

Personalized Time-Aware Tweets Summarization

Zhaochun Ren
ISLA, University of Amsterdam
Amsterdam, The Netherlands
z.ren@uva.nl

Shangsong Liang
ISLA, University of Amsterdam
Amsterdam, The Netherlands
s.liang@uva.nl

Edgar Meij
Yahoo! Research
Barcelona, Spain
emeij@yahoo-inc.com

Maarten de Rijke
ISLA, University of Amsterdam
Amsterdam, The Netherlands
derijke@uva.nl

ABSTRACT

We focus on the problem of selecting meaningful tweets given a user’s interests; the dynamic nature of user interests, the sheer volume, and the sparseness of individual messages make this an challenging problem. Specifically, we consider the task of time-aware tweets summarization, based on a user’s history and collaborative social influences from “social circles.” We propose a time-aware user behavior model, the Tweet Propagation Model (TPM), in which we infer dynamic probabilistic distributions over interests and topics. We then explicitly consider novelty, coverage, and diversity to arrive at an iterative optimization algorithm for selecting tweets. Experimental results validate the effectiveness of our personalized time-aware tweets summarization method based on TPM.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Information filtering

Keywords

Twitter, tweets summarization, data enrichment, topic modeling

1. INTRODUCTION

Twitter has amassed over half a billion users in 2012, who produce (“tweet”) over 300 million tweets per day.¹ Twitter users can subscribe to updates from other users by following them, essentially forming a unidirectional friend relationship. Moreover, tweets can be “retweeted,” basically copying a tweet posted by another user to one’s own timeline. From an information retrieval point of view, the sheer volume of users and tweets presents interesting challenges. On the one hand, interesting, relevant, or meaningful tweets can easily be missed due to a large number of followed users. On the other hand, users may miss interesting tweets when none of the users they follow retweet an interesting piece of information.

¹<http://blog.twitter.com/2012/03/twitter-turns-six.html>.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR’13, July 28–August 1, 2013, Dublin, Ireland.

Copyright 2013 ACM 978-1-4503-2034-4/13/07 ...\$15.00.

One task that is aimed at addressing this dual problem is *tweets summarization* [6]: to extract a group of representative tweets from a set of tweets. The task is similar to tweet recommendation, but tweets summarization pays more attention to the *quality* of selected results, including notions such as representativeness and diversity. So far, tweets summarization methods are typically query and user-independent. How to adapt tweets summarization to a specific user is still a topic of ongoing research [4, 5, 7, 22, 31, 36]. Current methods, whether personalized or not, also neglect to explicitly model the temporal nature of the microblogging environment; time-awareness is a key feature of Twitter in general and tweets summarization in particular.

We put forward a model for personalized, time-aware tweets summarization (TaTS). We investigate three key aspects of tweets summarization: (a) novelty, preventing near-duplicate tweets to be included, (b) coverage, so as to be representative to candidate tweets, (c) diversity, covering as many aspects as possible. When working with Twitter data, several methodological challenges arise. In order to perform effective tweets summarization, we require a notion of a user’s interest. Most Twitter users, however, mostly *consume* information without *producing* a lot of information. That is, they rarely post tweets of their own [22]. Hence, in order to infer a user’s interest in a robust manner, we need to use other signals than just the user’s tweets. To address the issue, we incorporate intuitions from the field of collaborative filtering and base our estimation of a person’s interest on those of their friends on Twitter, following [5]. We assume that for each user there exist one or more “social circles,” in which three or more users follow each other and form cliques. We find that people are usually connected to specific communities and assume that each user’s behavior on Twitter is affected by: (a) a user’s private taste, (b) a collaborative effect from social circles, and (c) a bursty component, reflecting current events.

Clearly, a user’s interest can change over time. Topic modeling has proven effective for topic detection and user behavior modeling on Twitter [8, 25, 32]. As a dynamic extension of the author-topic model [26], our proposed Tweet Propagation Model (TPM) aims to track both a user’s interests and any topic drift arising with the passing of time. Based on “social circles”, TPM derives the user’s interest from a dirichlet mixture over interests of someone who share “social circles.” It does so by inferring distributions over topics and interests that change over time. Following existing topic modeling approaches for Twitter [8, 37], we extend TPM and classify the topics as (a) personal topics, (b) common topics, or (c) bursty topics. Gibbs Expectation Maximization (EM) sampling [29] is used to infer the posterior probabilities and to estimate the value of hyperparameters in our topic models. After inferring the probabilities of

each tweet, we employ an iterative algorithm to optimize the tweet selection procedure, considering coverage, novelty, and diversity.

Our contributions in this paper are as follows. (1) We propose the task of personalized time-aware tweets summarization, selecting personalized meaningful tweets from a collection of tweets. (2) We leverage a user’s “collaborative influence” in order to derive the user’s interests. (3) We introduce a tweet propagation model to address the potential drift in a user’s interests as well as topics over time. (4) We employ a tweet selection algorithm that jointly optimizes for coverage, diversity, and novelty.

The rest of this paper is organized as follows. We introduce related work in Section 2. Our problem formulation is detailed in Section 3. Our strategy for tweet summary generation, is described in Section 4. Section 5 details our experimental setup and Section 6 presents and discusses the experimental results and Section 7 concludes the paper.

2. RELATED WORK

Our approach builds on earlier work in tweets summarization, tweet recommendation and topic modeling.

2.1 Tweets Summarization

Several publications have focused on tweets summarization: the task of selecting a list of meaningful tweets that are most representative for some topic. Most work in the literature concerns tweets as basic constituents to compose a summary. Some authors bring feature-based or graph-based summarization technologies to bear on this task [6, 27], while other methods use a term-frequency based method [28] or a strategy based on mutual reinforcement between users’ influence and qualifications of tweets [9]. Recently, time-aware summarization has been studied by several authors, often in the form of timeline generation on Twitter. Chakrabarti and Punera [4] separate topic related tweets into various periods as an event evolution map, and generate an update-summarization result. Yan et al. [33] propose an evolutionary timeline summarization strategy based on dynamic programming. Evolutionary summarization approaches segment post streams into event chains and select tweets from various chains to generate a tweet summary; Nichols et al. [21] propose an effective method to separate timelines using Twitter. To the best of our knowledge, existing work on tweets summarization focuses on the extraction of representative tweets for specific topics, without considering personalization.

Other work integrates the task of selecting tweets with other web documents: Yang et al. [35] use mutual reinforcement to train both the selection of related web documents and tweets via a single graph factor model. Zhao et al. [37] extract representative keywords from tweets based on a topic model. Tweet ranking has also attracted attention: Weng et al. [31] proposed a graph-based ranking strategy for ranking tweets based on the author-topic model.

2.2 Collaborative Tweet Recommendations

In recent years, collaborative filtering on Twitter has attracted increased attention. Yang et al. [34] address recommendation and link prediction tasks based on a joint-propagation model, FTP, between social friendship and interests. Ye et al. [36] propose a generative model to describe users’ behavior, given influences from social communities, for recommendation [18, 19]. To track social influence of users in a social network, Xu et al. [32] propose a graphical mixture model to describe user’s behavior in posting tweets and analyze the original topic domain for a specific proposed tweet. Chen et al. [5] propose a collaborative filtering method to generate personalized recommendations in Twitter through a collaborative ranking procedure. Similarly, Pennacchiotti et al. [22]

Table 1: Glossary.

Symbol	Description
K	number of topics
U	number of users
V	the size of the vocabulary
T	number of time periods
\mathcal{D}_t	candidate tweets at time t
D_t	number of candidate tweets at time t , i.e., $ \mathcal{D}_t $
u	user u on Twitter, $u \in \mathcal{U}$
$\mathcal{C}_{u,t}$	social circle for user u at t
$\mathcal{D}_{u,t}$	tweets posted by u at time t
$D_{u,t}$	number of tweets posted by u at time t , i.e., $ \mathcal{D}_{u,t} $
$F_{u,t}$	number of friends of u at time t
$C_{u,t}$	number of social circles around u at time t
d_t	tweet published at time t , $d_t \in \mathcal{D}_t$
w	token/word present in some tweet, $w \in \mathcal{W}$
z_t	latent topic at t time, $z_t \in \mathcal{Z}_t$
$c_{u,t}$	social circle around u at time t , $c_{u,t} \in \mathcal{C}_{u,t}$
$\theta_{u,t}$	distribution of u ’s interests over topics at time t
ϑ_t	distribution of topics within a tweet at time t
ϕ_t	distribution of words over topics at time t
Z	classification of individual topics in θ or ϑ
β, α, σ, r	hyper-parameters in TPM
N	maximum number of tweets returned
$\lambda_{u,c,t}$	weight of social circle c for user u at t

propose a method to recommend “novel” tweets to users by following users’ interests and using the tweet content. However, many of these methods ignore the dynamic nature of the problem; with the change of time, user interests may also change.

