Tagging Your Tweets: A Probabilistic Modeling of Hashtag Annotation in Twitter

Zongyang Ma Aixin Sun Quan Yuan Gao Cong
School of Computer Engineering, Nanyang Technological University, Singapore 639798
{zma4, qyuan1}@e.ntu.edu.sg, {axsun, gaocong}@ntu.edu.sg

ABSTRACT

The adoption of hashtags in major social networks including Twitter, Facebook, and Google+ is a strong evidence of its importance in facilitating information diffusion and social chatting. To understand the factors (*e.g.,* user interest, posting time and tweet content) that may affect hashtag annotation in Twitter and to capture the implicit relations between latent topics in tweets and their corresponding hashtags, we propose two PLSA-style topic models to model the hashtag annotation behavior in Twitter. ***Content-Pivoted Model*** (CPM) assumes that tweet content guides the generation of hashtags while ***Hashtag-Pivoted Model*** (HPM) assumes that hashtags guide the generation of tweet content. Both models jointly incorporate user, time, hashtag and tweet content in a probabilistic framework. The PLSA-style models also enable us to verify the impact of social factor on hashtag annotation by introducing social network regularization in the two models. We evaluate the proposed models using perplexity and demonstrate their effectiveness in two applications: retrospective hashtag annotation and related hashtag discovery. Our results show that HPM outperforms CPM by perplexity and both user and time are important factors that affect model performance. In addition, incorporating social network regularization does not improve model performance. Our experimental results also demonstrate the effectiveness of our models in both applications compared with baseline methods.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

Keywords

Twitter, Hashtag, Topic model, Hashtag annotation

1. INTRODUCTION

Twitter is one of the most popular social networking and microblogging platforms. It has accumulated a tremendous amount of text data; as at January 2014, on average 58 million tweets are posted per day by more than 645 million active Twitter users.[1] These tweets cover a larger number of diverse topics, including comments on recent or ongoing events and emerging topics, personal activities, politics and many others. Due to the informal writing style and the 140-character length constraint, tweets are short, noisy, and are often posted with very limited context.

Hashtag (*i.e.,* keyword prefixed with # symbol) has demonstrated its effectiveness in bringing organization to the sparse information in Twitter. Hashtags associated with tweets enhance information diffusion and tweet search as well as facilitate social chatting. Reported in a recent survey by RadiumOne [20], 58% of Twitter users utilize hashtags on a regular basis. Because of its effectiveness, hashtag has been adopted as a key feature in other micro-blogging services like Tumblr and Sina Weibo, and recently has been officially supported in Google+ and Facebook.[2]

The effectiveness of hashtags in tweets, however, is limited by the freedom of users in deciding (i) whether or not to annotate tweets with hashtags, and (ii) which hashtags to use (*e.g.,* #cikm, #cikm14, or #cikm2014). In 2010, only about 11% of tweets were annotated with one or more hashtags [11]. Detailed in our literature survey, lots of studies related to hashtags in Twitter have been carried out. However, there is a lack of study on the formal modeling of the latent relationship between the important factors in affecting hashtag annotation. In this study, we consider and model user interest, posting time, tweet content, and hashtag in a probabilistic framework for better understanding hashtag annotation at topic level. A topic-level modeling of these factors in hashtag annotation benefits many applications such as retrospective hashtag annotation, related hashtag discovery, hashtag summarization, etc..

Each tweet contains at least three attributes, *i.e.,* tweet content, author (also known as user in this paper), and posting time. As previously mentioned, tweets cover a large number of diverse topics, such as personal activities and comments on recent or ongoing events. As a form of high-level topic abstraction, hashtags in a collection of tweets directly reflect these topics. In other words, hashtags reflect topics related to personal interests/activities of individual users, and also reflect the popular or trending topics in Twitter at that time period. We therefore aim to model the latent topical relationship between *tweet content*, *user*, *time*, and *hashtag*.

Tweet content. As an annotation, a hashtag is a high-level abstraction of the content of a tweet. Among all factors, tweet content is the most important factor affecting the usage of hashtags. However, there could be two kinds of possible associations between a hashtag and a tweet: (i) a user composes a tweet and then finds one or more appropriate hashtags to describe the tweet. In other words, before user finishing writing this tweet, she has no particular hashtag in mind to use. A hashtag is chosen because it best describes

[1] http://www.statisticbrain.com/twitter-statistics/

[2] http://en.wikipedia.org/wiki/Hashtag

CIKM'14, November 3–7, 2014, Shanghai, China.

ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2661903.

the tweet content. (ii) a user composes a tweet with a specific hashtag in mind. In this case, the tweet content could be considered as a detailed elaboration of the pre-chosen hashtag or comment on the event indicated by the hashtag. In this paper, we propose two models to model the two different generation processes between tweet content and hashtag.

User. In general, a large portion of tweets from a common user are about her personal interests/activities (*e.g.,* music, sports, food, travel). The hashtags adopted by a user often reflect such interests and activities. Some of the common hashtags (*e.g.,* #nowplaying, #nba) adopted by a large number of users sharing similar interests lead to informal social communities through these common hashtags as well as mention mechanism. It is reported that social network formed through mentions among users is essential for interaction in Twitter [12]. Intuitively, users who often mention each other are more likely to share similar interests (or similar topics). We therefore consider user as a factor in affecting hashtag annotation and also evaluate the impact of social factor on affecting hashtag annotation.

Time. Twitter is a real-time social media. Many of the tweets are about recent or ongoing events. Many tweets, hence their associated hashtags, published in a time period are about hot events at that time period. Take the royal wedding as an example, on April 29, 2011, hashtags like #royalwedding and #bbcwedding were used to annotate thousands of tweets reporting the wedding of Prince William and Catherine Middleton. The usage of both hashtags reduces significantly in a week's time. The time factor enables our models to better associate time-sensitive hashtags with tweets.

Considering the three factors and the two generation processes, we propose two PLSA-style models, namely, *Content-Pivoted Model* (CPM) and *Hashtag-Pivoted Model* (HPM), to jointly model the relationship between user, time, tweet content and hashtag, at topic level. CPM assumes that a user composes a tweet and then finds the appropriate hashtags to describe the tweet. HPM assumes that a user composes a tweet with pre-selected hashtag(s) in mind. We further incorporate and evaluate the impact of social factor in our models. Specifically, we evaluate CPM^{sn} (resp. HPM^{sn}) by introducing social network regularization to CPM (resp. HPM) with the assumption that users mention each other more often are more likely to adopt similar hashtags.

As case studies, we utilize our models in two example applications. Retrospective hashtag annotation aims to annotate existing tweets with the most appropriate hashtags because 90% of hashtags are not tagged, observed from our data and reported in other studies [11]. Related hashtag discovery is to search for most related hashtags of a given query hashtag. The related hashtags help in hashtag query refinement, query extension and query recommendation. To summarize, contributions arising from this paper are:

1. To the best of our knowledge, we are the first to model the relations between user interest, time, tweet content, and hashtag through latent topics. Two PLSA-style models, Content-Pivoted Model (CPM) and Hashtag-Pivoted Model (HPM) are proposed to simulate the two hashtag generation processes and both consider these important factors. Based on the assumption that users who often mention each other are more likely to share similar topics, we further introduce social network regularization into the two models to evaluate the impact of social factor.

2. Through extensive experiments, we evaluate our models by perplexity, a standard metric for estimating topic models, and demonstrate that HPM outperforms CPM by perplexity. We also show that the topics discovered by both CPM and HPM cover all major hot events and personal activities in our Twitter data. Our experimental results also indicate that incorporating social network regularization does not improve model performance.

3. We define the problems of retrospective hashtag annotation and related hashtag discovery and utilize our models to address the two problems. Compared against baseline methods, the topic-level representation brings significant improvements on the accuracies in both applications.

