking clustering topic model, UCIT. Shaded nodes represent observed variables.

4.3 Short-Term Dependency UCIT Topic Model

Modeling Interests over Short-Term Period. The goal of short-term dependency 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 short-term dependency UCIT model, where unshaded and shaded nodes indicate latent and observed variables, respectively.

Given a user u, to track the dynamics of her interests, we make the assumption that the mean of her 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 [10, 14, 16, 21, 35], 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 $\boldsymbol{\theta}_{t-1,u}$:

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

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 [36]. 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 $\boldsymbol{\beta}_t = \{\beta_{t,z}\}_{z=1}^{Z}$:

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

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

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

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 short-term UCIT model for documents in stream at time t, is as follows,

i) Draw Z multinomials $\boldsymbol{\phi}_{t,z}$, 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 $v_i \in b$ from multinomial $\boldsymbol{\phi}_{t,z_{t,u,b}}$;
   c) Draw another word $v_j \in b$ from multinomial $\boldsymbol{\phi}_{t,z_{t,u,b}}$.

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. The parameterization of the proposed short-term dependency UCIT topic model is as follows:

$$\begin{aligned}
\boldsymbol{\phi}_{t,z} \mid \boldsymbol{\gamma}_t\boldsymbol{\phi}_{t-1,z} &\sim \text{Dirichlet}(\boldsymbol{\gamma}_t\boldsymbol{\phi}_{t-1,z}) \\
\boldsymbol{\theta}_{t,u} \mid \boldsymbol{\alpha}_t\boldsymbol{\theta}_{t-1,u} &\sim \text{Dirichlet}(\boldsymbol{\alpha}_t\boldsymbol{\theta}_{t-1,u}) \\
\boldsymbol{\psi}_{t,u} \mid \boldsymbol{\beta}_t\boldsymbol{\psi}_{t-1,u} &\sim \text{Dirichlet}(\boldsymbol{\beta}_t\boldsymbol{\psi}_{t-1,u}) \\
z_{t,u,b} \mid ((1-\lambda)\boldsymbol{\theta}_{t,u} + \lambda\boldsymbol{\psi}_{t,u}) &\sim \text{Multinomial}((1-\lambda)\boldsymbol{\theta}_{t,u} \\
&\quad +\lambda\boldsymbol{\psi}_{t,u}) \\
v_i \in b \mid \boldsymbol{\phi}_{t,z_{t,u,b}} &\sim \text{Multinomial}(\boldsymbol{\phi}_{t,z_{t,u,b}}) \\
v_j \in b \mid \boldsymbol{\phi}_{t,z_{t,u,b}} &\sim \text{Multinomial}(\boldsymbol{\phi}_{t,z_{t,u,b}}).
\end{aligned}$$

Note that in the generative process described above, there is a fixed number of latent topics Z. A non-parametric Bayes version of our dynamic topic model that automatically integrates over the number of topics is possible, but we leave this as future work.

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 $P(z_{t,u,b}|\mathbf{z}_{t,-b}, \mathbf{d}_t, \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Psi}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t)$.
      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 $\boldsymbol{\Phi}_t$.

Inference for the Short-Term Dependency UCIT. We employ a collapsed Gibbs sampler [37] 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 $\boldsymbol{\phi}_{t,z}$. 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 for user u and the number of times word v is assigned to topic z 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; and N_{iter} is the total number of iterations.

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, $P(\mathbf{z}_t, \mathbf{d}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Psi}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t)$:

$$\begin{aligned}
&P(\mathbf{z}_t, \mathbf{d}_t|\boldsymbol{\Theta}_{t-1}, \boldsymbol{\Psi}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t) \quad (4)\\
&= (1-\lambda)P(\mathbf{z}_t, \mathbf{d}_t|\boldsymbol{\Theta}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t) + \lambda P(\mathbf{z}_t, \mathbf{d}_t|\boldsymbol{\Psi}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t) \\
&= (1-\lambda)\left(\prod_z \left(\frac{\Gamma(\sum_v(\varkappa_b))}{\prod_v \Gamma(\varkappa_b)} \frac{\prod_v \Gamma(\varkappa_a)}{\Gamma(\sum_v \varkappa_a)}\right)\right)^2 \cdot \prod_u \frac{\Gamma(\sum_z(\varkappa_2))}{\prod_z \Gamma(\varkappa_2)} \frac{\prod_z \Gamma(\varkappa_1)}{\Gamma(\sum_z \varkappa_1)} \\
&\quad +\lambda\left(\prod_z \left(\frac{\Gamma(\sum_v(\varkappa_b))}{\prod_v \Gamma(\varkappa_b)} \frac{\prod_v \Gamma(\varkappa_a)}{\Gamma(\sum_v \varkappa_a)}\right)\right)^2 \cdot \prod_u \frac{\Gamma(\sum_z(\varkappa_4))}{\prod_z \Gamma(\varkappa_4)} \frac{\prod_z \Gamma(\varkappa_3)}{\Gamma(\sum_z \varkappa_3)},
\end{aligned}$$

where $\Gamma(\cdot)$ 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:

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

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

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

$$\prod_{v\in b}\frac{n_{t,z,v}+\gamma_{t,v}\phi_{t-1,z,v}-1}{\sum_{v'=1}^{V}(n_{t,z,v'}+\gamma_{t,v'}\phi_{t-1,z,v'})-1}+$$
$$\lambda\frac{o_{t,u,z}+\beta_{t,z}\psi_{t-1,u,z}-1}{\sum_{z'=1}^{Z}(o_{t,u,z'}+\beta_{t,z'}\psi_{t-1,u,z'})-1}\times$$
$$\prod_{v\in b}\frac{n_{t,z,v}+\gamma_{t,v}\phi_{t-1,z,v}-1}{\sum_{v'=1}^{V}(n_{t,z,v'}+\gamma_{t,v'}\phi_{t-1,z,v'})-1},$$

for the proposed Gibbs sampling (step 5 in Algorithm 2). Our derivations of the joint probability in (4) and the conditional probability in (5) are detailed in Appendix A. 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 $P(\mathbf{z}_t, \mathbf{d}_t\,|\boldsymbol{\Theta}_{t-1}, \boldsymbol{\Psi}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t)$. We apply fixed-point iterations to obtain the optimal $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$. The following update rules 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 the two bounds in [38]:

$$\alpha_{t,z}\leftarrow\frac{(1-\lambda)\alpha_{t,z}\sum_u(\Delta(\varkappa_1)-\Delta(\varkappa_2))}{\sum_u(\Delta(\sum_z\varkappa_1)-\Delta(\sum_z\varkappa_2))},$$
$$\beta_{t,z}\leftarrow\frac{\lambda\beta_{t,z}\sum_u(\Delta(\varkappa_3)-\Delta(\varkappa_4))}{\sum_u(\Delta(\sum_z\varkappa_3)-\Delta(\sum_z\varkappa_4))},$$
$$\gamma_{t,v}\leftarrow\frac{\gamma_{t,v}\sum_z(\Delta(\varkappa_a)-\Delta(\varkappa_b))}{\sum_z(\Delta(\sum_v\varkappa_a)-\Delta(\sum_v\varkappa_b))}. \tag{6}$$

where $\Delta(x)=\frac{\partial\log\Gamma(x)}{x}$ is a Digamma function. After each iteration, we normalize these by $\alpha_{t,z}=\frac{\alpha_{t,z}}{\sum_{z'}\alpha_{t,z'}}$, $\beta_{t,z}=\frac{\beta_{t,z}}{\sum_{z'}\beta_{t,z'}}$ and $\gamma_{t,v}=\frac{\gamma_{t,v}}{\sum_{v'}\gamma_{t,v'}}$, respectively. Our derivations of the update rules for $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$ in (6) are detailed in Appendix B.

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 $\boldsymbol{\phi}_{t,z}$ in our short-term dependency UCIT, as follows, respectively:

$$\theta_{t,u,z}=\frac{m_{t,u,z}+\alpha_{t,z}\theta_{t-1,u,z}}{\sum_{z'=1}^{Z}m_{t,u,z'}+\alpha_{t,z'}\theta_{t-1,u,z'}},$$
$$\psi_{t,u,z}=\frac{o_{t,u,z}+\beta_{t,z}\psi_{t-1,u,z}}{\sum_{z'=1}^{Z}o_{t,u,z'}+\beta_{t,z'}\psi_{t-1,u,z'}}, \tag{7}$$
$$\phi_{t,z,v}=\frac{n_{t,z,v}+\gamma_{t,z}\phi_{t-1,z,v}}{\sum_{v'=1}^{V}n_{t,z,v'}+\gamma_{t,z'}\phi_{t-1,z,v'}}.$$

4.4 Long-Term Dependency UCIT-L Topic Model

Modeling Interests over Long-Term Period. In the previous section (Section 4.3), the most recent distributions $\boldsymbol{\theta}_{t,u}$, $\boldsymbol{\psi}_{t,u}$ and $\boldsymbol{\phi}_{t,z}$ are modeled to depend on the previous distributions, respectively. Previous work has shown that modeling the recent distributions to depend on longer histories can enhance the performance [13–16, 18]. Thus, we propose a long-term dependency user collaborative interest tracking topic model, UCIT-L, that infers the most recent distributions from multiple time periods in the past, i.e., the distributions at time periods $(t-1),(t-2),\ldots,(t-L)$, where L indicates the length of the histories we consider for the inference of the distributions at the most recent time t. Obviously, short-term dependency UCIT model is a special case of the long-term dependency one if $L=1$.