2.3 Topic Models

Topic models [2, 12] are employed to reduce the high dimensionality of terms appearing in text into low-dimensional, “latent” topics. Ever since Hofmann [12] presented probabilistic latent semantic indexing (pLSI), many extensions have been proposed. In latent Dirichlet allocation (LDA, [2]) each document is generated by choosing a distribution over topics and then each word in the document is chosen from a selected topic. To handle users’ connections with particular documents and topics, the author-topic model has been proposed [26]. However, for data with topic evolution the underlying “bag of words” representation may be insufficient. To analyze topic evolution, other models have been proposed, such as the Dynamic Topic Model [1], Dynamic Mixture Models [30] and the Topic Tracking Model [13]. Topic models have not yet been considered very frequently in the setting of Twitter. Twitter-LDA is an interesting exception; it classifies latent topics into “background” topic and “personal” topics [37], while an extension of Twitter-LDA has been proved to be effective in burst detection [8].

Our work is different from the related work mentioned above in the following important ways: (1) our work focuses on personalized and time-aware tweets summarization and (2) we propose a tweet propagation model by jointly modeling time-aware propagation and collaborative filtering from “social circles,” which is different from existing topic models.

3. PROBLEM FORMULATION

Before introducing our method for time-aware tweets summarization, we introduce our notation and key concepts. Table 1 lists the notation we use. Given two users u_i and u_j on Twitter, there are two main reasons for u_i and u_j to follow each other: either because

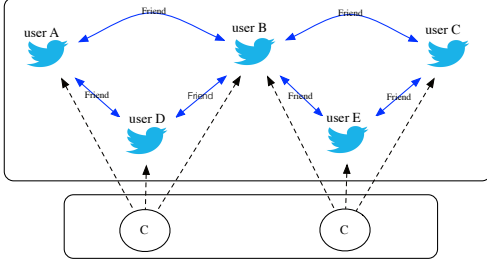


Figure 1: Example of social circles on Twitter: there are two social circles (indicated using the ‘c’) among the five users in this graph, where each pair of vertices in each social circle is connected through the “friend” relationship.

they have similar interests or they have some relationship outside Twitter [32]. If two users u_i and u_j follow each other, we define them to be *friends* on Twitter. Given this definition, we define a *social circle* around a user u to be a set of friends of u such that every pair of users in this set is in the *friend* relation. See Figure 1 for a schematic representation.

Similar to the author-topic model [26], we assume that each Twitter user’s interests are represented by a multinomial distribution $\theta_{u,t}$, which may, however, change over time. That is, the time-aware interests of user u are represented as a multinomial distribution $\theta_{u,t}$ over topics, where each topic is represented as a probabilistic distribution over words [2]. Formally, we have $\theta_{u,t} = \{\theta_{u,t,z_1}, \dots, \theta_{u,t,z_K}\}$, where θ_{u,t,z_i} denotes the distribution of topic z_i for user u at time t .

We further assume that each tweet can be represented as a probabilistic distribution over topics. To cater for the phenomenon of user interests changing over time, we assume that topic distributions are dynamic and may differ between time periods. Given a user u , we split the topic set Z_t at time t into three classes: $Z_t = Z_t^u \cup Z_t^{com} \cup Z_t^B$: there exist “private” topics Z_t^u that solely depend on the user, there are common topics Z_t^{com} that are influenced by friends from shared social circles, and there are topics from event-related, bursty sources, Z_t^B . The latter type of topic will typically transfer from initially being observed at time t into Z_t^{com} at some later time t' .

The dynamic interests of user u at time t , reflected by $\theta_{u,t}$, evolve in different ways depending on the class that a topic $z_t \in Z_t$ belongs to. For each user, $\theta_{u,t}$ is affected by the following three classes.

- (a) If $z_t \in Z_t^u$ is a “private” topic, then $\theta_{u,t,z}^u$ only depends on $\theta_{u,t-1,z}^u$ at time $t-1$.
- (b) If $z_t \in Z_t^{com}$ then the topic is dependent on friends in the user’s social circle(s). $\theta_{u,t,z}^{com}$ is computed from the collaborative effect $\theta_{u_i,t-1}^{com}$ at time $t-1$ from the social circles $\{u_i | u_i \in C_{u,t-1}\}$.
- (c) If $z_t \in Z_t^B$ is a “burst” topic, $\theta_{u_i,z,t}^B$ is generated according to a distribution of “burst” words in u_i ’s tweets at time t .

Typically, traditional summarization does not cover the evolution of a specific event. Given a split of a user’s history into time periods, the task of time-aware tweets summarization is to select the most representative tweets for each time period, covering the whole event evolution on a timeline. More precisely, given a set of tweets \mathcal{D} , a set of time periods T , and a maximum number of tweets per period, N , time-aware tweets summarization aims to extract multiple sets of tweets RT_t ($1 \leq t \leq T$) from \mathcal{D} , where for each time period t , $RT_t = \{d_{t,x_1}, d_{t,x_2}, \dots, d_{t,x_N}\}$ is a set of representative tweets that summarize the period. Furthermore, *personalized*

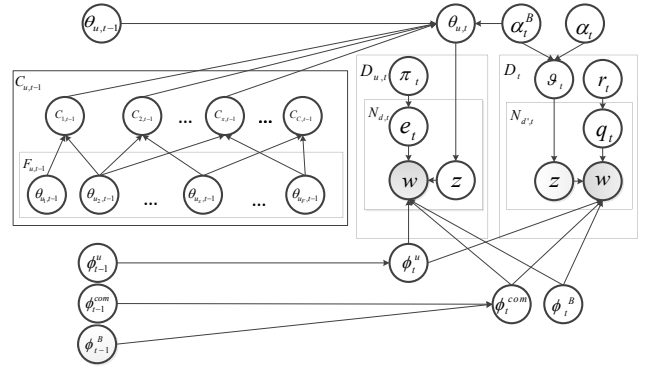


Figure 2: Graphical representation of TPM.

time-aware tweets summarization is defined similar to time-aware tweets summarization, but in this case the tweets selected for inclusion in RT_t need to be relevant based on u ’s interests θ_u at time t .

4. METHOD

In this section, we detail our tweets summarization method, including the required methods for joint user-tweets topic modeling, inference and parameter estimation. As input, our method has probabilistic distributions from topic modeling. The output is the time-aware tweets summary, i.e., a selection of tweets (per period) satisfying the user’s interest.

4.1 Topic Modeling: Tweets Propagation

We start by proposing the tweets propagation model (TPM) to jointly track dynamic user’s interests and topics. The interests of a user u are assumed to be reflected by a multinomial distribution $\theta_{u,t}$ over topics. We assume that the distribution of topics ϕ_t over words follows a dynamic propagation process with changes over time. Figure 2 provides a graphical overview of TPM.

In the graphical structure of TPM, we see a number of ingredients. Among the variables related to user u in the graph, z , θ and γ are random variables and w is the observed variable. In the candidate tweets part, ϑ , z and σ are random variables; $D_{u,t}$ and D_t indicate the number of variables in the model. As usual, directed arrows in a graphical model indicate the dependency between two variables; the variables $c_{u,t-1}$ depend on variables $\{\theta_{u_i,t-1} | u_i \in C_{u,t-1}\}$. The variables ϕ_t^{com} and ϕ_t^u depend on variables $\{\phi_{t-1}^{com}, \phi_{t-1}^B\}$ and ϕ_{t-1}^u , respectively.