The rest of the paper is structured as follows. We survey the related work in Section 2. Section 3 describes the proposed models and their inference algorithms for model parameter estimation. In Section 4, we evaluate the performance of the proposed models by perplexity and the discovered topics. Two applications,retrospective hashtag annotation and related hashtag discovery, are presented and evaluated in Section 5. Section 6 concludes this paper.

2. RELATED WORK

In this section, we begin with a brief overview of the studies on hashtags in Twitter. We then survey the related work on hashtag recommendation, followed by topic models proposed for Twitter.

Hashtags in Twitter. The wide adoption of hashtags in Twitter has attracted significant research attention. Hashtags have been studied from many different perspectives in the literature, such as hashtag adoption prediction [29], hashtag popularity prediction [15,27], hashtag diffusion [24], and hashtag sentiment analysis [2,28].

In the study of hashtag adoption prediction [29], Yang *et al.* stated that there are two main purposes for hashtag adoption, bookmarking tweet content and joining a community on the same topic or trend. The features used in the prediction include (i) *relevance* and *preference* derived from tweets annotated with hashtags, and (ii) *prestige* and *influence* derived from social graph formed by users who adopt a hashtag. In [27], Tsur and Rappoport predicted hashtag popularity on weekly basis using regression model. Both features derived from hashtag itself (*e.g.,* orthography, number of characters in a hashtag) and features derived from tweet content are used in the prediction. Their experiments showed that content features improve prediction performance. Both studies reveal a strong relationship between hashtag adoption and tweet content. In our work, we jointly model hashtag and tweet content to capture their relations at topic level.

Romero *et al.* [24] categorized hashtags into 8 classes (*e.g.,* politics, celebrity, and game) and analyzed the differences in the mechanics of information diffusion of hashtags from different classes. They reported that hashtags on politics are adopted for a longer time period and the exposure times of a hashtag (*i.e.,* how many times a user observes this hashtag in her Twitter stream) plays an important role in hashtag diffusion. A user graph based on mention relationship was constructed in their work to trace information diffusion. Our work also considers the impact of user factor on hashtag adoption and models the social network as a regularization, assuming that users who often mention each other share similar topics.

Hashtag Recommendation. Tag recommendation has established itself into an important research topic. Many techniques like tensor factorization [22,23,26] and graph model [5,7] have been proposed and applied to different social tagging systems like Flickr and Delicious. For Twitter, both user-based recommendation [4, 14] and tweet-based recommendation have been proposed for hashtag recommendation [13, 16, 25, 32]. Next, we briefly survey tweet-based recommendation for being more relevant to our work.

To recommend hashtags for a tweet, Zangerle *et al.* [32] searched for similar tweets to the given tweet by content similarity, then ranked the hashtags by their usage on the similar tweets. Mazzia *et al.* [16] also utilized tweet content for hashtag recommendation using a Bayesian model. Kywe *et al.* [13] further incorporated user preference into the model in [32]. That is, hashtags to be recommended to a tweet d by user u are the hashtags used to annotated many similar tweets to d and the hashtags adopted by many similar users to u. In our experiments, we use this method to be our baseline method in the retrospective hashtag annotation application.

Topic Models for Twitter. Topic models, including Probabilistic Latent Semantic Analysis (PLSA) [9] and Latent Dirichlet Allocation (LDA) [1], are widely employed in text mining and information retrieval. PLSA [9] is a classic topic model and has been used to model various types of data, with or without regularization. Yin *et al.* [30] proposed a model to discover regional topics in Flickr and incorporated GPS information into PLSA with the assumption that topics of nearby regions are more coherent. In [8], a PLSA-style model was presented to mine topics of search queries. The authors incorporated regularization in the model with the assumption that topic distribution of two users are similar if they click on similar documents retrieved. Recently, PLSA models have been applied to mine Twitter data [10, 31]. In [10], a PLSA-style model was proposed to discover topics and identify user's interests from geo-tagged tweets for both topic tracking and location estimation. The PLSA model proposed in [31] further incorporated time factor in addition to geo-location information for location prediction.

Many LDA extensions have been proposed to Twitter. Due to the shortness of tweets, it is often assumed that each tweet has one unique topic [33]. To find bursty topics in Twitter, Diao *et al.* [3] proposed a TimeUserLDA model. This model assumes that tweets posted around similar time are more likely to share similar topics and tweets posted by the same user are more likely to share similar topics. Labeled LDA, a semi-supervised learning model, was proposed in [21], to model the latent relationship between users and tweets in Twitter. Topic models have also been applied to hashtag recommendation [6]. Given a tweet, Godin *et al.* employed LDA to generate its topic distribution, and then recommended top keywords from the dominant topics to this tweet as hashtags. In our proposed solution, we recommend existing hashtags rather than keywords to tweets.

Although both PLSA and LDA extensions have been used to model tweet data, we choose to adopt the PLSA framework for its flexibility in introducing social network regularization.

3. HASHTAG ANNOTATION MODELS

In this section, we present the two hashtag annotation models: *Content-Pivoted Model* (CPM) and *Hashtag-Pivoted Model* (HPM). Both models jointly model tweet content, user, time, and hashtag, but with different assumptions on the generation of hashtag and tweet content. In the following, we start with the notations used in our models and the intuitions in our models. We then present the two models and their inference algorithms. Lastly, we detail the inference algorithms considering social network regularization in the two models CPM and HPM. The models with social network regularization are denoted by CPM^{sn} and HPM^{sn} respectively.

Notations. Let d be a tweet and D be a collection of tweets. Let U be a collection of users each of which has published at least one tweet. We partition time into a sequence of time slots of fixed length and map the publication time of a tweet to a time slot t.[3] Let T be the collection of time slots, V be the word vocabulary, and E be the hashtag vocabulary. A tweet d is a 4-tuple $d = \{u, t, \boldsymbol{w}_d, \boldsymbol{h}_d\}$: $u \in U$ is the author of the tweet; $t \in T$ is the time slot within which d was published; $\boldsymbol{w}_d$ is the word collection in d, where the words are drawn from V; and $\boldsymbol{h}_d$ is the set of distinct hashtags annotated to tweet d, where the hashtags are drawn from E. Note that, a tweet may have more than one hashtag and even duplicated hashtags. In our work, we only consider distinct hashtags for the same tweet.

Intuitions and Assumptions. All our models are designed based on the following two intuitions:

- The topic of a tweet is guided by the personal interest (or activity) of the user who has published this tweet. As discussed in Section 1, a large portion of tweets reflect the users' personal interests or activities. Based on this intuition, we model each user as a topic probability vector. The topic of a tweet from a user is generated based on her corresponding topic probability distribution.

- The topic of a tweet may be affected by time. A large number of tweets are related to recent and ongoing events or trending topics. In other words, each time slot is associated with some major events or popular topics happened or discussed within that time slot. Tweets published in different time slots reflect different topic distributions. Based on this intuition, we model each time slot as a topic probability vector and the topic of a tweet published in that time slot could be generated based on its corresponding topic probability.

As discussed in Section 1, a hashtag is a high-level abstraction of the tweet content. Among all factors, words in a tweet is the most import factor affecting hashtag annotation. However, when composing a tweet with hashtag(s), there could be two possible cases: (i) user composes the tweet first and then finds appropriate hashtags to annotate this tweet, or (ii) user has a hashtag (*e.g.,* a hashtag created for a popular event) in mind and writes a tweet for the hashtag. To model the difference in the order of generating tweet content and hashtags, we propose two models: Content-Pivoted Model which assumes the tweet content is drafted first and the generation (or selection) of the hashtag is guided by the tweet content, and (ii) Hashtag-Pivoted Model which assumes that the user has selected the hashtag and then drafts the tweet content based on her understanding of this hashtag. In both models, we assume that each tweet has only one topic due to its short length. The same assumption has been adopted in many other works [3, 33]. In the following, we detail the two models and their inference algorithms.