Given a user u, to track the dynamics of their interests in our UCIT-L, we let the user's current interests at time period t depend on a longer time-step history. In particular, we model such a long-term (L-steps) dependency UCIT-L model on the basis of the distribution priors for user u's interests $\boldsymbol{\theta}_{t,u}$ as follows:

$$P(\boldsymbol{\theta}_{t,u}|\{\boldsymbol{\theta}_{t-l,u},\boldsymbol{\alpha}_{t,l}\}_{l=1}^{L})\propto\prod_{z=1}^{Z}\theta_{t,u,z}^{\left(\sum_{l=1}^{L}\alpha_{t,z,l}\theta_{t-l,u,z}\right)-1}, \tag{8}$$

where the mean of $\boldsymbol{\theta}_{t,u}$ in UCIT-L is modeled to be proportional to the weighted sum of the past L "topic trends" in the user u's interests, and $\boldsymbol{\alpha}_{t,l}=\{\alpha_{t,z,l}\}_{z=1}^{Z}$ represents how the user u's interest on topic z at the current time period t is related to the L-previous interests. Previous work [39] shows that recent distributions may matter more than distant ones, and thus a temporal weight π_l being applied to each $\theta_{t-l,u,z}$ in (8), i.e., changing $\alpha_{t,z,l}\theta_{t-l,u,z}$ to $\pi_l\alpha_{t,z,l}\theta_{t-l,u,z}$ accordingly, would yield better performance. However, to keep the focus of this work, we leave this as future work. For a comparison between the short-term and long-term dependency UCIT models for modeling users' dynamic interests we refer to (1) and (8). In contrast to short-term dependency UCIT model, long-term dependency UCIT-L model reduces the information loss and the bias of the inference due to the multiple estimates.

Similarly, to track the dynamic changes of a user u's collaborative interests in our UCIT-L model, we assume a Dirichlet prior, in which the mean of the user u's current collaborative interests $\boldsymbol{\psi}_{t,u}$ evolves from the past L "topic trends" in the user u's collaborative interests:

$$P(\boldsymbol{\psi}_{t,u}|\{\boldsymbol{\psi}_{t-l,u},\boldsymbol{\beta}_{t,l}\}_{l=1}^{L})\propto\prod_{z=1}^{Z}\psi_{t,u,z}^{\left(\sum_{l=1}^{L}\beta_{t,z,l}\psi_{t-l,u,z}\right)-1}, \tag{9}$$

where $\boldsymbol{\beta}_{t,l}=\{\beta_{t,z,l}\}_{z=1}^{Z}$ represents how the user u's collaborative interest on topic z at the current time period t is related to the past L-previous collaborative interests. Similar to (8) we do not apply any temporal weight. For a comparison between the short-term and long-term dependency models for modeling users' collaborative dynamic interests we refer to (2) and (9).

In a similar way, in our long-term dependency UCIT-L model, the Dirichlet prior of the trend over words $\boldsymbol{\phi}_{t,z}$ on topic z at the current time t can be revised such that $\boldsymbol{\phi}_{t,z}$ is modeled to depend on the past L-previous trends over words:

$$P(\boldsymbol{\phi}_{t,z}|\{\boldsymbol{\phi}_{t-l,z},\boldsymbol{\gamma}_{t,l}\}_{l=1}^{L})\propto\prod_{v=1}^{V}\phi_{t,z,v}^{\left(\sum_{l=1}^{L}\gamma_{t,v,l}\phi_{t-l,z,v}\right)-1}, \tag{10}$$

where $\boldsymbol{\gamma}_{t,l}=\{\gamma_{t-l,z,v}\}_{v=1}^{V}$ represents how the dynamic word distributions over topics at the current time period t are related to the past L-previous ones. Again, here we do not apply any temporal weight. For a comparison between the short-term and long-term dependency UCIT models for modeling the dynamic word distributions we refer to (3) and (10).

The graphical representation of the UCIT-L model is shown in Fig. 2, where unshaded and shaded modes indicate latent and observed variables, respectively. The parameterization of the proposed long-term dependency UCIT-L topic model is as follows:

$$\boldsymbol{\phi}_{t,z}\mid\sum_{l=1}^{L}\boldsymbol{\gamma}_{t,l}\boldsymbol{\phi}_{t-l,z}\sim\quad\text{Dirichlet}(\sum_{l=1}^{L}\boldsymbol{\gamma}_{t,l}\boldsymbol{\phi}_{t-l,z})$$
$$\boldsymbol{\theta}_{t,u}\mid\sum_{l=1}^{L}\boldsymbol{\alpha}_{t,l}\boldsymbol{\theta}_{t-l,u}\sim\quad\text{Dirichlet}(\sum_{l=1}^{L}\boldsymbol{\alpha}_{t,l}\boldsymbol{\theta}_{t-l,u})$$
$$\boldsymbol{\psi}_{t,u}\mid\sum_{l=1}^{L}\boldsymbol{\beta}_{t,l}\boldsymbol{\psi}_{t-l,u}\sim\quad\text{Dirichlet}(\sum_{l=1}^{L}\boldsymbol{\beta}_{t,l}\boldsymbol{\psi}_{t-l,u})$$

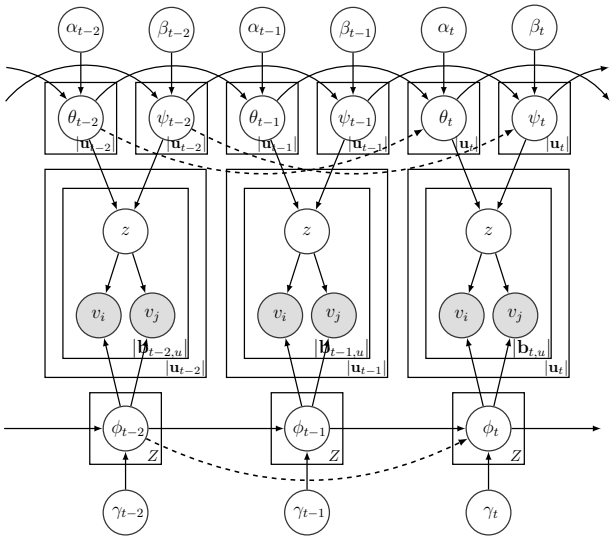

Fig. 2. Graphical representation of our long-term dependency user interest tracking clustering topic model, UCIT-L with $L = 2$. Shaded nodes represent observed variables. The model will return to short-term dependency UCIT if we remove the dashed curved lines.

$$
\begin{aligned}
z_{t,u,b} \mid ((1-\lambda)\boldsymbol{\theta}_{t,u} + \lambda\boldsymbol{\psi}_{t,u}) \sim \quad & \text{Multinomial}((1-\lambda)\boldsymbol{\theta}_{t,u} \\
& +\lambda\boldsymbol{\psi}_{t,u}) \\
v_i \in b \mid \boldsymbol{\phi}_{t,z_{t,u,b}} \sim \quad & \text{Multinomial}(\boldsymbol{\phi}_{t,z_{t,u,b}}) \\
v_j \in b \mid \boldsymbol{\phi}_{t,z_{t,u,b}} \sim \quad & \text{Multinomial}(\boldsymbol{\phi}_{t,z_{t,u,b}}).
\end{aligned}
$$

Inference for the Long-Term Dependency UCIT-L. Our proposed collapsed Gibbs sampling for long-term dependency UCIT-L is similar to that for short-term dependency UCIT; See Algorithm 2. The only difference between the inference for long-term dependency UCIT-L and that for short-term dependency UCIT lies in the way we sample the latent topic for each word pair (step 5 in Algorithm 2) and the update rules for the priors (step 7 in Algorithm 2). Similar to (5), we sample a latent topic for a word pair b in our long-term dependency UCIT-L model by:

$$
\begin{aligned}
&P(z_{t,u,b} = z | \mathbf{z}_{t,-b}, \mathbf{d}_t, \{\boldsymbol{\Theta}_{t-l}, \boldsymbol{\Psi}_{t-l}, \boldsymbol{\Phi}_{t-l}, \boldsymbol{\alpha}_{t,l}, \boldsymbol{\beta}_{t,l}, \boldsymbol{\gamma}_{t,l}\}_{l=1}^{L}) \\
&\propto (1-\lambda)\frac{m_{t,u,z} + \sum_{l=1}^{L}\alpha_{t,z,l}\theta_{t-l,u,z} - 1}{\sum_{z'=1}^{Z}(m_{t,u,z'} + \sum_{l=1}^{L}\alpha_{t,z',l}\theta_{t-l,u,z'}) - 1} \times \\
&\quad \prod_{v\in b}\frac{n_{t,z,v} + \sum_{l=1}^{L}\gamma_{t,v,l}\phi_{t-l,z,v} - 1}{\sum_{v'=1}^{V}(n_{t,z,v'} + \sum_{l=1}^{L}\gamma_{t,v',l}\phi_{t-l,z,v'}) - 1} + \\
&\quad \lambda\frac{o_{t,u,z} + \sum_{l=1}^{L}\beta_{t,z,l}\psi_{t-l,u,z} - 1}{\sum_{z'=1}^{Z}(o_{t,u,z'} + \sum_{l=1}^{L}\beta_{t,z',l}\psi_{t-l,u,z'}) - 1} \times \\
&\quad \prod_{v\in b}\frac{n_{t,z,v} + \sum_{l=1}^{L}\gamma_{t,v,l}\phi_{t-l,z,v} - 1}{\sum_{v'=1}^{V}(n_{t,z,v'} + \sum_{l=1}^{L}\gamma_{t,v',l}\phi_{t-l,z,v'}) - 1}. \qquad (11)
\end{aligned}
$$

The derivation of (11) is similar to that of (5) (see Appendix A). The update rules for $\alpha_{t,z,l}$, $\beta_{t,z,l}$ and $\gamma_{t,v,l}$ in (11) using the two bounds in [38] with fixed-point iterations are as follows, respectively:

$$
\begin{aligned}
\alpha_{t,z,l} &\leftarrow \frac{(1-\lambda)\alpha_{t,z,l}\sum_u(\Delta(\varkappa_1^L) - \Delta(\varkappa_2^L))}{\sum_u(\Delta(\sum_z \varkappa_1^L) - \Delta(\sum_z \varkappa_2^L))}, \\
\beta_{t,z,l} &\leftarrow \frac{\lambda\beta_{t,z,l}\sum_u(\Delta(\varkappa_3^L) - \Delta(\varkappa_4^L))}{\sum_u(\Delta(\sum_z \varkappa_3^L) - \Delta(\sum_z \varkappa_4^L))},
\end{aligned}
$$

$$
\gamma_{t,v,l} \leftarrow \frac{\gamma_{t,v,l}\sum_z(\Delta(\varkappa_a^L) - \Delta(\varkappa_b^L))}{\sum_z(\Delta(\sum_v \varkappa_a^L) - \Delta(\sum_v \varkappa_b^L))}. \qquad (12)
$$

where the parameters $\varkappa^L$ are defined as the following:

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

The derivations of the update rules for $\alpha_{t,z,l}$, $\beta_{t,z,l}$ and $\gamma_{t,v,l}$ in the long-term dependency UCIT-L model are similar to those for $\alpha_{t,z}$, $\beta_{t,z}$ and $\gamma_{t,v}$ in the short-term dependency UCIT one (see Appendix B).