Now, let us give a more detailed technical account of our model. Around user u , there exist multiple social circles. For each social circle $c_{u,t}$ in time period t , there is a random parameter $\lambda_{c_{u,t}}$ indicating the importance of $c_{u,t}$ to u at t . User u ’s interests $\theta_{u,t}$ are composed of three parts: the personal aspect, the common topic aspect and the bursty aspect, i.e., $\theta_{u,t} = \{\theta_{u,t}^{com}, \theta_{u,t}^u, \theta_{u,t}^B\}$, where the common topics are not only influenced by the user’s social circles, but also by his own previous interests. Therefore, we use a Dirichlet distribution to derive the probability of $\theta_{u,t}^{com}$ over $x_{u,t}^{com}$ as:

$$x_{u,t}^{com} = \alpha_{u,t} \theta_{u,t-1}^{com+B} + (1 - \alpha_{u,t}) \sum_{c_i \in C_{u,t-1}} \lambda_{c_i} \theta_{c_i,t-1}^{com+B} \quad (1)$$

where $\theta_{u,t-1}^{com+B}$ refers to the set $\{\theta_{u,t-1}^{com}, \theta_{u,t-1}^B\}$ at period $t-1$, which reflects user u ’s interests for common and burst topics at time $t-1$, and $\theta_{c_i,t-1}^{com+B}$ refers to the set $\{\theta_{c_i,t-1}^{com}, \theta_{c_i,t-1}^B\}$ at period $t-1$. The hyperparameter $\alpha_{u,t}$ indicates the weight of $\theta_{u,t-1}^{com+B}$

1. For each topic $z, z \in Z_t^{com} \cup Z_t^B \cup Z_t^u$:
 - Draw $\phi_t^B \sim Dir(\beta_t^B)$;
 - Draw $\phi_t^{com} \sim Dir(\beta_t^{com} \{\phi_{t-1}^{com}, \phi_{t-1}^B\})$;
 - Draw $\phi_t^u \sim Dir(\beta_t^u \phi_{t-1}^u)$
2. For each candidate tweet $d_t \in \mathcal{D}_t$:
 - Draw $\vartheta_t \sim Dir(\alpha_t, \alpha_t^B)$; $r_t \sim Dir(\gamma_t)$;
 - For each word w in d_t
 - Draw $q \in Multi(r)$; $z_w \sim Multi(\vartheta_t)$;
 - * if $q = 0$: Draw $w \sim Multi(\phi_{z,t}^{com})$;
 - * if $q = 1$: Draw $w \sim Multi(\phi_{z,t}^u)$;
 - * if $q = 2$: Draw $w \sim Multi(\phi_t^B)$;
3. For user $u, u \in \mathcal{U}$:
 - Draw $\theta_{u,t} \sim Dir(\{x_{u,t}^u, x_{u,t}^{com}, \alpha_t^B\})$;
 - Draw $\pi_t \sim Dir(\sigma_t)$;
 - For each word $w \in d_{u,t}$, where $d_{u,t} \in D_{u,t}$:
 - Draw $e \sim Multi(\pi)$; $z_{w,t} \sim Multi(\theta_{u,t})$
 - * if $e = 0$: Draw $w \sim Multi(\phi_{z,t}^{com})$;
 - * if $e = 1$: Draw $w \sim Multi(\phi_{z,t}^u)$;
 - * if $e = 2$: Draw $w \sim Multi(\phi_t^B)$;

Figure 3: Generative process for the TPM model.

in Equation 1 that we use to calculate $\theta_{u,t}^{com+B}$. Here, the value of $\theta_{c_i,t-1}^{com+B}$ is equal to $\frac{1}{|C_{u,t}|} \sum \theta_{u_i,t-1}$, where $u_i \in C_{u,t-1}$.

For private topical aspects $\theta_{u,t}^u$, we use a Dirichlet distribution over $x_t^u = \theta_{u,t-1}^u$ that is derived from values in period $t-1$. For bursty topics in period t , we only focus on those ‘‘burst’’ words that have a high term frequency within period t . Similar to [32], we define a keyword to be ‘‘bursty’’ if its frequency $n_{w,t}$ at time t is above a threshold value. We derive $\theta_{u,t}^B$ from a Dirichlet distribution over the hyperparameter α_t^B .

For a tweet in \mathcal{D}_t that is posted during time period t , a probabilistic distribution ϑ_t over topics $Z_t = Z_t^u \cup Z_t^{com} \cup Z_t^B$ is derived from a Dirichlet distribution over the hyperparameter α_t .

For each word w in tweet $d_t, d_t \in \{\mathcal{D}_{u,t}, \mathcal{D}_t\}$ proposed during period t , we assign a specific topic z from u ’s interests $\theta_{u,t}$ or distribution ϑ_t for candidate documents. For topic aspects $z (z \in Z_t^{com} \cup Z_t^B \cup Z_t^u)$, we introduce three kinds of multinomial distribution ϕ_t^{com}, ϕ_t^u and ϕ_t^B to reflect the probability over Z^{com}, Z^B and Z^u , respectively. Based on [13, 30], we assume that the common and personal topic propagations follow a Dirichlet distribution over the value from the previous interval’s distributions, with a weighted prior $\beta_t = \{\beta_t^{com}, \beta_t^u\}$: for common topics $z \in Z_t^{com}$, we use the Dirichlet distribution to infer from $\{\phi_{t-1}^{com}, \phi_{t-1}^B\}$; for private topics $z \in Z_t^u$, ϕ_t^u is derived from ϕ_{t-1}^u .

This concludes the technical account of the graphical model depicted in Figure 2. After computing the models for period t for all users in \mathcal{U} , we update the edge weights for the social circles ($\lambda_{u,c_i,t}$), using related users’ interests θ and current social circles. Inference for our topic modeling process will then move on to period $t+1$. The generative process for the TPM model at time interval $t, 0 < t < T$, is described in Figure 3.

4.2 Inference and Parameter Estimation

Sampling-based methods for LDA rarely include methods for optimizing hyper-parameters. In the TPM model, since $\alpha_{u,t}$ and $\beta_{z,t}^{f_l}$ indicate the weight of the results for period $t-1$ for computations for period t , it is necessary to find an optimized process for hyper-parameters $\alpha_{u,t}$ and $\beta_{z,t}^{f_l}$ during our posterior inference. Therefore, unlike many previous dynamic topic models, to infer weighted priors we use a Gibbs EM algorithm [29] to handle the approximate posterior inference step. For user u at time interval t , we first jointly sample topic z_i and parameter q_i from the i th word in tweet $d (d \in D_{u,t})$ over other variables. So for u ’s tweets we obtain:

$$p(e_i = l, z_i = z | \mathcal{W}, e_{-i}, Z_{-i}, x_{u,t}, \sigma, \beta_t^u) \propto \frac{n_{d,l,-i}^u + \sigma}{n_{d,-i}^u + 3\sigma} \cdot \frac{n_{d,z,-i}^u + x_{u,z,t}^{f_l}}{\sum_{z' \in Z^{f_l}} (n_{d,z',-i}^u + x_{u,z',t}^{f_l})} \cdot \frac{n_{w,z,-i}^u + \beta_t^{f_l}}{\sum_{w' \in N_{u,t}} n_{w',z,-i}^u + N_{u,t} \beta_t^{f_l}}, \quad (2)$$

where l indicates the possible values of variable e for the i th word in tweet p , and the f_l indicate the corresponding kind of topics when $e_i = l$. For private and common topics in u , i.e., $l = 0, 1$, in Equation 2, $n_{d,l,-i}^u$ indicates the number of times that words in d are assigned to label l except for the i th word, whereas $n_{d,-i}^u$ indicates the sum of $n_{d,l,-i}^u$ for all values of l . Furthermore, $n_{d,z,-i}^u$ is the number of times that tweet d is assigned to topic z excluding the i th word in d , whereas $n_{w,z,-i}^u$ indicates the number of times that word w is assigned by topic z excluding the i th word. According to Figure 3, if $e_i = 2$, we are dealing with a ‘‘bursty’’ topic, so the vocabulary only refers to the set of ‘‘bursty’’ keywords in $\{\mathcal{D}_{u,t}, \mathcal{D}_t\}$, then $x_{u,t}^B$ in Eq. 2 equals to α_t^B .