3.1 Content-Pivoted Model (CPM)

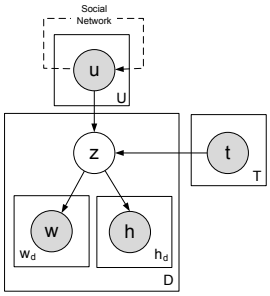
Figure 1(a) (without dotted line) illustrates the Bayesian graphical representation of CPM. Some of the notations used in the model are summarized in Figure 1(c).

The topic z of tweet d is generated from personal interest of u or topic distribution of time slot t. That is, when user u publishes a tweet d in time slot t, she first decides whether to write anything related to her personal interests/activities or to comment on some hot topics in that time slot. More specifically, the topic z of tweet d can be generated from the user topic distribution $p(z|u)$ and the time topic distribution $p(z|t)$. We use a parameter α to balance the importance between $p(z|u)$ and $p(z|t)$:

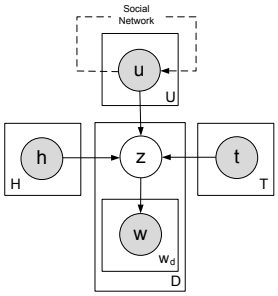
$$p(z|u,t) = \alpha p(z|u) + (1-\alpha)p(z|t)$$

After a topic z is generated, all words $\boldsymbol{w}_d$ in the tweet d are sampled from $p(w|z)$. Then the hashtags of tweet d, $\boldsymbol{h}_d$, are sampled from $p(h|z)$. The generative process of CPM is summarized as follows:

[3] In this paper, the length of a time slot is a day.



(a) CPM and CPM^{sn}



(b) HPM and HPM^{sn}

Symbol	Description
D	Collection of tweets
H	Collection of hashtags
U	Collection of users
T	Collection of time slots
D_u	Collection of tweets posted by user u
D_t	Collection of tweets posted at time slot t
D_h	Collection of tweets having hashtag h
D_w	Collection of tweets having word w
$\boldsymbol{w}_d$	Collection of words in tweet d
$\boldsymbol{h}_d$	Set of hashtags in tweet d

(c) Notations used in the models

Figure 1: Graphical model representations of CPM and HPM (without dotted line), graphical model representations of CPM^{sn} and HPM^{sn} (with dotted line), and the notations used in the models.

- For each tweet $d \in D$, written by user u at time t
  - Draw a topic $z \sim p(z|u,t)$
  - For each word w in $\boldsymbol{w}_d$, draw $w \sim p(w|z)$
  - For each hashtag h in $\boldsymbol{h}_d$, draw $h \sim p(h|z)$

Observe that CPM model incorporates all factors user, time, tweet content, and hashtag into a PLSA framework. Further, the tweet content and hashtag are generated after a topic has been determined based on the user interests and (popular) topics of that time slot.

3.2 Hashtag-Pivoted Model (HPM)

Similar to CPM, the HPM model also jointly considers user, time, tweet content and hashtag. However, as illustrated in Figure 1(b) (without dotted line), HPM models hashtags as a high-level feature partially guiding the generation of tweet content. That is, when drafting a tweet, a user may choose to report her personal interests or comment on some hot events in that time slot as in CPM; in HPM a user may also choose to directly comment on a specific hashtag. In short, the topic z of a tweet may be drawn from user topic distribution $p(z|u)$, time topic distribution $p(z|t)$, or hashtag topic distribution $p(z|\boldsymbol{h}_d)$. Note that, one tweet might have multiple hashtags. We assume that all hashtags of a tweet $\boldsymbol{h}_d$ share equal importance to the tweet d:

$$p(z|\boldsymbol{h}_d) = \frac{1}{|\boldsymbol{h}_d|} \sum_{h' \in \boldsymbol{h}_d} p(z|h')$$

Considering the three factors $p(z|u)$, $p(z|t)$, $p(z|\boldsymbol{h}_d)$, the topic z of a tweet d written by a user u at time slot t with hashtag(s) $\boldsymbol{h}_d$ in mind is:

$$p(z|u,t,\boldsymbol{h}_d) = \beta(\alpha p(z|u) + (1-\alpha)p(z|t)) + (1-\beta)p(z|\boldsymbol{h}_d) \quad (1)$$

Here, the two parameters α and β balance the importance of the three factors in selecting the topic of the tweet. Similarly, after generating the topic z of tweet d, all words $\boldsymbol{w}_d$ are sampled from $p(w|z)$. The generative process of HPM is as follows:

- For each tweet $d \in D$, written by user u at time t for hashtags $\boldsymbol{h}_d$
  - Draw a topic $z \sim p(z|u,t,\boldsymbol{h}_d)$
  - For each word w in $\boldsymbol{w}_d$, draw $w \sim p(w|z)$

3.3 Inference Algorithms

For both CPM and HPM, there is one latent variable topic z to be inferred. The exact inference algorithm is intractable. We propose an Expectation-Maximization (EM) algorithm for appropriately inferring z in both models. Next, we first detail the inference algorithm for CPM. In CPM, the joint probability over tweet d and topic z can be represented as:

$$\begin{aligned} p(d,z) =& p(u,t,z,\boldsymbol{w}_d,\boldsymbol{h}_d) \\ =& p(u)p(t)p(z|u,t)p(\boldsymbol{w}_d|z)p(\boldsymbol{h}_d|z) \end{aligned} \quad (2)$$

where

$$p(z|u,t) = \alpha p(z|u) + (1-\alpha)p(z|t) \quad (3)$$

$$p(\boldsymbol{w}_d|z) = \prod_{w' \in \boldsymbol{w}_d} p(w'|z) \quad (4)$$

$$p(\boldsymbol{h}_d|z) = \prod_{h' \in \boldsymbol{h}_d} p(h'|z) \quad (5)$$

Accordingly, the log-likelihood in CPM is $L = \sum_d \log \sum_z p(d,z)$. We train the model using EM algorithm as follows:

- In **E-step**,

$$p(z|d) = \frac{p(d,z)}{p(d)} = \frac{p(d,z)}{\sum_z p(d,z)} \quad (6)$$

- In **M-step**, it is complicated to estimate $p(z|u)$ and $p(z|t)$, because they are coupled by the sum in logarithm in log-likelihood, *i.e.,* $\log(\alpha p(z|u) + (1-\alpha)p(z|t))$. We apply Jensen's inequality to get a lower bound: $\log(\alpha p(z|u) + (1-\alpha)p(z|t)) \geq \alpha \log p(z|u) + (1-\alpha)\log p(z|t)$. We now maximize the log-likelihood to estimate the following parameters:

$$p(z|u) = \frac{\sum_{d \in D_u} p(z|d)}{\sum_{d \in D_u} \sum_{z'} p(z'|d)} \quad (7)$$

$$p(z|t) = \frac{\sum_{d \in D_t} p(z|d)}{\sum_{d \in D_t} \sum_{z'} p(z'|d)} \quad (8)$$

$$p(w|z) = \frac{\sum_{d \in D_w} n(d,w)p(z|d)}{\sum_{w'} \sum_{d \in D_{w'}} n(d,w')p(z|d)} \quad (9)$$

$$p(h|z) = \frac{\sum_{d \in D_h} p(z|d)}{\sum_{h'} \sum_{d \in D_{h'}} p(z|d)} \quad (10)$$

where $n(d,w)$ represents number of appearances of word w in d, or w's term frequency in d.