After 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 $\boldsymbol{\phi}_{t,z}$ in our long-term dependency UCIT-L model, as follows, respectively:

$$
\begin{aligned}
\theta_{t,u,z} &= \frac{m_{t,u,z} + \sum_{l=1}^{L}\alpha_{t,z,l}\theta_{t-l,u,z}}{\sum_{z'=1}^{Z}(m_{t,u,z'} + \sum_{l=1}^{L}\alpha_{t,z',l}\theta_{t-l,u,z'})}, \\
\psi_{t,u,z} &= \frac{o_{t,u,z} + \sum_{l=1}^{L}\beta_{t,z,l}\psi_{t-l,u,z}}{\sum_{z'=1}^{Z}(o_{t,u,z'} + \sum_{l=1}^{L}\beta_{t,z',l}\psi_{t-1,u,z'})}, \\
\phi_{t,z,v} &= \frac{n_{t,z,v} + \sum_{l=1}^{L}\gamma_{t,z,l}\phi_{t-l,z,v}}{\sum_{v'=1}^{V}(n_{t,z,v'} + \sum_{l=1}^{L}\gamma_{t,z',l}\phi_{t-l,z,v'})}. \qquad (13)
\end{aligned}
$$

As it can be seen, if we let $L = 1$, (8), (9) and (10) will reduce to (1), (2) and (3), respectively. Thus the short-term dependency UCIT model is a special case of the long-term dependency UCIT-L model.

4.5 Clustering Users

Clustering Previously Seen Users. After we obtain each user's and her collaborative topic distributions, $\boldsymbol{\theta}_{t,u}$ and $\boldsymbol{\psi}_{t,u}$ from either (7) in the short-term dependency UCIT model or (13) in the long-term dependency UCIT-L model, we use the following mixture distribution $\boldsymbol{\rho}_{t,u}$ to represent each user:

$$
\boldsymbol{\rho}_{t,u} = (1-\lambda)\boldsymbol{\theta}_{t,u} + \lambda\boldsymbol{\psi}_{t,u}. \qquad (14)
$$

Then, we can conveniently cluster users based on their interests $\boldsymbol{\rho}_{t,u}$ using the K-means algorithm [22, 40]. For previously unseen users, however, we can not directly utilize (7) in the short-term dependency UCIT or (13) in the long-term dependency UCIT-L for the clustering, as $\boldsymbol{\theta}_{t-1,u}$ and $\boldsymbol{\psi}_{t-1,u}$ are not defined at the current time 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.

Clustering Previously Unseen Users. For newly arriving users, we can not directly utilize either (7) or (13) for the clustering, as $\boldsymbol{\theta}_{t-1,u}$ and $\boldsymbol{\psi}_{t-1,u}$ are not available at t. We obtain the probability of a new user u_{new} being interested in topic z at time t, i.e., $\theta_{t,u_{\text{new}},z}$, as:

$$
\theta_{t,u_{\text{new}},z} = P(z|t,u_{\text{new}}) = \sum_{b\in\mathbf{b}_{t,u_{\text{new}}}} P(z|t,b)P(b|t,u_{\text{new}}), \quad (15)
$$

where the first term $P(z|t,b)$ is obtained as:

$$\begin{aligned} P(z|t,b) &= \frac{P(v_i|t,z)P(v_j|t,z)P(z|t)}{P(b|t)} \\ &= \frac{P(v_i|t,z)P(v_j|t,z)P(z|t)}{\sum_{z'} P(z'|t)P(v_i|t,z')P(v_j|t,z')} \\ &= \frac{P(z|t)\phi_{t,z,v_i}\phi_{t,z,v_j}}{\sum_{z'} P(z'|t)\phi_{t,z',v_i}\phi_{t,z',v_j}}, \end{aligned} \tag{16}$$

where $P(v|t,z)$ is the probability of word v associated with topic z at time t, i.e., $\phi_{t,z,v}$, and $P(z|t)$ is the probability of topic z at time t. We obtain $P(z|t)$ for (16) as:

$$P(z|t) = \frac{n_t(z,v)}{n_t(v)}, \tag{17}$$

where $n_t(z,v)$ and $n_t(v)$ are the total number of words assigned to topic z and the total number of words at time t, respectively.

Then we estimate the second term $P(b|t,u_{\text{new}})$ in (15) as:

$$P(b|t,u_{\text{new}}) = \frac{n_{t,u_{\text{new}}}(b)}{\sum_{b'} n_{t,u_{\text{new}}}(b')}, \tag{18}$$

where $n_{t,u_{\text{new}}}(b)$ is the number of biterm b in $\mathbf{b}_{t,u_{\text{new}}}$.

Finally, after applying (16), (17) and (18) to (15), we can obtain the new user's interests $\boldsymbol{\theta}_{t,u_{\text{new}}}$ at time t. Following the same way, we can obtain the new user's collaborative interests $\boldsymbol{\psi}_{t,u_{\text{new}}}$. We then cluster this new user into a cluster $\mathbf{c}_{t,u_{\text{new}}}$ where they share similar interests with other users in the cluster:

$$\mathbf{c}_{t,u_{\text{new}}} = \arg\max_{\mathbf{c}_{t,u}} \sum_{u \in \mathbf{c}_{t,u}} \frac{\cos\left(\boldsymbol{\rho}_{t,u}, \boldsymbol{\rho}_{t,u_{\text{new}}}\right)}{|\mathbf{c}_{t,u}|}. \tag{19}$$

where as denoted in (14), $\boldsymbol{\rho}_{t,u} = (1-\lambda)\boldsymbol{\theta}_{t,u} + \lambda\boldsymbol{\psi}_{t,u}$. We then update the user set $\mathbf{u}_t$ as $\mathbf{u}_t \leftarrow \mathbf{u}_t \cup \{u_{\text{new}}\}$.

5 Experimental Setup

In what follows, we detail our research questions, dataset, baselines and evaluation metrics.

5.1 Research Questions

The research questions that guide the remainder of the paper are: **(RQ1)** How do UCIT and UCIT-L 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 and UCIT-L? **(RQ3)** What is the contribution of the collaborative information for user clustering? **(RQ4)** What is the clustering performance of long-term dependency UCIT-L model compared to that of the short-term dependency one? **(RQ5)** What is the quality of the topical representation inferred by UCIT and UCIT-L? **(RQ6)** Can UCIT/UCIT-L infer users' dynamic interests for user clustering and make the clustering results explainable? **(RQ7)** Is the performance of UCIT/UCIT-L sensitive to the number of latent topics? **(RQ8)** What is the generalization performance of UCIT compared to state-of-the-art topic models? **(RQ9)** How does the complexity of UCIT/UCIT-L compared to state-of-the-art methods?

5.2 Dataset

In order to answer our research questions, we work with a dataset collected from Twitter. [1] The dataset contains 1,375 active users and their tweets spanning a time period that starts on each user's registration date 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).

[1] The dataset can be downloaded from https://bitbucket.org/sliang1/uct-dataset/get/UCT-Dataset.zip

5.3 Baselines

We compare our UCIT and UCIT-L models 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 [41].

Latent Dirichlet Allocation (LDA). This model infers topic distributions specific to each document via the LDA model.

Author topic model (AuthorT). This model [42] infers topic distributions specific to each user in a static dataset.

Dynamic topic model (DTM). This model [13] utilizes a Gaussian distribution for inferring topic distribution of long text documents in streams.

Continuous time dynamic topic model (cDTM). cDTM [43] is a dynamic topic model that uses Brownian motion to model latent topics through a sequential collection of long texts.

Topic over time model (ToT). This model [15] normalizes timestamps of long documents in a collection and then infers topics distribution for each document.

Topic tracking model (TTM). This model [16] 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, cDTM, 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_{avg} and $\text{UCIT}_{\text{avg+}\boldsymbol{\psi}}$, 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 UCIT model. Note that we use $\text{UCIT}_{\boldsymbol{\psi}}$ 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_{avg} and $\text{UCIT}_{\boldsymbol{\psi}}$ will reduce to the state-of-the-art user clustering baseline, UCT [6], where each user's friends' posts are not taken into account, and similarly, when both $\lambda_1 = 0$ and $\lambda_2 = 0$, $\text{UCIT}_{\text{avg+}\boldsymbol{\psi}}$ will reduce to UCT.

5.4 Evaluation Metrics and Settings

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 [44]. Higher Precision, Purity, NMI scores indicate better user clustering performance. We further use H-score [45] to evaluate the quality of topical representations of user clusters generated by our models 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 models we adopt Perplexity. This metric, used by convention in many topic models [7], 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.

To tune the parameter λ we split the dataset into two parts by timestamps: the first half of the dataset for training, and the remaining for testing. Traditional k-fold cross-validation is not applicable to temporally ordered data since it would corrupt the order [46]. The statistical significance of the observed differences between the performance of two models is tested using a two-tailed paired t-test and is denoted using $^{\blacktriangle}$ (or $^{\blacktriangledown}$) for $\alpha = .01$, and $^{\triangle}$ (and $^{\triangledown}$) for $\alpha = .05$.