For the process of sampling candidate tweets from \mathcal{D}_t , we have a similar procedure, as follows:

$$p(q_i = l, z_i = z | \mathcal{W}, d_{-i}, Z_{-i}, \alpha_t, \gamma, \beta_t^u) \propto \frac{n_{d,l,-i}^u + \gamma}{n_{d,-i}^u + 3\gamma} \cdot \frac{n_{d,z,-i}^u + \alpha_t^{f_l}}{\sum_{z' \in Z^{f_l}} n_{d,z',-i}^u + Z^{f_l} \alpha_t^{f_l}} \cdot \frac{n_{w,z,-i}^u + \beta_t^{f_l}}{\sum_{w' \in N_t} n_{w',z,-i}^u + N_{u,t} \beta_t^{f_l}}. \quad (3)$$

Meanwhile, every time after sampling for $p(e_i = l, z_i = z)$ and $p(q_i = l, z_i = z)$, we optimize $\hat{\alpha}_{u,t}$ and $\hat{\beta}_{z,t,t-1}^{f_l}$ by maximizing the likelihood posterior distribution

$$p(W | \Phi_{t-1}, x_{u,t-1}, \alpha^B, \beta_t, \sigma, \gamma),$$

so we get

$$\hat{\alpha}_{u,t} = \hat{\alpha}_{u,t} \cdot \frac{\sum_{z \in Z_t^{com}} (\theta_{u,t-1,z} - \sum_{c_i \in C_{u,t-1}} \lambda_{\theta_{c_i,t-1}}) A_{u,z,t}}{\Psi(n_{u,t}^{com} + \alpha_{u,t}) - \Psi(\alpha_{u,t})} \quad (4)$$

and

$$\hat{\beta}_{z,t}^{f_l} = \hat{\beta}_{z,t}^{f_l} \cdot \frac{\sum_{w \in N_t} \phi_{t-1,w}^{f_l} (\Psi(n_{w,z,t}^{f_l} + y_{t,t-1}^{w,z}) - \Psi(y_{t,t-1}^{w,z}))}{\Psi(n_{z,t}^{f_l} + \beta_{z,t}) - \Psi(\beta_{z,t})} \quad (5)$$

where $\Psi(x)$ is defined by $\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$, $A_{u,z,t}$ refers to

$$\Psi(n_{u,z,t}^{com} + x_{u,z,t}^{com}) - \Psi(x_{u,z,t}^{com}),$$

and $y_{t,t-1}^{w,z}$ is defined as $\beta_t^{com} \phi_{t-1}^{com+B}$.

Algorithm 1 summarizes the Gibbs EM sampling inference based on the equations that we have just derived. During the Gibbs EM

Algorithm 1: Gibbs EM Sampling Process during period t

Input: $\beta_t, \beta^B, \alpha^B, \alpha_t, X_{u,t}, \Phi_{t-1}^f, d_t, \mathcal{U}, \mathcal{D}_t$ and R
Output: $\hat{\beta}_t^{f_l}, \hat{\alpha}_t, \langle e, z \rangle$ and $\langle q, z \rangle$
Initialize $\beta_t, \beta^B, \alpha^B, \alpha_t$; Topic assignment for all words
for $u \in \mathcal{U}$ **do**
 $r = 0$;
 for $r < R$ **do**
 E-Step:
 for $d = 1$ to $D_{u,t}$ **do**
 for $i = 1$ to N_d **do**
 Draw $\langle e_i, z_i \rangle$ from Eq. 2
 Update $n_{e,0,i}^{u,t}, n_{e,z,i}^{u,t}$ and $n_{w,z,i}^{u,t}$
 end
 end
 for $d = 1$ to D_t **do**
 for $i = 1$ to N_d **do**
 Draw $\langle q_i, z_i \rangle$ from Eq. 3
 Update $n_{q,l,i}^t, n_{d,z,i}^t$ and $n_{w,z,i}^t$;
 end
 end
 M-Step:
 Calculate $\theta_{u,t}^{f_l}, \phi_{w,t}^{f_l}, \vartheta_{d,t}^{f_l}$ and λ_{u,c_i} from Eq. 6, 8;
 Maximize $\hat{\alpha}_{u,t}^{(r)}$ and $\hat{\beta}_{z,t}^{f_l,(r)}$ from Eq. 4, 5;
 $r = r + 1$ and go to **E-Step**;
 end
end

sampling process, we estimate the parameters of user u 's interests $\theta_{u,z,t}^{e=l}, \theta_{u,z,t}^{q=l}$, the probability of topics over candidate tweets $\vartheta_{d,z,t}^{q=l}$, topic distributions over words $\phi_{w,z,t}^{f_l}$ and $\{\pi_{d,l,t}, r_{d,l,t}\}$ as follows:

$$\begin{aligned}
\theta_{u,z,t}^{f_l} &= \frac{n_z^{u,t} + x_{u,z,t}^{f_l}}{\sum_{z \in Z^{f_l}} n_z^{u,t} + x_{u,z',t}^{f_l}} \\
\vartheta_{d,z,t}^{f_l} &= \frac{n_{d,z,t} + \alpha_{z,t}}{\sum_{z \in Z^{f_l}} n_{d,z',t} + \alpha_{z',t}} \\
\phi_{w,z,t}^{f_l} &= \frac{n_{w,z,t} + \beta_{w,t}^{f_l}}{\sum_{z \in Z^{f_l}} n_{w,z,t} + \beta_{w,t}^{f_l}} \\
\pi_{d,l,t} &= \frac{n_{d,l}^{u,t} + \sigma}{n_d^{u,t} + 3\sigma} \\
r_{d,l,t} &= \frac{n_{d,l}^t + \gamma}{n_d^t + 3\gamma}
\end{aligned} \tag{6}$$

To compute the weight λ_{u,c_i} , we use a Markov random walk strategy, which calculates saliency of a social circle based on "voting" from others. Since each social circle can be considered as a set of users, an interest distribution $\theta_{c_i,t-1}^{com+B}$ for each social circle c_i can be computed as $\sum_{u' \in c_i} \theta_{u',t-1}^{com+B}$. Thus we compute a $\theta_{u,t}^{com+B}$ -based similarity matrix $SIM_{i,j}^{u,t}$ among different social circles, where each item $SIM_{i,j}^{u,t}$ is computed based on the divergence between two items:

$$div(\theta_{c_i}, \theta_{c_j} | \theta_u) = \sum_{z \in Z} \left| \theta_{c_i,z} \ln \frac{\theta_{c_i,z}}{\theta_{u,z}} - \theta_{c_j,z} \ln \frac{\theta_{c_j,z}}{\theta_{u,z}} \right| \tag{7}$$

We calculate the saliency of c_i after normalizing SIM into \widehat{SIM} :

$$\lambda_{u,c_i,t} = \mu \sum_{i \neq j} \widehat{sim}(\theta_{c_i,t}, \theta_{c_j,t} | \theta_{u,t}) \cdot \lambda_{u,c_i,t} + \frac{(1-\mu)}{|C_{u,t}|} \tag{8}$$

4.3 Time-Aware Summarization

After Gibbs EM sampling, for each candidate tweet d_t at time t , we have two parametric distributions ϑ_t and ϕ_t that reflect the topic-tweet distribution and the word-topic distribution, respectively. I.e., $P(z_t | d_t) = \theta_{z_t, d_t}$ and $P(w | z_t) = \phi_{z_t, w}$. For user u at time t , we now derive the distribution of interests over topics $\theta_{u,t}$, i.e., $P(z_t | u, t)$.

Given the distribution $\theta_{u,t}$, one intuitive way to get the most meaningful tweets is to extract the most similar tweets with $\theta_{u,t}$ from among a candidate set \mathcal{D}_t . However, a high-degree relevance in latent topic distributions cannot be taken as the only criterion in our tweet selection. Thus after extracting a set of relevant tweets \mathcal{R}_t from \mathcal{D}_t , there are three key requirements for an ideal summary [16] that we need to consider in generating a tweet summary: *novelty*, the *coverage* and the *diversity*.

Novelty calculates the semantic divergence between the currently selected set $RT_{u,t}$ and the results in previous time periods $RT_{t'}$. Our intention is to make the current results as different as possible from previous results as much as possible. Therefore, we have:

$$\mathcal{L}_N(RT_t | RT_{t'}) = \sum_{p \in RT} \min_{p' \in RT_{t'}} (div(\vartheta_p, \vartheta_{p'} | \theta_{u,t})) \tag{9}$$

where the divergence $div(\vartheta_p, \vartheta_{p'} | \theta_{u,t})$ between ϑ_p and $\vartheta_{p'}$ are calculated based on Equation 7.