The joint probability for HPM over tweet d and topic z is defined in the following equation, where $p(z|u,t,\boldsymbol{h}_d)$ is defined in Equation 1:

$$\begin{aligned} p(d,z) =& p(u,t,\boldsymbol{h}_d,z,\boldsymbol{w}_d) \\ =& p(u)p(t)p(\boldsymbol{h}_d)p(z|u,t,\boldsymbol{h}_d)p(\boldsymbol{w}_d|z) \end{aligned} \quad (11)$$

In Equation 11, $p(\boldsymbol{h}_d) = \prod_{h' \in \boldsymbol{h}_d} p(h')$. The inference algorithm for HPM is similar to that of CPM. Specifically, the E-steps

for both models are the same. In M-step, the estimations of $p(z|u)$, $p(z|t)$, and $p(w|z)$ in CPM also apply to HPM. The additional parameter $p(z|h)$ in HPM is estimated as follows:

$$p(z|h) = \frac{\sum_{d \in D_h} p(z|d)/|\boldsymbol{h}_d|}{\sum_{d \in D_h} \sum_{z'} p(z'|d)/|\boldsymbol{h}_d|} \quad (12)$$

3.4 Social Network Regularization

As discussed in Section 1, users who often mention each other are more likely to share similar topics. With the aim of obtaining more accurate topics in Twitter data, we utilize the mention relationship in Twitter as a regularization R over the topic distribution of a pair of Twitter users who have mentioned each other. More specifically, we minimize the proximity of topic distributions $p(z|u)$ and $p(z|v)$ of two users u and v who have mentioned each other for C_{uv} number of times in their tweets (regardless u mentions v or v mentions u):

$$R = \sum_{u,v \in U} \sum_{z} C_{uv} (p(z|u) - p(z|v))^2$$

The two models CPM and HPM with social network regularization are denoted by CPM^{sn} and HPM^{sn} respectively. The dotted lines in Figures 1(a) and 1(b) denote the social network regularization. Regularized log-likelihood is expressed as $RL = L - \lambda R$, where λ is the regularization parameter ($\lambda = 10$ in our evaluation following the setting in [8]). We maximize the regularized log-likelihood using Generalized EM algorithm [18].

Except for $p(z|u)$, all other parameters in CPM^{sn} and HPM^{sn} are estimated in the same way as their corresponding models CPM and HPM. Next, we use CPM^{sn} as an example to estimate $p(z|u)$ and the same applies to HPM^{sn}. Let $p^i(z|u)$ be the estimation obtained in the i-th iteration of CPM^{sn}, $p^{i+1}(z|u)$ in the $(i+1)$-th iteration is computed using Equation 13 based on the Newton-Raphson method [19]. Note that $p^0(z|u)$ is the $p(z|u)$ estimated in CPM (see Equation 7).

$$p^{i+1}(z|u) = (1-\gamma) p^i(z|u) + \gamma \frac{\sum_{v \in U} C_{uv} p^i(z|v)}{\sum_{v \in U} C_{uv}} \quad (13)$$

In the above equation, γ is the step parameter ($\gamma = 0.1$ in our implementation following the setting in [8]) and C_{uv} is the number of times users u and v who have mentioned each other in their tweets. More details of the algorithm can be found in [8].

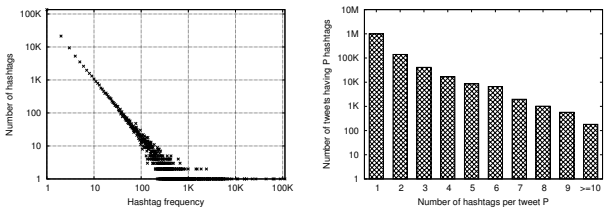
4. EXPERIMENT

We conduct experiments to evaluate the performance of CPM and HPM using perplexity and show example topics discovered by the two models. We also evaluate the impact of introducing social network regularization in both models.

4.1 Data Set

The tweets used in our evaluation are published by Singapore-based usersfrom January 1, 2011 to August 31, 2011.[4] Because our work focuses on the modeling of hashtag annotation, tweets without hashtags are not considered in our evaluation. In other words, each tweet used in our experiments contains at least one hashtag. Stopwords and non-English words are also removed from all tweets and tweets with empty content are then dropped. To ensure that each hashtag has a reasonable number of tweets for topic modeling, tweets annotated with extremely infrequent hashtags (*i.e.,* each is used to annotate fewer than 5 tweets in the whole collection) are removed from our collection. As the result, every hashtag in our final collection has been used to annotate at least 5 non-empty tweets written in English.

[4]User location information is based on the location specified in user profile.



(a) Hashtag frequency distribution (b) Number of hashtags per tweet

Figure 2: Hashtag frequency distribution and number of hashtags per tweet

Table 1: Statistics of the data set

Number of tweets $\|D\|$	1,217,928
Number of distinct hashtags $\|E\|$	14,055
Number of distinct words or vocabulary size $\|V\|$	61,274
Number of users $\|U\|$	13,711
Number of time slots (days) $\|T\|$	243

After preprocessing, the data set used in our experiments contains more than 1.2 million tweets published by over 13 thousand users in 243 days. The tweets are annotated by more than 14 thousand distinct hashtags. Table 1 reports the statistics of our processed data set.

Plotted in Figure 2(a), the hashtag frequency distribution follows a power-law like distribution. That is, most hashtags are used few times by few users, while a small number of hashtags are extremely popular and have been used to annotate many tweets. Observe that 82.2% of tweets in our collection are associated with one hashtag each (see Figure 2(b)). The remaining 17.8% of tweets, each is annotated by more than one hashtag. A small number of tweets are annotated by more than 10 hashtags each.

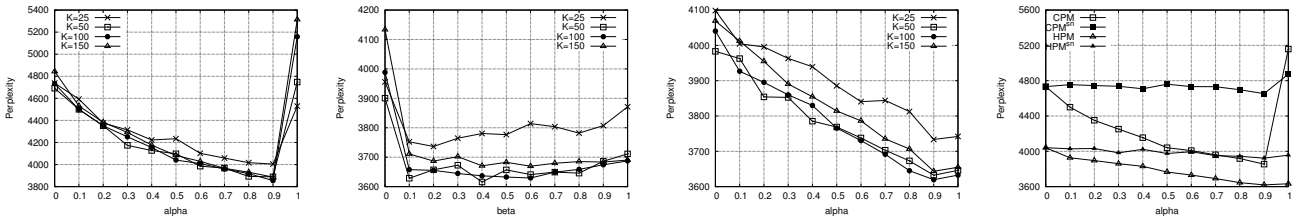
4.2 Evaluation by Perplexity

Perplexity is a standard metric for evaluating topic models [1]. Defined in Equation 14, perplexity measures the ability of a model in generating unseen data (*i.e.,* D_{test} in the equation, which is a set of documents not used in model training). In this equation, $p(\boldsymbol{w}_d)$ indicates the probability of generating all the words in a test document $d \in D_{test}$, and N_d denotes the number of words in document d. Lower perplexity indicates better model performance.

$$Perplexity(D_{test}) = exp\{-\frac{\sum_{d \in D_{test}} \log p(\boldsymbol{w}_d)}{\sum_{d \in D_{test}} N_d}\} \quad (14)$$

In our evaluation, we randomly select 200,000 tweets to be the testing data set, and the remaining 1,017,928 tweets are used to train the models. Next, we first examine the impact of user factor, time factor and the number of topics on the model performance of CPM and HPM respectively by perplexity. We then evaluate the effectiveness of social network regularization on the two models by comparing their perplexity with that of CPM^{sn} and HPM^{sn}. In all our experiments, the number of iteration in training the models is fixed to 100.

CPM Model Performance. Recall that in CPM, the topic z of a tweet d is generated from the user topic distribution $p(z|u)$ and the time topic distribution $p(z|t)$, balanced with a parameter α: $p(z|u,t) = \alpha p(z|u) + (1-\alpha) p(z|t)$. To evaluate the impact of user interest $p(z|u)$ and time factor $p(z|t)$, we vary α from 0 to



(a) CPM (b) HPM (α=0.9) (c) HPM (β=0.6) (d) SN regularization (K =100)

Figure 3: Perplexity of CPM and HPM with varying α, β or K (Figures (a), (b), and (c)), and the impact of social network regularization on CPM and HPM (Figure (d)).