6 Results and Analysis

In the following, we discuss and analyze our experimental results and answer the research questions **RQ1** to **RQ9**.

6.1 Effectiveness of UCIT

We begin by answering research question **RQ1**. To better understand the performance of the proposed models, we use short-term dependency UCIT model as a representative, as the performance of long-term dependency UCIT-L is at least the same or better than that of the short-term dependency one. The performance comparisons between short-term dependency UCIT and long-term dependency UCIT-L are shown in subsection 6.4. Table 2 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 2: (1) All the three versions of UCIT model, UCIT_{avg}, $\text{UCIT}_{\text{avg+}\psi}$ and UCIT_{ψ}, 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. (2) Both UCIT_{ψ} and $\text{UCIT}_{\text{avg+}\psi}$ outperform UCIT_{avg}, which demonstrates that utilizing the inferred collaborative interests ψ can yield better performance compared to simply utilizing the average of followees' interests as collaborative information. (3) UCIT_{ψ} works better than $\text{UCIT}_{\text{avg+}\psi}$, which demonstrates that the contribution of ψ is more critical for user clustering compared to that of the average of the interests for user clustering. The reason UCIT_{ψ} works better than $\text{UCIT}_{\text{avg+}\psi}$ is, again, that using average interests as collaborative interests from followees is less effective than that explicitly inferred in the model.

6.2 Impact of Time Interval Length

We now turn to answer research question **RQ2**. We use short-term dependency UCIT model as a representative only. To understand the influence on UCIT of the length of the time period used for evaluation, in Fig. 3 we compare the performance for different time intervals: a week, a month, a quarter, half a year and a year, respectively.

TABLE 2
Clustering performance of UCIT and the baselines using a time period of a month. Statistically significant differences between UCIT_{ψ} and $\text{UCIT}_{\text{avg+}\psi}$, between UCIT_{ψ} and UCIT_{avg} are marked in the upper and lower right hand corner of UCIT_{ψ}'s score, respectively. The statistical significance is tested using a two-tailed paired t-test.

	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
cDTM	.340	.583	.510	.496
TTM	.344	.587	.522	.521
ToT	.359	.605	.552	.582
GSDMM	.398	.632	.592	.561
UCIT_{avg}	.505	.714	.718	.818
$\text{UCIT}_{\text{avg+}\psi}$	.560	.736	.762	.861
UCIT_{ψ}	.583$^{\blacktriangle}$	.746$^{\triangle\blacktriangle}$	.776$^{\triangle\blacktriangle}$	.883$^{\blacktriangle}$

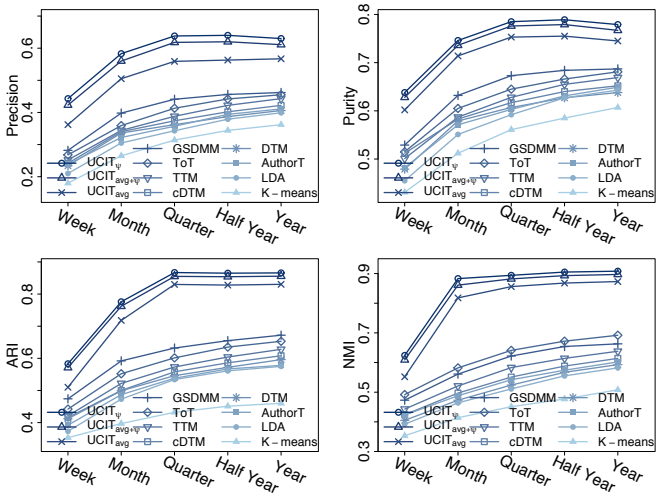

Fig. 3. Precision, Purity, ARI and NMI performance of our UCIT models and the baselines on time periods of a week, a month, a quarter, half a year, and a year, respectively.

According to Fig. 3, all the UCIT models, UCIT_{avg}, $\text{UCIT}_{\text{avg+}\psi}$ and UCIT_{ψ}, outperform 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 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.

6.3 Contribution of the Collaborative Interests

Subsequently, 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. Here, again, we take UCIT as a representative, as the experimental results with UCIT-

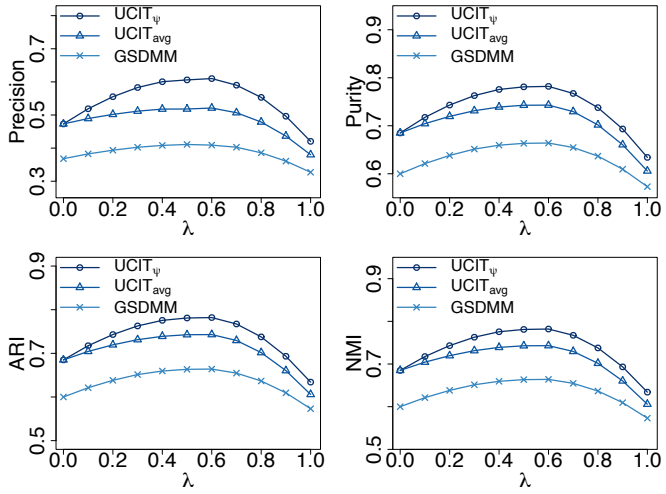

Fig. 4. Precision, Purity, ARI and NMI performance of our UCIT models, UCIT$_{avg}$ and UCIT$_\psi$, and GSDMM on varying scores of λ, respectively.

L is qualitatively the same to that of UCIT. We vary λ and show the performance of our models, UCIT$_\psi$ and UCIT$_{avg}$, and the best baseline model, GSDMM in Fig. 4. 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$_{avg+\psi}$, as it obtains quantitively similar to UCIT$_{avg}$ 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$_{avg}$ 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. 4 also shows that UCIT$_\psi$ that uses collaborative interests for clustering outperforms UCIT$_{avg}$ 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$_{avg}$ 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. 4 that both UCIT$_\psi$ and UCIT$_{avg}$ outperform UCT.

6.4 Comparison Between Short-term UCIT and Long-term Dependency UCIT-L models

Next, we turn to research question **RQ4** to examine the impact of dependency length on the long-term dependency UCIT-L model. Table 3 shows the comparisons between the performance of the UCIT-L model and that of the short-term UCIT model. In the table, we use UCIT$_\psi$-L as a representative only, as UCIT$_\psi$-L performs better than UCIT$_{avg}$-L and UCIT$_{avg+\psi}$-L in most cases. To be clear, we denote the short-term dependency model as UCIT$_\psi$-1 ($L = 1$) and the long-term version as UCIT$_\psi$-L ($L \geq 2$) with L being the dependency length, i.e., the number of pervious time periods under consideration for collaboratively inferring the current topic distributions.

As can be seen in Table 3, the long-term dependency UCIT$_\psi$-L with $L \geq 2$ statistically significantly outperforms the short-term dependency UCIT$_\psi$-1 on all the metrics when using a week as time slice. The performance of UCIT$_\psi$-L levels off when $L \geq 12$ and thus it is not shown in the table when $L \geq 12$. The performance of UCIT$_\psi$-L with $L \geq 2$ outperforms that of the short-term dependency UCIT$_\psi$-1 as well and it also levels off when $L \geq 3$ when using a month as time slice. These findings verify the merit of the proposed long-term dependency UCIT-L model that it can enhance the performance of user clustering when more past information of user's own and her friends' interest distributions are integrated into the model. In other words, the long-term dependency UCIT-L model works better than the short-term dependency version especially in terms of using a week and a month as time slices. When the time slices are set to be a quarter or longer–half a year, the performance of the long-term dependency model is almost the same as that of the short-term dependency one. This is because the interests of the users inferred by both the long-term and short-term dependency models seem to be the same when the time slices are sufficient long.

In the remainder of the analysis, to further study the performance of our collaborative user clustering models independently of the length of the dependency, we will continue to focus on the short-term dependency UCIT model only. The performance of long-term dependency UCIT-L with $L \geq 2$ is at least the same or better than that of the short-term dependency version.

TABLE 3
The impact of dependency length L on collaborative user clustering. UCIT$_\psi$-L is the long-term dependency UCIT model with L being the length of the dependency under consideration. UCIT$_\psi$-1 is the short-term dependency UCIT model. Statistically significant differences between UCIT$_\psi$-L when $L \geq 2$ and UCIT$_\psi$-1 per metric are tested using a two-tailed paired t-test and are denoted in the upper right hand corner of the UCIT$_\psi$-L scores, respectively.