Furthermore, a tweet summary should contain important aspects from all related tweets and minimize the information loss with the set of all candidate tweets. Thus, given $\theta_{u,z,t}$, the *coverage* between RT and \mathcal{D}_t is calculated as follows:

$$\mathcal{L}_C(RT | \mathcal{D}_t) = \sum_{d \in RT} e^{-\min_z \sum_{d' \in \mathcal{D}_t} div(\vartheta_{d,z}, \vartheta_{d',z} | \theta_{u,z,t})} \tag{10}$$

where the divergence $div(\vartheta_{d,z}, \vartheta_{d',z} | \theta_{u,z,t})$ is calculated as follows:

$$div(\vartheta_{d,z}, \vartheta_{d',z} | \theta_{u,z,t}) = \left| \vartheta_{d,z} \ln \frac{\vartheta_{d,z}}{\theta_{u,z,t}} - \vartheta_{d',z} \ln \frac{\vartheta_{d',z}}{\theta_{u,z,t}} \right| \tag{11}$$

Diversity calculates the information divergence among all tweets within the current candidate result set. Ideally, the tweet summary results have the largest possible difference in topic distributions with each other. The equation is as follows:

$$\mathcal{L}_D(RT) = \sum_{w, w' \in RT} \max_z div(\phi_{w,z,t}, \phi_{w',z,t} | \prod_{d \in \mathcal{D}_t} \vartheta_{d,z}) \tag{12}$$

where we compute the divergence $div(\phi_{w,z,t}, \phi_{w',z,t} | \prod_{d \in \mathcal{D}_t} \vartheta_{d,z})$ in the same way as Equation 11.

The exact process for generating $RT_{u,t}$ given user u is shown in Algorithm 2. Illuminated by a previous work [33], an iterative optimization algorithm is used to select the set $RT_{u,t}$. During each iteration n , we extract tweet d_x such that $d_x \in \mathcal{R}_t \cap d_x \notin RT_{u,t}$ to substitute $d_y \in RT_{u,t}^{(n)}$ when the saliency gain $\mathcal{S}((RT_{u,t} -) \cup d_x) - \mathcal{S}(RT_{u,t})$ gets a maximum value. The algorithm will converge when $\mathcal{S}(RT_{u,t})$ reaches its maximum value.

5. EXPERIMENTAL SETUP

For our experiments we employ a Twitter dataset that includes both social relations and tweets: we crawl tweets via the Twitter

Algorithm 2: Iterative Process for $RT_{u,t}$ Generation.

Input : $\mathcal{D}_t, RT_{u,t'}, \theta_{u,t}, \phi_t, N;$ **Output:** $RT_{u,t};$ Calculate Kullback-Leibler divergence $KL(\vartheta_{d,t}, \theta_{u,t});$
Rank and extract relevant tweets to \mathcal{R}_t by $e^{-KL(\vartheta_{d,t}, \theta_{u,t})};$ **Initialize:** Extract N tweets from \mathcal{R}_t to $RT_{u,t};$ **repeat** Extract $\mathcal{X}_t = \{d_x \in \mathcal{R}_t \cap d_x \notin RT_{u,t}\};$ **for** $d_x \in \mathcal{X}_t, \forall d_y \in RT_{u,t}$ **do** Calculate $\mathcal{S}_{RT_{u,t}} = F(\mathcal{L}_C \cdot \mathcal{L}_N \cdot \mathcal{L}_D);$

Calculate

 $\Delta \mathcal{S}_{d_x, d_y} = \mathcal{S}((RT_{u,t} - d_y) \cup d_x) - \mathcal{S}(RT_{u,t});$ **end** Get $\langle \hat{d}_x, \hat{d}_y \rangle$ that $\langle \hat{d}_x, \hat{d}_y \rangle = \arg \max_{d_x, d_y} \Delta \mathcal{S}_{d_x, d_y};$ $RT_{u,t} = (RT_{u,t} - \hat{d}_y) \cup \hat{d}_x;$ **until** $\forall \Delta \mathcal{S}_{d_x, d_y} < \varepsilon;$ **return** $RT_{u,t}.$

streaming API,² which contains a random sample of around 10% of all items posted on Twitter. Timestamps in our dataset are from November 1, 2009 to December 31, 2010; the 2009 part contains 47,373,408 tweets and 562,361 users, while the numbers for 2010 are 295,145,421 and 5,828,356, respectively. Figure 4(a) shows the statistics of the number of tweets per user in our dataset, where we can find that most users (75.2%) in our dataset wrote fewer than 100 tweets. For crawling the social relations, we use the dataset from [15], which includes social relations for all users on Twitter until July 2009. In our experiments, we use only those tweets and users that appear in both datasets. In our experiments we assume social relations among users to remain the same over the entire time period.

Since it is impossible to evaluate the effectiveness if a user posted nothing on Twitter, sparse postings obstruct our experimental evaluation. We therefore only consider users who posted a sufficient number of tweets for our evaluation: we collect users who post over 100 tweets in our dataset. This results in a subset containing 32,659 users. Thereafter we use social relations to build the social circles around those users. Figure 4(b) shows the number of tweets of these users (y-axis) versus the number of friends on the x-axis. We further remove non-English tweets through automatic language identification [3]. We remove stop words and apply Porter stemming [23].

5.1 Data Enrichment

Since each tweet is only up to 140 characters long, the amount of textual evidence to work with is very limited. To remedy this, we employ a state-of-the-art method for *linking* tweets to Wikipedia articles [20]. In particular, we employ the so-called CMNS method that uses the prior probability that Wikipedia article c is the target of a link with anchor text q within Wikipedia:

$$CMNS(c, q) = \frac{|L_{q,c}|}{\sum_{c'} |L_{q,c'}|}, \quad (13)$$

where $L_{q,c}$ denotes the set of all links with anchor text q and target c .

After we have obtained three Wikipedia articles with the highest CMNS score, we extract the most *central* sentences from these Wikipedia articles and append them to the tweet. In particular, we apply a query-sensitive graph-based summarization method, simi-

²<https://dev.twitter.com/docs/streaming-apis>.

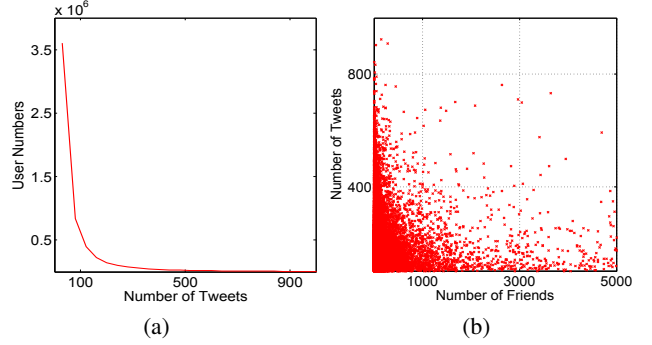


Figure 4: Histograms of the number of users and tweets in our dataset: the left (a) indicates the number of tweets per user in our dataset where the y-axis denotes the number of users; while the right (b) indicates the number of tweets per user with its number of friends in Twitter, where y-axis indicates the number of tweets the user wrote and the x-axis indicates the number of friends.

lar to [10], to each Wikipedia article to ranking sentences, using the tweet d_t as the query. This calculates the score of each sentence via “votes” from other sentences in a document. Figure 5 shows 4 example tweets and the appended sentences. Here, the left text box in each item is a tweet and on the right we show the identified sentences from the linked Wikipedia articles.

5.2 Experimental Setup

Following existing topic models [11], we set pre-defined values for the hyperparameters α_t and β_t in our graphical model: for the weighted parameter $\alpha_{u,t}$ and β_t , we set $50/K_t^u$ to $\alpha_{u,t}$ and 0.5 to β_t respectively. And we set $50/K_t^B$ to α^B and 0.5 to β^B respectively. For the hyperparameters γ and σ in TPM, as defined in [14], we set $\sigma_u = \gamma_{com} = 0.5$ and $\gamma_u = \sigma_{com} = 0.3$. For burst topics we set $\gamma_B = \sigma_B = 0.2$ in our experiments. The initial value of $\lambda_{u,c_i,t-1}$ for each social circle of u is set to $1/C_{u,t}$, the parameter μ is set as 0.85; and ε in Algorithm 2 is set to 0.0001. For the number of topics in our topic modeling process, the default values for Z_0^u and Z_0^{com} in our experiments are set to 100, respectively. To optimize the number of topics, we compare performance in various values and discuss it latter.

Statistical significance of observed differences between two comparisons is tested using a two-tailed paired t-test. In our experiments, statistical significance is denoted using \blacktriangle for significant differences for $\alpha = 0.01$, or \blacktriangle for $\alpha = 0.05$.

5.3 Evaluation Metrics

Evaluating the effectiveness of time-aware tweets summarization is a challenging task, especially in the absence of explicit user feedback. One possible solution is to use evidence from users themselves: we use a user’s retweeted post(s) at time $t+1$ as the ground truth to evaluate performance of comparisons at time t .

We measure the quality of summaries by counting overlapping textual *units* between the generated results and the ground truth results. In our experiments, we adopt the ROUGE evaluation metrics [17], a widely-used recall-oriented metric in the task of document summarization that evaluates the overlap between a gold standard and candidate selections.³ In our experiments, ROUGE-1 (*unigram based method*), ROUGE-2 (*bigram based method*) and ROUGE-W (*weighted longest common sequence*) are used as evaluation metrics.