1, with a step of 0.1. Observe that when $\alpha = 0$ topic z is generated from time topic distribution only; and when $\alpha = 1$, topic z is generated purely based on user interest. Figure 3(a) plots the perplexity of CPM with varying α from 0 to 1 for four topic number settings $K = \{25, 50, 100, 150\}$. We make three observations from this result.

First, parameter α has a significant impact on the perplexity of the model which is in the range from 3800 to 5400. With the four different topic number settings, the perplexity values follow very similar trends against the varying of α. When $\alpha = 0.9$, the lowest perplexity is achieved for all the four topic number settings. Either $\alpha = 0$ or $\alpha = 1$ results in much poorer model performance, indicating that (i) both user interest and time are important factors affecting the topic of tweets, and (ii) user interest is often the dominant factor in determining the topics of the tweets from a user.

Second, regarding the choice of number of topics, $K = 25$ or $K = 150$ leads to poorer performance than $K = 50$ or $K = 100$. Particularly, $K = 100$ and $\alpha = 0.9$ delivers the best perplexity in this set of experiments. In all our following experiments, we therefore set $K = 100$ and $\alpha = 0.9$ as the default settings.

Third, when tweet topics are purely drawn from time topic distributions (*i.e.,* $\alpha = 0$), the number of topics K has a limited impact on the perplexity. However, when tweet topics are solely generated based on user interest (*i.e.,* $\alpha = 1$), the smaller the number of topics (*i.e.,* $K = 50$), the better the perplexity. This observation suggests that a common user usually does not show interests in too many different topics.

HPM Model Performance. Compared with CPM, HPM considers one more factor $p(z|\boldsymbol{h}_d)$ in generating the topic of a tweet. More specifically, $p(z|u,t,\boldsymbol{h}_d) = \beta(\alpha p(z|u) + (1-\alpha)p(z|t)) + (1-\beta)p(z|\boldsymbol{h}_d)$. Note that $\beta = 0$ leads to tweet topic generation solely based on hashtags $p(z|\boldsymbol{h}_d)$.

Based on the results of CPM, we first set $\alpha = 0.9$ and evaluate the perplexity of HPM against the varying of β from 0 to 1 with a step of 0.1. Demonstrated in Figure 3(b) the impact of β on HPM is not significant when $\beta \geq 0.1$ for all K values. When $K = 100$, HPM achieves the best perplexity when $\beta = 0.6$. However, it is observed that the perplexity is much poorer when $\beta = 0$, *i.e.,* the topic of a tweet is purely generated based on hashtags.

Next, we fix $\beta = 0.6$ and vary the values of α from 0 to 1 (see Figure 3(c)). Similar to that in CPM, the perplexity of HPM is best when $\alpha = 0.9$ for all the four different numbers of topics. Compared with CPM, HPM performs better by perplexity, with perplexity ranging from 3600 to 4100. One reason is that HPM treats hashtags as topic vectors which could better cluster the words in tweets leading to better topic cohesion.

Social Network Regularization. We now evaluate the impact of considering social factor in the two models. In this set of experiments, we set number of topics $K = 100$ for all four models: CPM, CPM^{sn}, HPM and HPM^{sn}. For both HPM and HPM^{sn}, β is set to 0.6 based on earlier experimental results. The two additional parameters γ and λ in CPM^{sn} and HPM^{sn} are experimentally set to $\gamma = 0.1$ and $\lambda = 10$ (see Section 3.4).

Figure 3(d) shows the perplexity of all four models with α varying from 0 to 1. Note that when $\alpha = 0$, user interest is not considered in the model and therefore no social factor is considered as well. As shown in Figure 3(d), the introduction of social network regularization makes both models much worse in terms of perplexity. One possible reason is that, two users may mention each other because of common interests in some but not all the topics. The assumption that a pair of users who mention more about each other are more likely to share similar topic distributions might be too strong. However, on the other hand, predetermining a subset of common topics for a given pair of users is infeasible in generative models.

Table 3: Example topics found by HPM but not CPM

Topic label	Words with highest generative probability
royal wedding	wedding kate club quay clarke royal river demi rd valley
food	singapore food news paying restaurant world bill cash hotel free
business	ltd pte singapore manager executive sales assistant services tfeeds jeffs
shopping	parade tampines st blk 33 corner 314 gaming sunnys marine

4.3 Topic Discovery

We now present 8 sample topics discovered by the two models CPM and HPM. For both models, we set the number of topics to be 100. From the 100 topics, we select 8 topics as examples.

Table 2 lists these 8 topics. We further manually label these 8 topics to better explain them. For each topic CPM generates word probability $p(w|z)$ and hashtag probability $p(h|z)$ (see Section 3.1). We therefore list both the top words and the top hashtags according to their generative probabilities for each of the 8 sample topics. For clarity, we name these two kinds of topics *word topic* and *hashtag topic* respectively. For HPM, the model only generates word topic based on $p(w|z)$ (see Section 3.2). Twitter topics can be categorized into *exogenous topics* and *endogenous topics* [17]. Exogenous topics (*e.g.,* #earthquake and #flood) are originated outside of Twitter and endogenous topics (*e.g.,* #10thingsihate and #nowplaying) are originated within Twitter. As shown in Table 2, both CPM and HPM models capture the major topics discussed by Singapore users in Twitter from January to August 2011. Among them Singapore General Election[5] and Japan earthquake[6]

[5] http://en.wikipedia.org/wiki/Singaporean_general_election,_2011
[6] http://en.wikipedia.org/wiki/2011_Tohoku_earthquake_and_tsunami

Table 2: Example topics with CPM topical words, CPM topical hashtags and HPM topical words

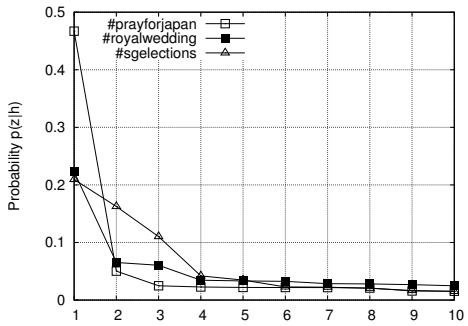
Topic label	Topic type	Top-10 words/hashtags with highest generative probability
job	CPM Hashtag	#job #jobs #career #interview #hr #jobhunt #jobsearch #sg #interviews #recruit
	CPM Word	questions interviewer job difference photo seeking using hour building rapport
	HPM Word	job singapore basic based industry questions executive days interviewer sales
singapore election	CPM Hashtag	#sgpresident #sgelections #singapore #sgelection #sgpolitics #fb #news #cars2race #yamahmee #pe2011
	CPM Word	tan tony president cheng dr jee bock kin vote lian
	HPM Word	tan tony cheng dr jee president bock kin presidential lian
japan earthquake	CPM Hashtag	#prayforjapan #japanlife #fb #japan #tsunami #sgelections #singapore #prayfortheworld #earthquake #oscars
	CPM Word	japan life live earthquake hope join people goal please tsunami
	HPM Word	japan god please hope earthquake singapore people safe news tsunami
digital devices	CPM Hashtag	#technews #technology #singapore #apple #fb #socialmedia #google #news #simonvideo #jobs
	CPM Word	apple iphone ipad video google app social facebook android media
	HPM Word	apple iphone ipad app google android video mobile phone mac
music	CPM Hashtag	#nowplaying #replacesongnameswithcurry #replacesongnameswithbangla #np #replacebandnameswithbangla #singapore #fb #nowlistening #lastfm #thingsbrokepeopledo
	CPM Word	curry love perry katy bangla adele rock black rolling party
	HPM Word	curry love song listening mars 987 bruno perry lt3 katy
football	CPM Hashtag	#lfc #mufc #fb #manutd #ynwa #arsenal #sleague #singapore #ff #sgfootball
	CPM Word	play game win time match singapore friends united goal liverpool
	HPM Word	united win game cant team fans cup liverpool manchester arsenal
daily life	CPM Hashtag	#100factsaboutme #fml #fb #likeaboss #fail #nowplaying #justsaying #foreveralone #sosingaporean #random
	CPM Word	school lol time day sleep cant haha people gonna home
	HPM Word	school day time tomorrow homework study week gonna days doing
harry porter	CPM Hashtag	#nowwatching #replacemovienameswithbacon #nowplaying #fb #harrypotterlive #singapore #glee #trueblood #royalwedding #replacemovienameswithvoldemort
	CPM Word	watching bacon harry watch potter love episode season movie cant
	HPM Word	bacon pancakes watching harry potter pants thinking voldemort green investigators

are major exogenous events in our data set. Jobs, music and daily life are example continuous endogenous topics discussed in Twitter. In short, CPM and HPM are able to explore both exogenous topics and endogenous topics.