	a week				a month			
	Pre.	Purity	ARI	NMI	Pre.	Purity	ARI	NMI
UCIT$_\psi$-1	.443	.638	.583	.624	.583	.746	.776	.883
UCIT$_\psi$-2	.455$^\triangle$	.647$^\triangle$	.598$^\blacktriangle$	.642$^\blacktriangle$	.610$^\blacktriangle$	.754$^\triangle$	.804$^\blacktriangle$	.885
UCIT$_\psi$-3	.467$^\blacktriangle$	.649$^\blacktriangle$	.604$^\blacktriangle$	.658$^\blacktriangle$	.623$^\blacktriangle$	.765$^\blacktriangle$	.823$^\blacktriangle$	.890$^\triangle$
UCIT$_\psi$-4	.472$^\blacktriangle$	.654$^\blacktriangle$	.613$^\blacktriangle$	.663$^\blacktriangle$	.624$^\blacktriangle$	.766$^\blacktriangle$	.823$^\blacktriangle$	.890$^\triangle$
UCIT$_\psi$-5	.478$^\blacktriangle$	.662$^\blacktriangle$	.627$^\blacktriangle$	.677$^\blacktriangle$	.624$^\blacktriangle$	.766$^\blacktriangle$	.823$^\blacktriangle$	.890$^\triangle$
UCIT$_\psi$-6	.484$^\blacktriangle$	.670$^\blacktriangle$	.635$^\blacktriangle$	.684$^\blacktriangle$	.624$^\blacktriangle$	.766$^\blacktriangle$	.823$^\blacktriangle$	.890$^\triangle$
UCIT$_\psi$-7	.489$^\blacktriangle$	.676$^\blacktriangle$	.644$^\blacktriangle$	.706$^\blacktriangle$	.624$^\blacktriangle$	.766$^\blacktriangle$	.823$^\blacktriangle$	.890$^\triangle$
UCIT$_\psi$-8	.593$^\blacktriangle$	.680$^\blacktriangle$	.652$^\blacktriangle$	.712$^\blacktriangle$	.623$^\blacktriangle$	.765$^\blacktriangle$	.821$^\blacktriangle$	.889$^\triangle$
UCIT$_\psi$-9	.502$^\blacktriangle$	.684$^\blacktriangle$	.663$^\blacktriangle$	.723$^\blacktriangle$	.623$^\blacktriangle$	.766$^\blacktriangle$	.823$^\blacktriangle$	.890$^\triangle$
UCIT$_\psi$-10	.508$^\blacktriangle$	.795$^\blacktriangle$	.675$^\blacktriangle$	.731$^\blacktriangle$	.624$^\blacktriangle$	.766$^\blacktriangle$	.823$^\blacktriangle$	.890$^\triangle$
UCIT$_\psi$-11	.517$^\blacktriangle$	.704$^\blacktriangle$	.684$^\blacktriangle$	.739$^\blacktriangle$	.624$^\blacktriangle$	.766$^\blacktriangle$	.823$^\blacktriangle$	.890$^\triangle$
UCIT$_\psi$-12	.521$^\blacktriangle$	.708$^\blacktriangle$	.692$^\blacktriangle$	.745$^\blacktriangle$	.624$^\blacktriangle$	.766$^\blacktriangle$	.823$^\blacktriangle$	.890$^\triangle$
	a quarter				half a year			
UCIT$_\psi$-1	.638	.785	.867	.894	.640	.789	.865	.905
UCIT$_\psi$-2	.643	.787	.868	.896	.644	.789	.865	.907
UCIT$_\psi$-3	.640	.785	.867	.895	.643	.787	.866	.906
UCIT$_\psi$-4	.639	.785	.867	.896	.642	.789	.866	.905

6.5 Quality of Topic Representation

Here, we turn to research question **RQ5**. In order to answer the question 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 and thus each cluster in the ground-truth clusters of users has lower average intra-cluster distance and higher inter-cluster distance. Fig. 5 shows the result. It is clear from Fig. 5 that the UCIT models outperform all other baselines, i.e., the average inter-cluster distance in the clusters generated by UCIT is smaller than that in the clusters generated

by any of the baseline models, which demonstrates a better quality of topical representation of UCIT models in contrast to other baselines. Note that the H-score cannot be computed for the baseline GSDMM, as it assigns one single topic to each short document and each user.

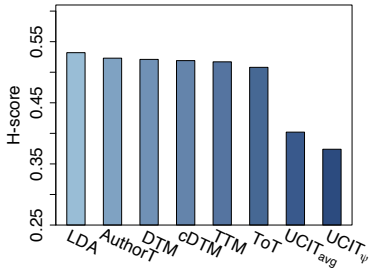

Fig. 5. Quality of topic representations evaluated by H-score, for UCIT and the baselines using time periods of a quarter.

To further analyze the quality of topic representation, we show the top-K words from an example output cluster $\mathbf{c}$ and two users in this cluster, respectively. The top-K words from an example output cluster is generated by this way: the words are first ranked by $P(v \mid t, \mathbf{c})$, i.e., the probability of word v given a cluster $\mathbf{c}$ at time t. We compute $P(v \mid t, \mathbf{c})$ as:

$$P(v \mid t, \mathbf{c}) = \frac{1}{|\mathbf{c}|} \sum_{u \in \mathbf{c}} \sum_{z} P(v \mid t, z) P(z \mid t, u)$$
$$= \frac{1}{|\mathbf{c}|} \sum_{u \in \mathbf{c}} \sum_{z} \phi_{t,z,v} \left((1-\lambda)\theta_{t,u,z} + \lambda \psi_{t,u,z} \right),$$

where $|\mathbf{c}|$ is the total number of users in the cluster $\mathbf{c}$. The top-K words with the highest probabilities $P(v \mid t, \mathbf{c})$ are then selected to represent the cluster at time t. Similarly, we rank the words in decreasing order of the probability $P(v \mid t, u)$ to obtain the top-K words to represent user u. $P(v \mid t, u)$ is computed as $P(v \mid t, u) = \sum_z P(v \mid t, z) P(z \mid t, u) = \sum_z \phi_{t,z,v} \left((1-\lambda)\theta_{t,u,z} + \lambda \psi_{t,u,z} \right)$. Table 4 shows the top 30 words extracted from example output clusters generated by UCIT and the baseline ToT (Again, we do not use GSDMM for comparisons here as it assigns one single topic to each short document), and the two users in the clusters, respectively. Words in the first row of the table are for representing a cluster, while those in the second and the third are for representing the two users, respectively. As can be seen from Table 4, the two users in the same cluster generated by our UCIT share more similar interests represented by words such as "landscape", "city" and "bridge" from the topic "city" and words such as "house", "design" and "apartment" from the topic "home", compared to those generated by the baseline model, ToT. This again, illustrates that topic representation of UCIT is better than that of the baseline models.

6.6 Dynamic Topic Representation of Users

To address **RQ6**, we conduct a qualitative analysis and examine if the clustering results produced by UCIT is explainable. We randomly choose two example users and show their interests tracked by UCIT$_{\psi}$ over the five quarters from April 2014 to May 2015. We use UCIT$_{\psi}$ as a representative here. Table 5 shows top 30 words at each quarter for each user, respectively, where the top words are from the most probable topics of the user and the 30 most probable words from the topics. In Table 5, the first row shows the top 30 words per quarter to represent an example user's interests, whereas the second row shows the top 30 words per quarter to represent another example user's interests.

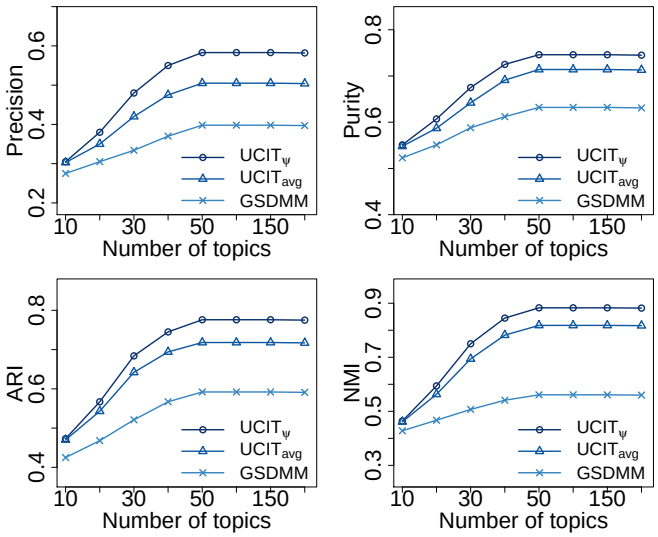

Fig. 6. Precision, Purity, ARI and NMI performance of our UCIT$_{avg}$ and UCIT$_{\psi}$ models, and GSDMM on varying number of topics, respectively.

As can be seen in Table 5, the first user's interests vary from time to time. In the first three quarters, i.e., from April 2014 to December 2014, the first user is interested in the topic "food" represented by the words "food", "pork", "beef" etc. As time processes, in the last two quarters, i.e., from January 2015 to May 2015, the user's interests shift to the topic "sport" represented by the words "sports", "team", "championship" etc. In contrast, the second user's interests seem to be stable over the whole period. The second user' interests mainly focus on the topic "internet" represented by the words "internet", "network", "Google" etc. over these five quarters. This example demonstrates that our UCIT/UCIT-L models can capture dynamic topic distributions to represent the interests of users and the result of dynamic clustering is explainable and understandable in streams of short texts.

6.7 Effect of Number of Topics

To answer **RQ7**, we examine the effect of the number of latent topics passed as an input parameter to UCIT and the baselines on the performance. We vary the number of latent topics from 10 to 200, and compare the performance in terms of all the metrics. Again, we use GSDMM as the representative of other baselines.

Fig. 6 shows the comparison result. As shown in the figure, when only 10 latent topics are modeled, UCIT$_{\psi}$ and UCIT$_{avg}$ yields almost the same performance, but both of them still outperform the best baseline GSDMM. With the number of latent topics increasing from 10 to 50, the performance of all the models increases and the positive performance differences between UCIT$_{\psi}$ and UCIT$_{avg}$, between UCIT$_{\psi}$/UCIT$_{avg}$ and GSDMM, also increase. When the number of latent topics further increases from 50, the performance of all the models reaches a plato, which demonstrates another merit of the proposed models: they are robust and insensitive to the number of topics and once enough topics are used they are able to improve the performance.

6.8 Perplexity Performance

In order to answer **RQ8** and understand the generalization performance of UCIT and the baseline models, we use perplexity for

TABLE 4
Top 30 words representing a cluster and two users extracted by UCIT$_\psi$ and ToT, respectively. Words in the first row represent a cluster, while words in the second and third rows represent two users in the cluster, respectively. Words in bold represent the most coherent words for topics; those in italic represent less coherent words and others neither in bold nor in italic represent irrelevant words.