³Version 1.5.5 is used in this paper.

<p>whenever i see lily tomlin, i call her debbie fiderer</p>	<p>She is Executive Assistant to the President, and is portrayed by noted actor and comedienne Lily Tomlin.</p> <p>She appeared on the dramatic series The West Wing for four years (2002-2006) in the recurring role of presidential secretary Deborah Fiderer. Nominated—Screen Actors Guild Award for Outstanding Performance by an Ensemble in a Drama Series (2003, 2005)</p>	<p>Tyranny Disguised As Health Care Reform & How To Truly Reform Health Insurance:</p>	<p>Health care reform in the United States</p> <p>President Obama gave a speech at a rally in Pennsylvania explaining the necessity of health insurance reform and calling on Congress to hold a final up or down vote on reform.</p> <p>The law includes health-related provisions that take effect over several years, including expanding Medicaid eligibility for people making up to 133% of the federal poverty level (FPL)</p>
<p>Waiting for Charlie Sheen to say the Bush administration on was behind the attack on his wife.</p>	<p>Sheen has since become a prominent advocate of the 9/11 Truth movement. He was characterized by the press as believing the 9/11 Commission was a whitewash and that the administration of former President George W. Bush may have been responsible for the attacks.</p> <p>After the terrorist attacks on September 11, 2001, Bush declared a global War on Terrorism and, in October 2001, ordered an invasion of Afghanistan to overthrow the Taliban, destroy Al-Qaeda, and to capture Osama bin Laden.</p> <p>On December 25, 2009, Sheen was arrested for assaulting his wife, Brooke Mueller in Aspen, Colorado.</p>	<p>Hyundai Enhances Assurance for 2010: This complimentary service is termed Hyundai Assurance</p>	<p>On January 6, Hyundai reported sales of December 2008 fell to 24,037, from 46,487 in previous year and sales for the year dropped 14%, a day after the company launched 'Hyundai Assurance' in order to spark sales amid tough economic conditions.</p> <p>In 2010, a Consumer Reports reliability survey ranked Hyundai (including Kia) as the fourth-best automaker.</p> <p>U.S. Hydrogen Highway Paved With Public-Private Research Funds</p>

Figure 5: Four examples for entity linking and ranking corresponding to four individual tweets, where the textbox on left side indicates the original tweet while the textbox on the right side shows the extracted related sentences. A mixture of the tweet and extracted wiki sentences will replace the original tweet in our experiments.

5.4 Baseline Comparisons

Given the TPM modeling introduced in Section 4.1, our contribution is twofold: (1) we introduce collaborative influence to user’s interests detection; (2) we adopt time-aware propagation to infer topics. To evaluate the influence of social circles and time-aware topics, besides our overall TPM-based strategy, we also evaluate the performance of the model that only includes (1) the collaborative influence or only the (2) time-aware propagation, respectively.

We write **TPM-ALL** for the overall process as described in Section 4.1, which includes both the social influence modeling and time-aware topic and interests tracking. We write **TPM-SOC** for the model that only considers users’ social influence (so excluding time-aware topic propagation and it doesn’t consider if some topic is private or not). We write **TPM-TOP** for the model that uses a user’s own tweets (without social circles but considering topic and interests propagation with the time).

To evaluate our proposed method in more detail, in our experiments the baselines not only include widely-used topic models, but also recent user behavior models on Twitter. For those topic models, we use the Author-Topic Model (**AT**) [26] and the Twitter-LDA [37] as baselines for topic models: (AT) focuses on various users’ interests in one static corpus. Since each tweet only has one author, AT’s process on Twitter coincides with the LDA modeling process on all tweets written by a specific user. As an extension of the author-topic model, Twitter-LDA (**TLDA**) classifies topics into private topics and background topic by introducing one binomial distribution. For comparison, we use one more state-of-the-art user behavior model, **UBM** [32]; here, a user’s interest is tracked by a mixture graphical model that considers background knowledge, social interests and the user’s own interest. The final baseline that we consider is **TF-IDF**, which uses TF-IDF to re-calculate $\mathcal{S}_{RT_{u,t}}$ in Algorithm 2. Finally, we also use **SUM-TF**, a baseline used in [4] that extract tweets by ranking tf scores, and **Random**, which extracts tweets randomly in each period.

For the baseline topic models, we use a similar tweet selection method as in Algorithm 2 to select tweets in each time interval. For

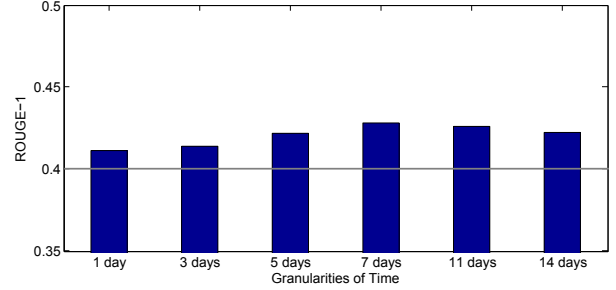


Figure 6: Performance of TPM-ALL with various granularities of time periods.

static topic models, results at time t , $1 \leq t \leq T$ are calculated after re-modeling for all past data before period t .

To evaluate the effectiveness of results to *personalized* aspect, we introduce several other sentence extraction procedures from the area of document summarization (without personalization) as baselines: **LexRank** and **Centroid** are two widely-used unsupervised document summarization methods, where **LexRank** [26] is a graph-based method for ranking tweet as “votes” from other tweets, and **Centroid** [24] applies the MEAD summarization method that uses statistical and structural features in tweets selection.

5.5 Granularities and Number of Topics

To test the optimal granularity of time intervals, we examine ROUGE-1 performance of TPM-ALL with different values for granularities, shown in Figure 6. The performance of TPM-ALL in terms of ROUGE-1 peaks when the granularity is set to 7 days. With fewer than 7 days, performance keeps increasing because adding more days reduces sparseness; but after 7 days, due to the increase in irrelevant and noisy tweets, the ROUGE-1 score decrease. Thus, we set the granularity to 7 days in the remainder of our experiments.

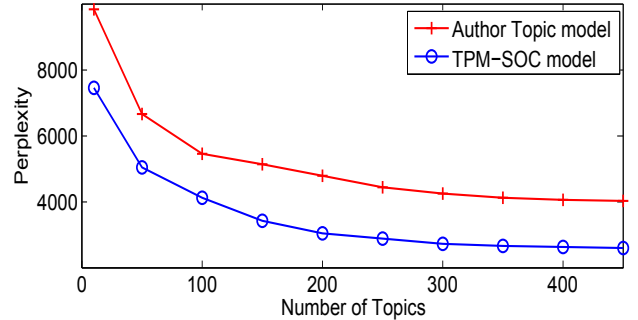


Figure 7: Perplexity performance with different number of topics in Author topic model and TPM-SOC model;

Optimizing the number of topics is a problem shared between all topic modeling approaches. Similar to previous work [2, 11, 32], we introduce the perplexity of a held-out test set to evaluate the performance of our topic models. The perplexity, usually used in language modeling, focuses on the inverse of the geometric mean per-word likelihood, which is calculated as follows:

$$Perplexity(W) = \exp \left[- \frac{\sum_{t \in T} \sum_{w \in D_t} \log p(w)}{\sum_{t \in T} d_t} \right] \quad (14)$$

where $p(w)$ indicates $p(w) = p(w|z)p(z)$. Thus, a lower perplexity score indicates a better generalization performance [2]. Figure 7 shows the results of perplexity values for the author-topic model

and the TPM-SOC model with differing numbers of topics on our held-out test set. After the number of topics becomes larger than 300, the perplexity of both approaches starts to flatten out. We find that TPM-SOC outperforms the author-topic model with better generalization performance. For TPM-ALL and TPM-TOP we set the number of “private” topics and “common” topics to 150, separately.

6. RESULTS AND DISCUSSION

In this section we provide an answer to the following research questions. (1) How does the TPM-based TaTS strategy perform on time-aware tweets summarization (§6.1)? (2) How does the TPM-based TaTS strategy perform on social-aware tweets summarization? (§6.2)? And (3) what is the overall performance for TPM on the task of personalized TaTS (§6.3)?

6.1 Time-Aware Comparisons

To illustrate the performance at different time periods, the evaluation results of the TPM-ALL, TPM-TOP, UBM and AT strategies at different time periods are shown in Figure 8, in terms of ROUGE-1, ROUGE-2 and ROUGE-W, respectively. We select 10 contiguous weeks from November 1, 2009 onwards as the test period and separate it into 10 periods.