Next we discuss the relationship between word topic and hashtag topic generated by CPM (see rows labeled by "CPM Hashtag" and "CPM Word" in Table 2). Observe that most top-ranked hashtags of hashtag topic are well associated semantically with the corresponding word topics. Take the first topic job as an example, the top-ranked hashtags (*i.e.,* #job, #jobs, #career, #interview) and the top-ranked words (*i.e.,* questions, interviewer, job, difference) are closely associated semantically. Generally speaking, topical words of CPM are relatively more specific while topical hashtags of CPM are more general. However, because a topic is usually annotated by few dominant hashtags only, the top-10 hashtags listed for each topic in Table 2 might not all describe the corresponding topic. For instance, hashtag #oscars is not very relevant to Japan earthquake and #royalwedding is irrelevant to Harry Porter movie. Some hashtags are extremely popular (*e.g.,* #fb, #singapore) and are often used to annotate many different topics.

HPM only generates word topics (see rows labeled by "HPM Word" in Table 2). Some of the top-ranked topical words of CPM and HPM are very similar. The topic labeled digital devices is an example. However, HPM discovers several topics which can not be found in CPM, listed in Table 3. These topics include royal wedding, food, business, and shopping. HPM is more powerful in finding less popular topics like business and shopping, which also partially explains why the perplexity of HPM is better than that of CPM.

Next, we show the topic distribution generated by HPM for three example hashtags: #sgelections, #royalwedding and #prayforjapan. For each example hashtag, Figure 4(a) lists their top-10 topics ranked by probability $p(z|h)$ in descending order. Observe that both #prayforjapan and #royalwedding were popular for about two weeks, a relatively short time period. For each of the two hashtags, there is one dominant topic, with the highest probability. For instance, the probability of the top topic for #prayforjapan is nearly



(a) Top-10 topics to each hashtag ranked by probability $p(z|h)$ in descending order

Hashtag	$p(z\|h)$	Top-5 Keywords
#sgelections	0.209	pap vote grc aljunied wp
	0.163	pap rally grc tan wp
	0.110	lee pap singapore pm minister
#royalwedding	0.223	wedding kate club quay clarke
	0.065	sleep time cant gonna watch
	0.060	trending lol omg lt3 happy
#prayforjapan	0.467	japan god please hope earthquake
	0.050	love people life youre time
	0.025	news tan reuters singapore japan

(b) Topical keywords of the top-3 topics for each hashtag

Figure 4: Topic distribution of top-3 most relevant topics

50%. For #sgelections, it was popular for a few months and was adopted to annotate tweets for two elections (parliamentary general election and presidential election). Three topics are observed to have high probabilities for this hashtag. For all the three example hashtags, the topical keywords of the top-3 topics with the highest probabilities are listed in Figure 4(b).

5. APPLICATIONS

In this section, we present two applications as case studies to illustrate the effectiveness of our models in addressing practical

problems in Twitter. We first motivate the two problems, namely *Retrospective Hashtag Annotation* and *Related Hashtag Discovery*, and then present experimental results.

5.1 Retrospective Hashtag Annotation

Hashtag facilitates tweet search and information diffusion. However, only about 10% of tweets are annotated by hashtags, observed from our data and also reported in other studies [11]. As surveyed in Section 2, many studies have been carried out on hashtag recommendation. Most hashtag recommendation methods target on online recommendation (*i.e.,* to recommend one or more hashtags when a user posts a new tweet) because Twitter is widely accepted as a real-time media. However, the historical data accumulated in Twitter remains an important and rich information source for more advanced tweet search options and other applications like retrospective event detection. Annotating historical tweets also helps to finding relevant tweets for less popular hashtags. As a case study, we evaluate the effectiveness of our models in *Retrospective Hashtag Annotation* which aims to annotate existing tweets without hashtags. More specifically, given a tweet d published by user u at time t, the task of retrospective hashtag annotation is to annotate this tweet with the most appropriate hashtag(s). That is, we recommend hashtags to historical tweets. Next, we present the baseline method proposed in [13] and discuss the solutions using our models.

Baseline methods: CF and CFU. A collaborative-filtering (CF) based method proposed in [13] recommends hashtags to a tweet by considering both the tweet content and the user. Given a tweet d, the method finds the top-x most similar tweets with hashtags by content similarity (*e.g.,* cosine similarity). The most frequent hashtags used by these top-x tweets are recommended. We name this method the CF method. The authors in [13] also propose a method which considers user factor, which we call the CFU method. In CFU, each user is represented as a hashtag vector. This hashtag vector is weighted by the $TF-IDF$ scheme where the TF is the number of times this user has used a hashtag in all her tweets, and IDF is computed from the number of distinct users who have used this hashtag. With this hashtag vector, the top-y most similar users to a user u are retrieved. Then the hashtag to be recommended to a tweet d by user u is based on (i) the number of times a hashtag is used to annotate the top-x most similar tweets (from all users), and (ii) the number of times a hashtag has been adopted by the top-y most similar users.

Our proposed methods: CFU+CPM and CFU+HPM. Both CPM and HPM model the three factors user, time, and tweet content in hashtag annotation. Given a tweet d written by user u at time t, the two models are able to directly estimate $p(h|u,t,\boldsymbol{w}_d)$. The most straightforward method for retrospective hashtag annotation is therefore to rank hashtags by this probability. This method, however, delivers poorer accuracy than the baseline methods. The reason is that many hashtags are under-represented because of their very limited usage in tweets. Recall that, the usage of hashtag follows a power-law like distribution (see Figure 2(a)) and most hashtags are used to annotate a small number of tweets, making the estimation $p(h|u,t,\boldsymbol{w}_d)$ less accurate for these hashtags.

To address this issue, we combine the recommendation by our models and the recommendation by the baseline methods. Generally speaking, the combined method recommends hashtags by considering both the global factors (*i.e.,* the latent relationship between hashtag and user, time, and tweet content based on our models) and the local factors (*i.e.,* the most similar tweets and most similar users based on the baseline methods). In this following, we use $CFU+CPM$ as an example to illustrate the combined method.

Let r_h be the number of times a hashtag h is recommended by the baseline method CFU for tweet d. Let $p_n(h|u,t,\boldsymbol{w}_d)$ be the normalized recommendation score from CPM:

$$p_n(h|u,t,\boldsymbol{w}_d) = \frac{p(u,t,\boldsymbol{w}_d,h)}{\sum_{h'} p(u,t,\boldsymbol{w}_d,h')}$$

where the joint probability $p(u,t,\boldsymbol{w}_d,h) = \sum'_z p(u,t,z,\boldsymbol{w}_d,h)$ and $p(u,t,z,\boldsymbol{w}_d,h)$ can be estimated with Equation 2 by replacing $\boldsymbol{h}_d$ with h in the equation. For HPM, $p(u,t,\boldsymbol{w}_d,h)$ is computed in a similar manner based on the joint probability $p(u,t,\boldsymbol{h}_d,z,\boldsymbol{w}_d)$ defined in Equation 11.