UCIT$_\psi$	ToT
architecture *house* home design railway greater **urban museum buildings** *studio* **tower** headdesk **residence** *pavilion* **centre bridge** *skyscraper art* **construction** expo **waterfront** *stadium* tiny *skyline* **landscape** partners *apartment headquarters* **garden city**	**design buildings landscape urban** *art* videos **railway** london **station** *skyscraper studio* **tower** japan *stadium pavilion* **centre** *apartment* **bridge** symbols **construction** france *neighbors* **waterfront** science chemistry *skyline* **landscape** call water molecular
landscape *windows* **urban** holidays **house** *art* postcard **home design centre** bicycle *underground* **city** conference *apartment* **garden railway residence** stories *weather* mile *school* **waterfront** musician dancing **bridge** *skyline* competition *pavilion* **museum**	**urban** *stadium* help government report global **landscape** future *kitchen* internet youtube open *university* **design** *education* **station** digital work research history *park peace* fruit **home** *headquarters* apple **bridge** technology security health
buildings construction city landscape virtual *beauty square* **waterfront** obama police **design museum house** *entries* **flat** *art education* calculate currency *apartment headquarters* million *CASA* **bridge** ambassador atelier **transportation** *culture* computer ipad	**buildings landscape** banner style **construction** facebook chicken fried *apartment* pillow floor photographer **city** *studio* **museum** nikon consume dish *headquarters* shooter fitness *square CASA beauty* **house** northwestern catholic *education* pattern freespirit

TABLE 5
Top 30 words representing two users' interests over time tracking by UCIT$_\psi$, covering five quarters from April 2014 to May 2015 in the first and the second rows, respectively. Words in bold represent the most coherent words for topics; those in italic represent less coherent words and others neither in bold nor in italic represent irrelevant words.

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
food pork beef taste disease **chicken rice restaurant fish soup** *Penang* mountain **sauce noodles** *phenol China* science *Thailand* **hotpot** *boston Japan* society **soos whiskey** *London* university **salad recipe** swimming **healthcity**	**food restaurant soup curry spicy** falsum **garlic** chicklette *chef Korean* ticktock **rice** urban *Leuven Singapore* **seafood dumpling** *China* **hotpot duck fried noodles** *Kuala Hongkong* panda market **cookie** *Japan* **soos prawn**	**food garlic pork smoked fish** *Thailand* **rice** bouldering **meal seafood cuisine chilli cookie** sky **pudding noodles** neighbor *China* **hotpot** weather **masala sausage** *devilled* **cake soos dumpling** *Japan* map **nutrition** navigator	**game team score** *season* **championship** Android **fans football coach sports basketball NBA baseball players** *brave college* jewellery **bowling** *league final beat field* **win** *net* **congrats defense playoff offense stadium sportscenter**	**championship team sports** time *national* **NBA fans basketball** *quarter mark* **players coach** *final* languages **halftime** rays **defense** *league* video **stadium** *hoops* **win defense shot tournament** *yankees former* **soccer record** *national*
www http internet center **Facebook Google Twitter retweets tweets** location **API technology apple mobile windows Android Microsoft** *game* **youtube release** opinion **wikipedia app media** *style* **software** person **iphone** people *search*	**video internet tweets** law **www** freedom **http** *interface* **Facebook Google** *innovation* factory *online* **digital** fiction **Twitter source cloud startup amazon blog** *company* **Google** *search* **Linkedlin design security code** human **anime**	**http www service network tools internet wikipedia** history *entities* **system Twitter management** target *world* **tweets Linux devices** errors *store* **email Facebook test support** time **version software iphone Microsoft** work *game*	**Twitter tweets www http** time **API internet app** interest **platform wikipedia** *free* **office** *industry* **html website computer science** *experience* **Facebook programming** people **browser ipad Google** association **customers iphone mobility media**	**internet www http launch Twitter Google tweets wikipedia** task **computer LinkedIn** *entities* **Facebook science privacy Java javascript** errors **gmail github blogs tabs** *download* **iwatch ipad** people **iphone amazon** *company search*

the evaluation. Fig. 7 shows the result. A lower perplexity score indicates better generalization performance. As it can be observed, the performance of UCIT$_\psi$ is almost the same as that of the best baseline GSDMM and better than that of all other baseline models. 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.

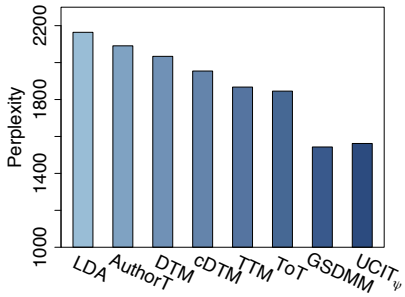

Fig. 7. Generalization performance evaluated by Perplexity, for UCIT and the baselines using time periods of a quarter.

6.9 Complexity Analysis

Finally, we turn to research question **RQ9** to make the comparison between the complexity of UCIT model and that of the state-of-the-art baseline ones. At each time interval t, the time complexity of short-term UCIT$_{\text{avg}}$, UCIT$_{\text{avg+}\psi}$ or UCIT$_\psi$ is proportional to $O(N_{iter} \times |\mathbf{u}_t| \times \frac{1}{|\mathbf{u}_t|}\sum_{u\in\mathbf{u}_t}\mathbf{b}_{t,u} \times Z \times O_{\text{cluster}})$ where $\frac{1}{|\mathbf{u}_t|}\sum_{u\in\mathbf{u}_t}\mathbf{b}_{t,u}$ is the average number of biterms for the users, and O_{cluster} is the time complexity of the clustering algorithm integrated into the UCIT models. The time complexity of our UCIT models is on par with that of the baseline models, DTM, cDTM, ToT, TTM, LDA/AuthorT, where the complexity of each

TABLE 6
Average running time comparisons measured by mins.

	Week	Month	Quarter	Half Y.	Year
GSDMM	28.71	120.01	358.42	715.33	1535.25
LDA	30.82	131.75	393.35	789.76	1576.63
ToT	31.73	135.45	405.90	812.34	1628.39
AuthorT	32.67	139.32	417.13	835.44	1672.21
TTM	33.60	142.57	412.74	862.43	1721.36
cDTM	34.21	147.25	435.72	872.35	1753.37
DTM	34.53	148.72	440.73	878.59	1762.18
UCIT	40.13	170.57	513.45	1028.71	2052.14
UCIT$_{\psi}$-10	41.07	175.42	525.35	1051.34	2101.83

is $O(N_{iter} \times |\mathbf{u}_t| \times \frac{1}{|\mathbf{u}_t|}\sum_{u\in\mathbf{u}_t}\mathbf{v}_{t,u} \times Z \times O_{\text{cluster}})$, where $\frac{1}{|\mathbf{u}_t|}\sum_{u\in\mathbf{u}_t}\mathbf{v}_{t,u}$ is the average number of words in documents associated with a user, and slightly worse than that of GSDMM, the complexity of which is $O(N_{iter} \times |\mathbf{u}_t| \times \frac{1}{|\mathbf{u}_t|}\sum_{u\in\mathbf{u}_t}\mathbf{v}_{t,u} \times O_{\text{cluster}})$. At each time interval t, the complexity of the long-term dependency UCIT model is proportional to $O(N_{iter} \times |\mathbf{u}_t| \times \frac{1}{|\mathbf{u}_t|}\sum_{u\in\mathbf{u}_t}\mathbf{b}_{t,u} \times Z \times O_{\text{cluster}} \times L)$, which is somewhat worse than that of the short-term dependency UCIT model. Table 6.9 makes comparisons on average processing time of our UCIT and UCIT-L (UCIT$_{\psi}$-10) and the baselines that were run on a computer with Scientific Linux release 6.9 (Carbon) operation system and 32GB memory for different time periods from a month to a year. It shows that the running time of all the models linearly increases as larger time periods are considered. Performing clustering online with update-to-date parameters of the models definitely not need to impact the real-time response, as we can perform updating the parameters offline.

7 Conclusion

In this paper, we studied the problem of dynamically clustering users in the context of streams of short texts. We have proposed two user collaborative interest tracking topic models that can infer and track each user's and their followees' dynamic interests for user clustering. Our models can be either short-term dependency (UCIT) or long-term dependency (UCIT-L) topic models. Our two models can effectively handle both the textual sparsity of short documents, and the dynamic nature of users' and their followees' interests over time. Short-term dependency UCIT model collaboratively tracks users' dynamic interests based on users' topic distributions at the previous time period only. In contrast, long-term dependency UCIT-L model collaboratively tracks users' dynamic interests based on users' topic distributions at not only the last time period but also other multiple time periods in the past. To effectively infer users' dynamic interests, we proposed two collapsed Gibbs sampling algorithms for the two collaborative interest tracking topic models. We evaluated the performance of the proposed models in terms of clustering, topical representation and generalization effectiveness, and made comparisons with state-of-the-art models. Our experimental results demonstrated that the models can effectively cluster users in streams of short texts and tracking users' interests based on their topic distributions at longer previous time periods helps to enhance the clustering performance.

As future work, we intend to incorporate other information such as the users' location information 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.