In Figure 8 we observe that the AT model obtains the worst performance, while both TPM-ALL and TPM-TOP outperform all other strategies in terms of ROUGE metrics at all time intervals. This demonstrates the advantage of TPM-based strategies in time-aware comparisons. In Figure 8, we observe a “cold-start” phenomenon, which results from the sparseness of the context in the first time period. In that condition, TPM-ALL and TPM-TOP are nearly equivalent to the UBM and AT since there are neither social circles nor burst topics during the first time period. After that, the performance of the TPM based methods keeps increasing over time until it achieves a stable performance after $t = 3$. We find that TPM based strategies are sensitive to time-aware topic drifting. Meanwhile, we find that TPM-ALL performs better than TPM-TOP in Figure 8. TPM-ALL detects user’s interests using social circles whereas TPM-TOP ignores them.

6.2 Social-Aware Comparisons

To evaluate the influence of social circles in our proposed strategy, we investigate the performance under various numbers of social circles. From our dataset, we extract users with different numbers of social circles and compare the performance of our methods on these data sets in terms of ROUGE. In Figure 9 we plot the values of ROUGE-1, ROUGE-2 and ROUGE-W in (a) to (c), respectively. For each figure, we compare our strategies that do consider social circles, TPM-ALL and TPM-SOC, against the TPM-TOP and UBM methods under varying number of social circles.

We observe from Figure 9(a) that the performance in terms of ROUGE-1 changes with the number of social circles, and the value increases and achieves a maximal value between 3 and 5 social circles. After that, the value decreases rapidly; redundant and irrelevant “relations” seem to enter the picture. Another plausible explanation concerns the difference of user characteristics in various social circles. Since the UBM and TPM-TOP models do not consider the social influence, their ROUGE values keep constant for different numbers of social circles. We observe a similar behavior in Figure 9(b) and 9(c) in terms of ROUGE-2 and ROUGE-W.

To evaluate the effect of collaborative filtering in TPM for various classes of users, especially for “passive” users on Twitter who rarely write a tweet, we compare the performance of different users in terms of ROUGE metrics with varying values of the number

of tweets selected per period (40 or 60). We separate users into 3 classes by counting their tweets: (1) less than 400 tweets; (2) between 400 to 800; and (3) more than 800 tweets. As shown in Figure 10(a) and (c) that focusing on ROUGE-1, the difference between TPM-ALL and TPM-TOP is bigger for users with up to 400 tweets than for those with more than 400. This can be explained by the fact that the collaborative filtering used in TPM-ALL becomes more effective when there is a bigger data sparseness issue to overcome. In terms of ROUGE-2, similar results can be found in Figure 10(b) and (d).

6.3 Overall Performance

Table 2 shows the average performance of our TPM-based strategies and baselines, in terms of ROUGE-1, ROUGE-2 and ROUGE-W, based on all candidate tweets in all time periods. We find that our method outperforms the baselines in every case. Except for our TPM-based strategies, UBM get the best performance than others. Since summarization baselines are not sensitive to users’ interests, thus we find that Centroid, Lex-R (short for LexRank), and SUM-TF do not perform well. Among the topic models, we found that the AT-based method yields almost the worst performance. This can be explained by the fact that the topic modeling procedure in AT does not capture topic drift and users’ social circles.

We evaluated the performance of the various approaches in terms of the three ROUGE metrics for a varying number of tweets selected per period, i.e., $N = 40$ and $N = 60$. As shown in Table 2, TPM-ALL performs better than all baselines on all metrics. For $N = 40$, TPM-ALL achieves an increase of 10.6%, 11.6% and 8.9% over UBM in terms of ROUGE-1, ROUGE-2, and ROUGE-W respectively. For $N = 60$, TPM-ALL gives an increase of 11.2%, 11.2% and 10.1% over UBM. For the dynamic version without social influence, TPM-TOP outperforms all other baselines also, which indicates the effectiveness of detecting dynamic topics. We further compare TPM-TOP with UBM: for $N = 40$, TPM-TOP offers relative performance improvements of 4.1%, 6.25% and 4.8%, respectively, for the ROUGE-1, ROUGE-2 and ROUGE-W metrics, while the relative improvements are 7.8%, 6.7% and 7.3% on the same metrics for $N = 60$. We find that TPM-ALL outperforms the UBM baselines with a statistical significance difference at level $\alpha < 0.01$ in terms of all ROUGE metrics, whereas TPM-TOP and TPM-SOC outperforms UBM with a statistical significance difference at level $\alpha < 0.05$.

7. CONCLUSION AND FUTURE WORK

We have considered the task of personalized time-aware tweets summarization, based on user history and influences from “social circles.” To handle the dynamic nature of topics and user interests along with the relative sparseness of individual messages, we have proposed a time-aware user behavior model. Based on probabilistic distributions from our proposed topic model, the tweets propagation model, we have introduced an iterative optimization algorithm to select tweets subject to three key criteria: novelty, coverage and diversity. In our experiments we have verified the effectiveness of our proposed method, showing significant improvements over various state-of-the-art baselines.

As to future work, we aim to employ a user-study to enhance the accuracy of interest detection, e.g., via an online evaluation. Another future direction is to take more information and features into account for our task: our current experiments ignore, e.g., URLs appearing in tweets which could enhance our entity linking setup. It will also be interesting to consider other features for modeling, such as geographic or profile information. Finally, our current model is evaluated based on fixed time intervals, which might not accurately

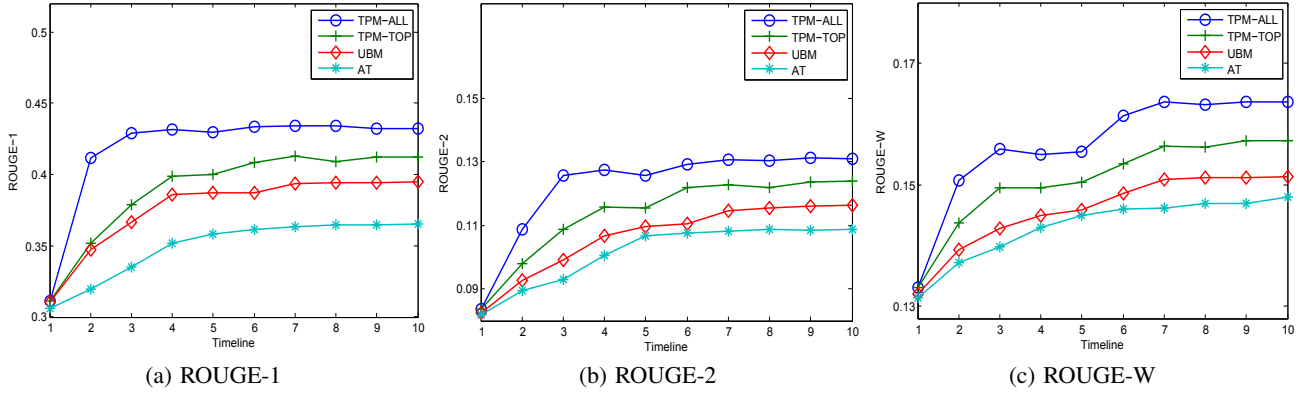


Figure 8: Time-aware performance in terms of ROUGE metrics.

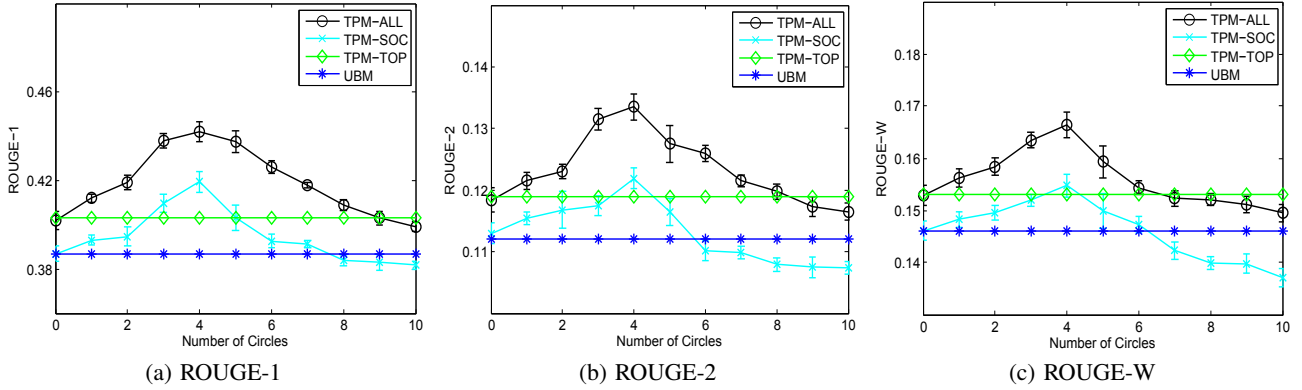


Figure 9: ROUGE-1 and ROUGE-2 performance with increasing numbers of social circles.