The recommendation score of hashtag h, denoted by $Score(h)$, by the combined method $CFU+CPM$ is:

$$Score(h) = \log(r_h+1) \times p_n(h|u,t,\boldsymbol{w}_d) \tag{15}$$

In the above equation, the logarithm function is introduced to reduce the impact of extremely popular hashtags. Note that, if a hashtag h does not receive any recommendation from CFU, then $Score(h)=0$ and this hashtag will not be recommended.

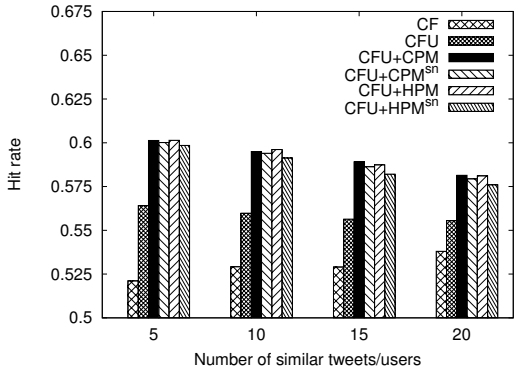
Experimental Setting. We randomly select 200,000 tweets as test set and the hashtags adopted by these tweets are considered as the ground truth. We use *Hit Rate* to evaluate the annotation accuracy. Given a tweet, a hit occurs if at least one of the top-n recommended hashtags matches the ground truth hashtags of the tweet. The hit rate for a method is computed by the number of hits divided by the number of test tweets. We report the hit rate for top-5 and top-10 recommendations for all methods. We evaluated six methods in total: CF, CFU, $CFU+CPM$, $CFU+CPM^{sn}$, $CFU+HPM$, and $CFU+HPM^{sn}$.

Experimental Results. Recall that in CFU, top-x most similar tweets and top-y most similar users are retrieved for hashtag recommendation. In our experiments, we set x and y to be the same and evaluated 4 settings: $x=y=5$, 10, 15, or 20. The hit rates of top-5 and top-10 recommendations are reported in Figures 5(a) and 5(b) respectively for the six methods. We make the following three observations from the results.

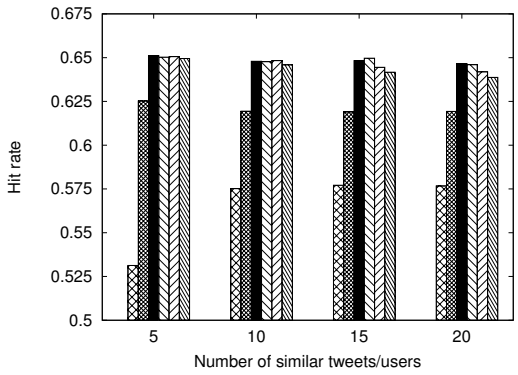
First, for both top-5 and top-10 hashtag recommendations, the methods with either CPM or HPM perform better than both baseline methods CF and CFU. In particular, in terms of hit rate for top-5 hashtag recommendation with 5 similar tweets/users, $CFU+CPM$ outperforms CFU by 6.72% and CF by 14.34% respectively. We also observe that $CFU+HPM$ yields very similar results as $CFU+CPM$, despite that HPM achieves better perplexity than CPM in our earlier experiments.

Second, $CFU+CPM^{sn}$ performs slightly worse than $CFU+CPM$ and the same observation holds for $CFU+HPM^{sn}$ against $CFU+HPM$. In other words, considering social network regularization does not improve the hit rate for hashtag recommendation. One possible reason is that the social network regularization introduces noises in estimating $p(z|u)$. Consequently, the poorer estimation of $p(z|u)$ results in less accurate $p(h|u,t,\boldsymbol{w}_d)$. This result is consistent with the results reported in Section 4.2 where the considering social network regularization leads to poorer perplexity to both models.

Third, evaluated by hit rate of top-5 hashtag recommendation, the hit rate for all methods decreases along with increasing the number of similar tweets/users. This observation suggests a larger number of similar tweets/users likely brings in irrelevant hashtags to the given tweet, particularly when the ground truth hashtag is an infrequent hashtag. Recall that hashtag frequency distribution follows a power-law like distribution and a large number of hashtags appear only 5 times in our dataset (see Section 4.1).



(a) Hit rate of top-5 hashtags



(b) Hit rate of Top-10 hashtags

Figure 5: Hit rate of the four methods for top-5/top-10 hashtags

5.2 Related Hashtags Discovery

Hashtags are chosen by Twitter users from an uncontrolled vocabulary. For the same event or the same topic, multiple hashtags might be chosen by users, *e.g.,* #cikm, #cikm14,or #cikm2014 for the same conference. Hashtags might also be related because of other types of relationships such as subsumption relation. For example the hashtag #sgelections has been used to annotate tweets related to both the Singapore Parliamentary General Election[7] in May 2011 and the Singaporean Presidential Election[8] in August 2011, while a more specific hashtag #sgpresident was also widely adopted for the latter. Discovering related hashtags helps users in refining, extending or reformulating hashtag-based queries.

Specifically, given a hashtag h and hashtag vocabulary E, related hashtag discovery is to locate the top-n hashtags from E (without h itself) that are most related to h. In this set of experiments, we evaluate four methods for their effectiveness in finding most related hashtags of a given hashtag.

Co-occurrence (COO). A straightforward method in finding the related hashtags is through co-occurrence. If a hashtag h' often co-occurs with the given hashtag h in tweet annotation, then h' is believed to be related with h.

Content-based Similarity (CBS). Two hashtags are related if they share similar semantic meanings defined by the sets of tweets annotated by them. Given a hashtag h, all tweets annotated by h combined together form a virtual document. Then the similarity between two hashtags is computed based on the cosine similarity of the two corresponding virtual documents.

Content- and Topic-based Similarity (CTS). In this method, we use the topic-based feature representation to enhance the hashtag similarity computation. More specifically, each hashtag can be represented by a topic vector, where each dimension is one of the K topics and is weighted by $p(z_i|h)$, $0 \leq i \leq K$. Let $S_c(h,h')$ be the content-based similarity between hashtags h and h' computed in CBS, and let $S_t(h,h')$ be the cosine similarity between the topic vector representations of the two hashtags. The CTS similarity between the two hashtags is: $S_{ct}(h,h') = \eta \times S_c(h,h') + (1-\eta) \times S_t(h,h')$, where η is a parameter for the combination. The following question is: how to compute $p(z|h)$ using the two models?

- In CPM, $p(z|h) = \frac{\sum_{d \in D_h} p(z|d)}{|D_h|}$ where $p(z|d)$ is computed using Equation 6.
- In HPM, $p(z|h)$ is estimated directly from the model (see Equation 12).

To summarize, we have four methods for evaluation: COO, CBS, CTS_{CPM}, and CTS_{HPM} where for the latter two CPM and HPM denote the model for computing the topic vector for hashtags.

To evaluate the effectiveness of the four methods in finding related hashtags, we randomly selected 50 hashtags among the top-500 most popular hashtags to be the query hashtags.[9] For each of the 50 query hashtags, a method returns the top-5 most related hashtags for manual assessment. In CTS, η is set to 0.6 in our experiments based on observations using a few sample hashtags (not included in the 50 query hashtags). We employ three volunteers to label the relatedness of the top-5 hashtags returned by each method and each hashtag receives a binary score: 0 for not-related and 1 for related. The kappa scores of the agreement between any pair of the volunteers are reported in Table 4. The average kappa score is 0.723 suggesting substantial agreement between our volunteers.