Appendix A
Derivation for Gibbs Sampling in UCIT

In the following, we show the derivation for Gibbs sampling in short-term dependency UCIT only, as the derivation for Gibbs sampling in the long-term dependency UCIT-L is similar. According to the graphical representation of our proposed short-term dependency UCIT model in Fig. 1, the joint distribution $P(\mathbf{z}_t, \mathbf{d}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Psi}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t, \boldsymbol{\gamma}_t)$ can be estimated as:

$$\begin{aligned}&(1-\lambda)P(\mathbf{z}_t, \mathbf{d}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t)\\&+\lambda P(\mathbf{z}_t, \mathbf{d}_t \mid \boldsymbol{\Psi}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t).\end{aligned}$$

In the following, we show the derivation of $P(\mathbf{z}_t, \mathbf{d}_t|\boldsymbol{\Theta}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t)$ only. The derivation of $P(\mathbf{z}_t, \mathbf{d}_t|\boldsymbol{\Psi}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\gamma}_t, \boldsymbol{\gamma}_t)$ is quite similar to that of $P(\mathbf{z}_t, \mathbf{d}_t|\boldsymbol{\Theta}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t)$. We can take advantage of conjugate priors to simplify the integrals. All the symbols are defined in the body of the paper.

$$\begin{aligned}&P(\mathbf{z}_t, \mathbf{d}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t)\\&=P(\mathbf{d}_t \mid \mathbf{z}_t, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\gamma}_t)P(\mathbf{z}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{\alpha}_t)\\&=\int P(\mathbf{b}_t \mid \mathbf{z}_t, \boldsymbol{\Phi}_t)P(\boldsymbol{\Phi}_t \mid \boldsymbol{\Phi}_{t-1}, \boldsymbol{\gamma}_t)d\boldsymbol{\Phi}_t\times\\&\quad\int P(\mathbf{z}_t \mid \mathbf{u}_t, \boldsymbol{\Theta}_t)P(\boldsymbol{\Theta}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{\alpha}_t)d\boldsymbol{\Theta}_t\\&=\left(\int\prod_{z=1}^{Z}\prod_{v=1}^{V}\phi_{t,z,v}^{n_{t,z,v}}\prod_{z=1}^{Z}P(\boldsymbol{\phi}_{t,z} \mid \boldsymbol{\phi}_{t-1,z}, \boldsymbol{\gamma}_t)d\phi\right)^2\\&\quad\times\int\prod_{u=1}^{|\mathbf{u}_t|}\prod_{z=1}^{Z}\theta_{t,u,z}^{m_{t,u,z}}\prod_{u=1}^{|\mathbf{u}_t|}P(\theta_{t,u} \mid \theta_{t-1,u}, \boldsymbol{\alpha}_t)d\theta\\&=\left(\prod_{z=1}^{Z}\frac{\Gamma(\sum_{v=1}^{V}(\gamma_{t,v}\phi_{t-1,z,v}))}{\prod_{v=1}^{V}\Gamma(\gamma_{t,v}\phi_{t-1,z,v})}\right)^2\\&\quad\times\left(\prod_{z=1}^{Z}\frac{\prod_{v=1}^{V}\Gamma(n_{t,z,v}+\gamma_{t,v}\phi_{t-1,z,v}-1)}{\Gamma\left(\sum_{v=1}^{V}n_{t,z,v}+\gamma_{t,v}\phi_{t-1,z,v}-1\right)}\right)^2\\&\quad\times\prod_{u=1}^{|\mathbf{u}_t|}\frac{\Gamma(\sum_{z=1}^{Z}(\alpha_{t,z}\theta_{t-1,u,z}))}{\prod_{z=1}^{Z}\Gamma(\alpha_{t,z}\theta_{t-1,u,z})}\\&\quad\times\prod_{u=1}^{|\mathbf{u}_t|}\frac{\prod_{z=1}^{Z}\Gamma(m_{t,u,z}+\alpha_{t,z}\theta_{t-1,u,z}-1)}{\Gamma\left(\sum_{z=1}^{Z}m_{t,u,z}+\alpha_{t,z}\theta_{t-1,u,z}-1\right)}.\end{aligned}\tag{20}$$

To simplify, we let:

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

Then (20) can be rewrote as:

$$\begin{aligned}&P(\mathbf{z}_t, \mathbf{d}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t)\\&=\left(\prod_z\left(\frac{\Gamma(\sum_v(\varkappa_b))}{\prod_v\Gamma(\varkappa_b)}\frac{\prod_v\Gamma(\varkappa_a)}{\Gamma(\sum_v\varkappa_a)}\right)\right)^2\times\\&\quad\prod_u\frac{\Gamma(\sum_z(\varkappa_2))}{\prod_z\Gamma(\varkappa_2)}\frac{\prod_z\Gamma(\varkappa_1)}{\Gamma(\sum_z\varkappa_1)}.\end{aligned}\tag{21}$$

Following the same derivation, we have:

$$P(\mathbf{z}_t, \mathbf{d}_t \mid \boldsymbol{\Psi}_{t-1}, \boldsymbol{\Phi}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t)$$

$$= \left(\prod_z \left(\frac{\Gamma(\sum_v (\varkappa_b))}{\prod_v \Gamma(\varkappa_b)} \frac{\prod_v \Gamma(\varkappa_a)}{\Gamma(\sum_v \varkappa_a)} \right) \right)^2 \times \prod_u \frac{\Gamma(\sum_z (\varkappa_4))}{\prod_z \Gamma(\varkappa_4)} \frac{\prod_z \Gamma(\varkappa_3)}{\Gamma(\sum_z \varkappa_3)}. \quad (22)$$

With (21) and (22), the joint distribution can be estimated as:

$$\begin{aligned} & P(\mathbf{z}_t, \mathbf{d}_t | \mathbf{\Theta}_{t-1}, \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t) \quad (23) \\ =& (1-\lambda) P(\mathbf{z}_t, \mathbf{d}_t | \mathbf{\Theta}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t) \\ & + \lambda P(\mathbf{z}_t, \mathbf{d}_t | \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t) \\ =& (1-\lambda) \left(\prod_z \left(\frac{\Gamma(\sum_v (\varkappa_b))}{\prod_v \Gamma(\varkappa_b)} \frac{\prod_v \Gamma(\varkappa_a)}{\Gamma(\sum_v \varkappa_a)} \right) \right)^2 \times \\ & \prod_u \frac{\Gamma(\sum_z (\varkappa_2))}{\prod_z \Gamma(\varkappa_2)} \frac{\prod_z \Gamma(\varkappa_1)}{\Gamma(\sum_z \varkappa_1)} + \\ & \lambda \left(\prod_z \left(\frac{\Gamma(\sum_v (\varkappa_b))}{\prod_v \Gamma(\varkappa_b)} \frac{\prod_v \Gamma(\varkappa_a)}{\Gamma(\sum_v \varkappa_a)} \right) \right)^2 \times \\ & \prod_u \frac{\Gamma(\sum_z (\varkappa_4))}{\prod_z \Gamma(\varkappa_4)} \frac{\prod_z \Gamma(\varkappa_3)}{\Gamma(\sum_z \varkappa_3)}. \quad (24) \end{aligned}$$

Applying the chain rule and (24), we can obtain the conditional probability as:

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

where $\mathbf{z}_{t,-(u,b)}$ and $\mathbf{b}_{t,-(u,b)}$ are the topic assignments for all the word pairs (biterms) except the word pair b from user u and the set of word pairs except the word pair b from user u, respectively.

APPENDIX B
DERIVATION OF THE UPDATE RULES

Here, we show the derivation for short-term dependency UCIT only. We apply fixed-point iterations for estimating the parameters $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$ by maximizing the joint distribution $P(\mathbf{z}_t, \mathbf{d}_t | \mathbf{\Theta}_{t-1}, \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t, \boldsymbol{\gamma}_t)$:

$$\begin{aligned} & \max P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Theta}_{t-1}, \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t, \boldsymbol{\gamma}_t) \\ =& \max\{(1-\lambda) P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Theta}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t) + \\ & \lambda P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t)\}. \quad (25) \end{aligned}$$

Because $P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Theta}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t) \geq 0$ and $P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t) \geq 0$, given a $\lambda \geq 0$, (25) can be represented as:

$$\begin{aligned} & \max P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Theta}_{t-1}, \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t, \boldsymbol{\gamma}_t) \\ =& (1-\lambda) \max P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Theta}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t) + \\ & \lambda \max P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t). \quad (26) \end{aligned}$$

According to (26), the maximization problem of (25) goes to the maximization problems of (21) and (22), respectively. In the following, we only show the derivation of the update rules of $\boldsymbol{\alpha}_t$ for maximizing (21). The derivation of the update rules of $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_t$ are quite similar to that of the update rules of $\boldsymbol{\alpha}_t$.

Instead of maximizing the two joint probabilities, (21) and (22), we try to maximize their log-likelihoods, such that we aim at the following:

$$\begin{aligned} & (1-\lambda) \max\{\log P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Theta}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t)\} + \\ & \lambda \max\{\log P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t)\}, \quad (27) \end{aligned}$$

where $\log P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Theta}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t)$ can be represented as:

$$\begin{aligned} & \log P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Theta}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t) \\ =& 2 \sum_z \left(\log \Gamma \left(\sum_v (\varkappa_b) \right) - \log \Gamma \left(\sum_v \varkappa_a \right) \right) + \\ & 2 \sum_z \sum_v (\log \Gamma(\varkappa_a) - \log \Gamma(\varkappa_b)) + \\ & \sum_u \left(\log \Gamma(\sum_z (\varkappa_2)) - \log \Gamma(\sum_z \varkappa_1) \right) + \\ & \sum_u \sum_z (\log \Gamma(\varkappa_1) - \log \Gamma(\varkappa_2)), \quad (28) \end{aligned}$$

and $\log P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t)$ can be represented as:

$$\begin{aligned} & \log P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t) \\ =& 2 \sum_z \left(\log \Gamma \left(\sum_v (\varkappa_b) \right) - \log \Gamma \left(\sum_v \varkappa_a \right) \right) + \\ & 2 \sum_z \sum_v (\log \Gamma(\varkappa_a) - \log \Gamma(\varkappa_b)) + \\ & \sum_u \left(\log \Gamma(\sum_z (\varkappa_4)) - \log \Gamma(\sum_z \varkappa_3) \right) + \\ & \sum_u \sum_z (\log \Gamma(\varkappa_3) - \log \Gamma(\varkappa_4)). \quad (29) \end{aligned}$$