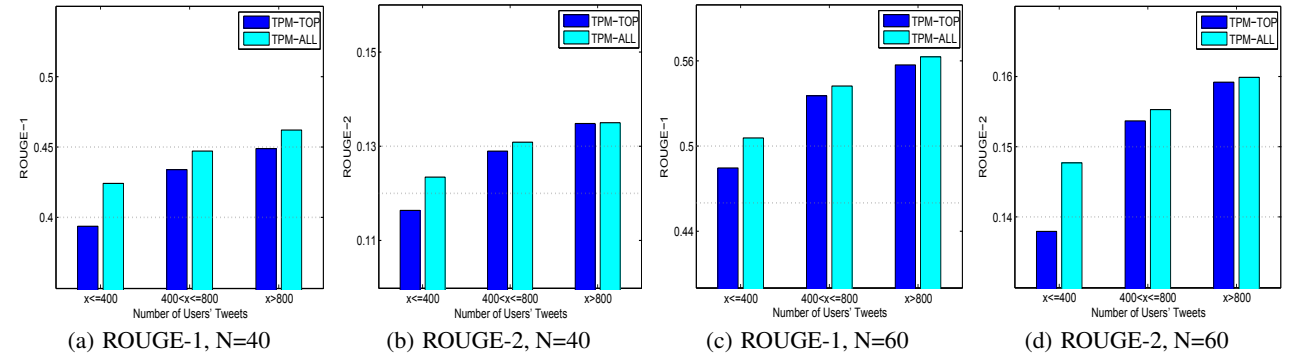


Figure 10: Performance for different kinds of users: users in our dataset are classified by their number of tweets.

reflect bursty topics on Twitter. Therefore, a novel graphical model that includes dynamic time bins instead of the fixed time granularities, will be another direction for future research.

Acknowledgements. This research was supported by the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSine project), the Netherlands Organisation for Scientific Research (NWO) under project nrs 640.004.802, 727.011.005, 612.001.116, HOR-11-10, the Center for Creation, Content and Technology (CCCT), the BILAND project funded by the CLARIN-nl program, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences

(KNAW), and the Netherlands eScience Center under project number 027.012.105.

8. REFERENCES

- [1] D. Blei and J. Lafferty. Dynamic topic models. In *ICML 2006*, pages 113–120, 2006.
- [2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.
- [3] S. Carter, W. Weerkamp, and M. Tsagkias. Microblog language identification: overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 2012.

Table 2: Overall ROUGE Performance for All Comparisons

Metrics	TPM-ALL	TPM-TOP	TPM-SOC	UBM	TLDA	AT	TF-IDF	Centroid	Lex-R	SUM-TF	Random
Cut-off of $N = 40$ tweets per period											
ROUGE-1	0.428 [▲]	0.403 [▲]	0.395 [▲]	0.387	0.374	0.355	0.341	0.302	0.291	0.274	0.252
ROUGE-2	0.125 [▲]	0.119 [△]	0.116 [△]	0.114	0.112	0.102	0.095	0.081	0.077	0.079	0.037
ROUGE-W	0.159 [▲]	0.153 [▲]	0.149 [△]	0.146	0.142	0.144	0.137	0.118	0.115	0.105	0.076
Cut-off of $N = 60$ tweets per period											
ROUGE-1	0.513 [▲]	0.497 [▲]	0.482 [▲]	0.461	0.457	0.423	0.411	0.362	0.369	0.329	0.281
ROUGE-2	0.149 [▲]	0.142 [△]	0.139 [△]	0.134	0.127	0.122	0.116	0.097	0.102	0.095	0.041
ROUGE-W	0.197 [▲]	0.191 [▲]	0.189 [▲]	0.178	0.176	0.166	0.161	0.131	0.135	0.119	0.081

- [4] D. Chakrabarti and K. Punera. Event summarization using tweets. In *ICWSM 2011*, pages 66–73, 2011.
- [5] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu. Collaborative personalized tweet recommendation. In *SIGIR 2012*, 2012.
- [6] B. Connor, M. Krieger, and D. Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. *ICWSM 2010*, pages 2–3, 2010.
- [7] G. De Francisci Morales, A. Gionis, and C. Lucchese. From chatter to headlines: harnessing the real-time web for personalized news recommendation. In *WSDM 2012*, 2012.
- [8] Q. Diao, J. Jiang, F. Zhu, and E. Lim. Finding bursty topics from microblogs. In *ACL 2012*, 2012.
- [9] Y. Duan, Z. Chen, F. Wei, M. Zhou, and H. Shum. Twitter topic summarization by ranking tweets using social influence and content quality. In *COLING 2012*, pages 763–779, 2012.
- [10] G. Erkan and D. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *JAIR*, 22: 457–479, 2004.
- [11] T. Griffiths and M. Steyvers. Finding scientific topics. *National Academy of Sciences*, 101:5228–5235, 2004.
- [12] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR 1999*, pages 50–57, 1999.
- [13] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda. Topic tracking model for analyzing consumer purchase behavior. In *IJCAI 2009*, volume 9, pages 1427–1432, 2009.
- [14] O. Jin, N. Liu, K. Zhao, Y. Yu, and Q. Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. In *CIKM 2011*, 2011.
- [15] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW 2010*, pages 591–600, 2010.
- [16] L. Li, K. Zhou, G. Xue, H. Zha, and Y. Yu. Enhancing diversity, coverage and balance for summarization through structure learning. In *WWW 2009*, 2009.
- [17] C. Lin. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*, pages 74–81, 2004.
- [18] H. Ma, I. King, and M. Lyu. Learning to recommend with social trust ensemble. In *SIGIR 2009*, pages 203–210, 2009.
- [19] H. Ma, D. Zhou, C. Liu, M. Lyu, and I. King. Recommender systems with social regularization. In *WSDM 2011*, pages 287–296, 2011.
- [20] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *WSDM 2012*, pages 563–572, 2012.
- [21] J. Nichols, J. Mahmud, and C. Drews. Summarizing sporting events using twitter. In *IUI 2012*, pages 189–198, 2012.
- [22] M. Pennacchiotti, F. Silvestri, H. Vahabi, and R. Venturini. Making your interests follow you on twitter. In *CIKM 2012*, 2012.
- [23] M. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 1980.
- [24] D. Radev, H. Jing, M. Styś, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing & Management*, 2004.
- [25] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *ICWSM 2010*, pages 130–137, 2010.
- [26] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI 2004*, pages 487–494, 2004.
- [27] B. Sharifi, M. Hutton, and J. Kalita. Summarizing microblogs automatically. In *NAACL 2010*, 2010.
- [28] H. Takamura, H. Yokono, and M. Okumura. Summarizing a document stream. *Advances in Information Retrieval*, pages 177–188, 2011.
- [29] H. Wallach. Topic modeling: beyond bag-of-words. In *ICML 2006*, pages 977–984, 2006.
- [30] X. Wei, J. Sun, and X. Wang. Dynamic mixture models for multiple time series. In *IJCAI 2007*, pages 2909–2914, 2007.
- [31] J. Weng, E. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM 2010*, pages 261–270, 2010.
- [32] Z. Xu, Y. Zhang, Y. Wu, and Q. Yang. Modeling user posting behavior on social media. In *SIGIR 2012*, pages 545–554, 2012.
- [33] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *SIGIR 2011*, pages 745–754, 2011.
- [34] S. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, and H. Zha. Like like alike: joint friendship and interest propagation in social networks. In *WWW 2011*, pages 537–546, 2011.
- [35] Z. Yang, K. Cai, J. Tang, L. Zhang, Z. Su, and J. Li. Social context summarization. In *SIGIR 2011*, 2011.
- [36] M. Ye, X. Liu, and W. Lee. Exploring social influence for recommendation: a generative model approach. In *SIGIR 2012*, 2012.
- [37] X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E. LIM, and X. Li. Topical keyphrase extraction from twitter. In *ACL 2011*, 2011.