Table 4: Kappa scores between three pairs of volunteers (v's)

Volunteer pair	$\langle v_1, v_2 \rangle$	$\langle v_1, v_3 \rangle$	$\langle v_2, v_3 \rangle$	Average
Kappa score	0.809	0.687	0.672	0.723

Table 5: Precision for related hashtag discovery with the best result in boldface

Method	COO	CBS	CTS_{CPM}	CTS_{HPM}
Precision	0.520	0.681	0.705	**0.729**

The average precision for the 50 query hashtags from the three volunteers is reported in Table 5. Observe that COO results in the poorest precision. This is because 82.2% of tweets each is annotated with only one hashtag (see Section 4.1). Consequently, there might be too few co-occurring hashtags for a given query hashtag. Among the other three methods, which utilize content similarity, CTS_{CPM} and CTS_{HPM} outperform the method not using topic vector. This demonstrates that the effectiveness of using topic vector as additional information in enhancing related hashtag discovery. Observe that CTS_{HPM} achieves the highest precision, probably because HPM discovers more meaningful topics reflected by the lowest perplexity (see Section 4.2).

We now use two examples hashtags #prayforjapan and #movies to illustrate the difference between the most related hashtags found

[7] http://en.wikipedia.org/wiki/Singaporean_general_election,_2011

[8] http://en.wikipedia.org/wiki/Singaporean_presidential_election,_2011

[9] Popular hashtags are expected to have higher chances of being co-occurred with other hashtags.

Table 6: Top-5 most related hashtags to #prayforjapan and #movies, discovered by the four methods

Method	Top related hashtags to #prayforjapan
COO	#japan #prayfortheworld #tsunami #fb #sleague
CBS	#japan #tsunami #quake #fukushima #japans
CTS_{CPM}	#prayfortheworld #japan #helpjapan #godblessjapan #quake
CTS_{HPM}	#prayfortheworld #helpjapan #japan #godblessjapan #quake
-	**Top related hashtags to #movies**
COO	#imdb #singapore #sg #singaporean #film
CBS	#imdb #celebrity #gossip #xinmsn #ryanreynolds
CTS_{CPM}	#imdb #movie #mfpgossip #seattle #eastboundanddown
CTS_{HPM}	#imdb #movie #sgfilm #trailer #video

by the four methods, listed in Table 6. Among the top-5 most related hashtags for #prayforjapan found by COO, #fb and #sleague are not related. All the remaining three methods CBS, CTS_{CPM}, and CTS_{HPM} are able to find related hashtags for #prayforjapan. Interestingly, the two methods with topic-level representation recommend the same set of hashtags in slightly different orders. Another example is #movies. All top-5 hashtags by CTS_{HPM} are relevant to #movie. The hashtags from the other three methods all contain some irrelevant hashtags such as #sg, #xinmsn and #eastboundanddown.

6. CONCLUSION AND FUTURE WORK

In this paper, we propose two PLSA-style topic models to model the latent relationship between tweet content, user interest, time, and hashtag at topic-level. We also evaluate the impact of considering social network regularization based on mention relationship in Twitter. Through extensive experiments, we show that Hashtag-Pivoted Model outperforms Content-Pivoted Model in terms of perplexity measure. We also show that the social network regularization based on mention relationship hurts the performance of both models. We further demonstrate the effectiveness of the two models in addressing two practical applications (*i.e.,* retrospective hashtag annotation and related hashtag discovery). The utilization of both models improves the effectiveness in addressing both applications compared to their corresponding baselines.

Recall that the two models follow different assumptions to simulate the two possible generation processes of hashtag and tweet content. However, given a tweet, there is no mechanism to predict which model best reflects the generation process between its hashtag and content. Research on such predicting mechanism is part of our future work. Another piece of future work is to evaluate the impact of social network regularization to the models based on other types of user relationships other than mention relationship. Furthermore, we will continue to apply our models to practical applications in tweets such as hashtag summarization.

7. REFERENCES

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

[2] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *COLING*, pages 241–249. ACL, 2010.

[3] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim. Finding bursty topics from microblogs. In *ACL*, pages 536–544. ACL, 2012.

[4] E. Diaz-Aviles, L. Drumond, L. Schmidt-Thieme, and W. Nejdl. Real-time top-n recommendation in social streams. In *RecSys*, pages 59–66. ACM, 2012.

[5] W. Feng and J. Wang. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In *KDD*, pages 1276–1284. ACM, 2012.

[6] F. Godin, V. Slavkovikj, W. De Neve, B. Schrauwen, and R. Van de Walle. Using topic models for twitter hashtag recommendation. In *WWW companion*, 2013.

[7] Z. Guan, J. Bu, Q. Mei, C. Chen, and C. Wang. Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. In *SIGIR*, pages 540–547. ACM, 2009.

[8] J. Guo, X. Cheng, G. Xu, and X. Zhu. Intent-aware query similarity. In *CIKM*, pages 259–268. ACM, 2011.

[9] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57. ACM, 1999.

[10] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *WWW*, pages 769–778. ACM, 2012.

[11] L. Hong, G. Convertino, and E. H. Chi. Language matters in twitter: A large scale study. In *ICWSM*, 2011.

[12] B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1), 2009.

[13] S. M. Kywe, T.-A. Hoang, E.-P. Lim, and F. Zhu. On recommending hashtags in twitter networks. In *SocInfo*, pages 337–350. Springer-Verlag, 2012.

[14] H. Liang, Y. Xu, D. Tjondronegoro, and P. Christen. Time-aware topic recommendation based on micro-blogs. In *CIKM*, pages 1657–1661. ACM, 2012.

[15] Z. Ma, A. Sun, and G. Cong. On predicting the popularity of newly emerging hashtags in twitter. *JASIST*, 64(7):1399–1410, 2013.

[16] A. Mazzia and J. Juett. Suggesting hashtags on twitter.

[17] M. Naaman, H. Becker, and L. Gravano. Hip and trendy: Characterizing emerging trends on twitter. *J. Am. Soc. Inf. Sci. Technol.*, 62(5):902–918, 2011.

[18] R. M. Neal and G. E. Hinton. Learning in graphical models. chapter A view of the EM algorithm that justifies incremental, sparse, and other variants, pages 355–368. MIT Press, 1999.

[19] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical recipes in C: the art of scientific computing*. Cambridge University Press, 1988.

[20] RadiumOne. #mobile hashtag survey, 2013. `http://radiumone.com/about/company-resources.html#research`.

[21] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing microblogs with topic models. In *ICWSM*. The AAAI Press, 2010.

[22] S. Rendle, L. Balby Marinho, A. Nanopoulos, and L. Schmidt-Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *KDD*, pages 727–736. ACM, 2009.

[23] S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *WSDM*, pages 81–90. ACM, 2010.

[24] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *WWW*, pages 695–704. ACM, 2011.

[25] S. Sedhai and A. Sun. Hashtag recommendation for hyperlinked tweets. In *SIGIR*, pages 831–834, 2014.

[26] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. Tag recommendations based on tensor dimensionality reduction. In *RecSys*, pages 43–50. ACM, 2008.

[27] O. Tsur and A. Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *WSDM*, pages 643–652. ACM, 2012.

[28] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *CIKM*, pages 1031–1040. ACM, 2011.

[29] L. Yang, T. Sun, M. Zhang, and Q. Mei. We know what @you #tag: does the dual role affect hashtag adoption? In *WWW*, pages 261–270. ACM, 2012.

[30] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical topic discovery and comparison. In *WWW*, pages 247–256. ACM, 2011.

[31] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann. Who, where, when and what: discover spatio-temporal topics for twitter users. In *KDD*, pages 605–613. ACM, 2013.

[32] E. Zangerle, W. Gassler, and G. Specht. Recommending#-tags in twitter. In *SASWeb*, volume 730, pages 67–78.

[33] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *ECIR*, pages 338–349. Springer-Verlag, 2011.