Applying the following two bounds from [38],

$$\begin{aligned} \log \Gamma(\hat{x}) - \log \Gamma(\hat{x}+n) \geq & \log \Gamma(x) - \log \Gamma(x+n) \\ & + (\Delta(x+n) - \Delta(x))(x - \hat{x}), \end{aligned}$$

and

$$\begin{aligned} \log \Gamma(\hat{x}+n) - \log \Gamma(\hat{x}) \geq & \log \Gamma(x+n) - \log \Gamma(x) \\ & + x(\Delta(x+n) - \Delta(x))(\log \hat{x} - \log x), \end{aligned}$$

into (27), (28) and (29), and assuming that $\hat{\alpha}_{t,z}$ is the optimal updating parameter in the next fixed-point iteration, we have:

$$\begin{aligned} & (1-\lambda) \log P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Theta}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t) + \\ & \lambda \log P(\mathbf{z}_t, \mathbf{d}_t \mid \mathbf{\Psi}_{t-1}, \mathbf{\Phi}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t) \geq L(\hat{\alpha}_{t,z}) \\ =& (1-\lambda) \left\{ \sum_u \left(\Delta(\sum_z \varkappa_1) - \Delta(\sum_z \varkappa_2) \right) (-\hat{\alpha}_{t,z}) + \right. \\ & \left. \alpha_{t,z} \sum_u (\Delta(\varkappa_a) - \Delta(\varkappa_b)) \log \hat{\alpha}_{t,z} \right\} + C, \end{aligned}$$

where C is a function not containing the variable $\hat{\alpha}_{t,z}$. Then to obtain the update rule for the optimal parameter $\hat{\alpha}_{t,z}$, we let:

$$\begin{aligned} \frac{\partial L(\hat{\alpha}_{t,z})}{\partial \hat{\alpha}_{t,z}} = & \frac{(1-\lambda) \alpha_{t,z} \sum_u (\Delta(\varkappa_1) - \Delta(\varkappa_2))}{\hat{\alpha}_{t,z}} \\ & - (1-\lambda) \sum_u \left(\Delta(\sum_z \varkappa_1) - \Delta(\sum_z \varkappa_2) \right) = 0 \end{aligned}$$

which results in the following update rule for $\alpha_{t,z}$ as:

$$\alpha_{t,z} \leftarrow \frac{(1-\lambda) \alpha_{t,z} \sum_u (\Delta(\varkappa_1) - \Delta(\varkappa_2))}{\sum_u (\Delta(\sum_z \varkappa_1) - \Delta(\sum_z \varkappa_2))}$$

Following the same derivation, we have the update rules for $\beta_{t,z}$ and $\gamma_{t,v}$ as:

$$\beta_{t,z} \leftarrow \frac{\lambda\beta_{t,z}\sum_u(\Delta(\varkappa_3)-\Delta(\varkappa_4))}{\sum_u(\Delta(\sum_z\varkappa_3)-\Delta(\sum_z\varkappa_4))}$$
$$\gamma_{t,v} \leftarrow \frac{\gamma_{t,v}\sum_z(\Delta(\varkappa_a)-\Delta(\varkappa_b))}{\sum_z(\Delta(\sum_v\varkappa_a)-\Delta(\sum_v\varkappa_b))}.$$

REFERENCES

[1] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *WWW*, 2010, pp. 499–508.
[2] G. Arru, D. F. Gurini, and F. Gasparetti, "Signal-based user recommendation on twitter," in *WWW*, 2013, pp. 941–944.
[3] J. Vosecky, K. W.-T. Leung, and W. Ng, "Collaborative personalized twitter search with topic-language models," in *SIGIR*, 2014, pp. 53–62.
[4] W. Chen, J. Wang, Y. Zhang, H. Yan, and X. Li, "User based aggregation for biterm topic model," in *ACL*, 2015, pp. 489–494.
[5] P. Xie, Y. Pei, Y. Xie, and E. Xing, "Mining user interests from personal photos," in *AAAI*, 2015, pp. 1896–1902.
[6] Y. Zhao, S. Liang, Z. Ren, J. Ma, E. Yilmaz, and M. de Rijke, "Explainable user clustering in short text streams," in *SIGIR*, 2016, pp. 155–164.
[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[8] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *UAI*, 2004, pp. 487–494.
[9] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *WWW*, 2013, pp. 1445–1455.
[10] S. Liang, Z. Ren, Y. Zhao, J. Ma, E. Yilmaz, and M. D. Rijke, "Inferring dynamic user interests in streams of short texts for user clustering," *ACM Trans. Inf. Syst.*, vol. 36, no. 1, pp. 10:1–10:37, 2017.
[11] S. Liang, Z. Ren, E. Yilmaz, and M. de Rijke, "Collaborative user clustering for short text streams," in *AAAI*, 2017, pp. 3504–3510.
[12] T. Hofmann, "Probabilistic latent semantic indexing," in *SIGIR*, 1999, pp. 50–57.
[13] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *ICML*, 2006, pp. 113–120.
[14] X. Wei, J. Sun, and X. Wang, "Dynamic mixture models for multiple time-series," in *IJCAI*, 2007, pp. 2909–2914.
[15] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *KDD*, 2006, pp. 424–433.
[16] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda, "Topic tracking model for analyzing consumer purchase behavior," in *IJCAI*, vol. 9, 2009, pp. 1427–1432.
[17] A. Ahmed, Q. Ho, C. H. Teo, J. Eisenstein, A. Smola, and E. Xing, "Online inference for the infinite topic-cluster model: Storylines from streaming text," in *AISTATS*, 2011, pp. 101–109.
[18] K. Caballero and R. Akella, "Dynamically modeling patient health state from electronic medical records: A time series approach," in *KDD*, 2015, pp. 69–78.
[19] S. Liang, E. Yilmaz, H. Shen, M. D. Rijke, and W. B. Croft, "Search result diversification in short text streams," *ACM Trans. Inf. Syst.*, vol. 36, no. 1, pp. 8:1–8:35, 2017.
[20] S. Liang, E. Yilmaz, and E. Kanoulas, "Dynamic clustering of streaming short documents," in *KDD*, 2016, pp. 995–1004.
[21] A. Bhadury, J. Chen, J. Zhu, and S. Liu, "Scaling up dynamic topic models," in *WWW*, 2016, pp. 381–390.
[22] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.
[23] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *SIGIR*, 2006, pp. 178–185.
[24] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2015.
[25] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: discovery and applications of usage patterns from web data," in *SIGKDD Explorations*. ACM, 2000, pp. 12–23.
[26] I. Li, Y. Tian, Q. Yang, and K. Wang, "Classification pruning for web-request prediction," in *WWW*. ACM, 2001.
[27] B. Mobasher, R. Cooley, and J. Srivastava, "Creating adaptive web sites through usage-based clustering of urls," in *IEEE KDEX workshop*. IEEE, 1999.
[28] G. Buscher, R. W. White, S. Dumais, and J. Huang, "Large-scale analysis of individual and task differences in search result page examination strategies," in *WSDM*. ACM, 2012, pp. 373–382.
[29] M. S. ElBamby and et al., "Content-aware user clustering and caching in wireless small cell networks," in *Wireless Communications Systems, International Symposium on*. IEEE, 2014, pp. 945–949.
[30] K. Balog and M. de Rijke, "Finding similar experts," in *SIGIR*. ACM, 2007, pp. 821–822.
[31] K. Hofmann, K. Balog, T. Bogers, and M. de Rijke, "Contextual factors for finding similar experts," *J. Am. Soc. Inf. Sci. Techn.*, vol. 61, no. 5, pp. 994–1014, May 2010.
[32] C. Van Gysel, M. de Rijke, and M. Worring, "Unsupervised, efficient and semantic expertise retrieval," in *WWW*. ACM, 2016, pp. 1069–1079.
[33] S. Liang, Z. Ren, and M. de Rijke, "Personalized search result diversification via structured learning," in *KDD*, 2014, pp. 751–760.
[34] S. Liang, F. Cai, Z. Ren, and M. de Rijke, "Efficient structured learning for personalized diversification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 11, pp. 2958–2973, 2016.
[35] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda, "Online multiscale dynamic topic models," in *KDD*. ACM, 2010, pp. 663–672.
[36] J. S. Liu, "The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem," *J. Am. Stat. Assoc.*, vol. 89, no. 427, pp. 958–966, 1994.
[37] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *PNAS*, vol. 101, pp. 5228–5235, 2004.
[38] T. Minka, "Estimating a dirichlet distribution," 2000.
[39] Y. Ren, E. B. Fox, and A. Bruce, "Clustering correlated, sparse data streams to estimate a localized housing price index," *Annals of Applied Statistics*, vol. 11, no. 2, pp. 808–839, 2017.
[40] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, pp. 281–297.
[41] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," in *KDD*. ACM, 2014, pp. 233–242.
[42] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *UAI*, 2004, pp. 487–494.
[43] C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models," in *Uncertainty in Artificial Intelligence*. AUAI Press, 2008, pp. 579–586.
[44] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press, 2008.
[45] I. Bordino, C. Castillo, D. Donato, and A. Gionis, "Query similarity by projecting the query-flow graph," in *SIGIR*, 2010, pp. 515–522.
[46] F. Cai, S. Liang, and M. de Rijke, "Prefix-adaptive and time-sensitive personalized query auto completion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 9, pp. 2452–2466, 2016.

Shangsong Liang received a B.Sc. and M.Sc. degree from the Northwest A&F University, Xi'an, China, in 2008 and 2011, respectively, and a Ph.D. degree from the University of Amsterdam in 2014, all related to computer science. His research interest is information retrieval and data mining. Dr. Liang has published at SIGIR, KDD, WWW, WSDM, CIKM, AAAI and in ACM transactions on Information Systems.

Emine Yilmaz is an Associate Professor at the University College London and a faculty fellow of the Alan Turing Institute. She also works as a research consultant for Microsoft Research Cambridge. She is the recipient of the Karen Sparck Jones 2015 Award, and she received the Google Faculty Research Award in 2014/2015. Dr. Yilmaz's current research interests include information retrieval, data mining and applications of information theory, statistics and machine learning. She is serving as the PC Chair for ACM SIGIR 2018 and ACM ICTIR 2017.

Evangelos Kanoulas is an Assistant Professor at the University of Amsterdam. His expertise lies in the fields of information retrieval and text mining. He has worked as a (visiting) research scientist in two of the leading companies in search technology, Google and Microsoft. In 2010 he was awarded the Marie Curie Fellowship and worked as a postdoctoral research scientist at the University of Sheffield. Dr. Kanoulas has extensively published his work in top-tier conferences, including SIGIR, CIKM and ECIR. He is the recipient of a Google Faculty Research Award in 2015/2016.