Hashtag Recommendation Using Dirichlet Process Mixture Models Incorporating Types of Hashtags

Yeyun Gong, Qi Zhang, Xuanjing Huang
Shanghai Key Laboratory of Data Science
School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, P.R.China
{12110240006, qz, xjhuang}@fudan.edu.cn

Abstract

In recent years, the task of recommending hashtags for microblogs has been given increasing attention. Various methods have been proposed to study the problem from different aspects. However, most of the recent studies have not considered the differences in the types or uses of hashtags. In this paper, we introduce a novel nonparametric Bayesian method for this task. Based on the Dirichlet Process Mixture Models (DPMM), we incorporate the type of hashtag as a hidden variable. The results of experiments on the data collected from a real world microblogging service demonstrate that the proposed method outperforms state-of-the-art methods that do not consider these aspects. By taking these aspects into consideration, the relative improvement of the proposed method over the state-of-the-art methods is around 12.2% in F1- score.

1 Introduction

Hashtags are used to mark keywords or topics in a microblog. Over the past few years, social media services have become some of the most important communication channels for people. According to the statistic reported by the Pew Research Centers Internet & American Life Project in Aug 5, 2013, about 72% of adult internet users are also members of at least one social networking site. Hence, microblogs have also been widely used as data sources for public opinion analyses (Bermingham and Smeaton, 2010; Jiang et al., 2011), prediction (Asur and Huberman, 2010; Bollen et al., 2011), reputation management (Pang and Lee, 2008; Otsuka et al., 2012), and many other applications (Sakaki et al., 2010; Becker et al., 2010; Guy et al., 2010; Guy et al., 2013).

In addition to the limited number of characters in the content, microblogs also contain a form of metadata tag (hashtag), which is a string of characters preceded by the symbol (#). Hashtags are used to mark the keywords or topics of a microblog. They can occur anywhere in a microblog, at the beginning, middle, or end. Hashtags have been proven to be useful for many applications, including microblog retrieval (Efron, 2010), query expansion (A.Bandyopadhyay et al., 2011), and sentiment analysis (Davidov et al., 2010; Wang et al., 2011). However, only a small percentages of microblogs contain hashtags provided by their authors. Hence, the task of recommending hashtags for microblogs has become an important research topic and has received considerable attention in recent years. Existing works have studied discriminative models (Ohkura et al., 2006; Heymann et al., 2008) and generative models (Blei and Jordan, 2003; Krestel et al., 2009; Ding et al., 2013; Godin et al., 2013) based on the textual information of a single microblog.

Since microblog users are free to develop and use their own hashtags, they may select hashtags for different purposes. Based on an analysis of the hashtags crawled from a real online service, we observe that hashtags are used for events, conferences, conversation, disasters, memes, recall, quotes, and so on. To illustrate it let us take the following examples:

Example 1:*#Apple_iOS_9 includes music feature, new security and support for older iPhones.*

Example 2:*#BREAKING: Missing cyclist Natalie Donoghue has been found alive after she went missing in the Hunter Valley.*

We can see that the hashtag *#Apple_iOS_9* used in the example summarize the main topics of the corresponding microblog. While, the aim of hashtag *#BREAKING* in the example 2 is used as a label of the microblog. The different uses greatly

impact the strategy of hashtag recommendation. However, there has been relatively few studies which take this issue into consideration.

In this paper, we propose a novel nonparametric Bayesian method to perform this problem. Inspired by the methods proposed by Liu et al. (2012), we assume that the hashtags and textual content in the corresponding microblog are parallel descriptions of the same thing in different languages. We adapt a translation model with topic distribution to achieve this task. Because of the ability of Dirichlet Process Mixture Models (DPMM) (Antoniak and others, 1974; Ferguson, 1983) to handle an unbounded number of topics, the proposed method is extended from them. Based on the different uses of hashtags, we incorporate the type of hashtag into the DPMM as a hidden variable.

The main contributions of this work can be summarized as follows:

- Through analyzing the microblogs, we propose the problem of influences of types of hashtags.
- We adopt a nonparametric Bayesian method to perform the hash tag recommendation task, which also takes the types of hashtags into consideration.
- Experimental results on the dataset we construct from a real microblogging service show that the proposed method can achieve significantly better performance than the state-of-the-arts methods.

2 The Proposed Method

In this section, we first give some brief descriptions about the Dirichlet process (DP) and Dirichlet Process Mixture Models (DPMM). Then, we detail the proposed hashtag recommendation method.

2.1 Preliminaries

2.1.1 Dirichlet Process

The Dirichlet process (DP) is a distribution over distributions. A DP, denoted by $G \sim DP(\alpha, H)$, is parameterized by a base measure H, and a concentration parameter α. After a discussion of basic definitions, we present two different perspectives on the Dirichlet process.

A perspective on the Dirichlet process is stick-breaking construction. The stick-breaking construction considers a probability mass function $\{\beta_k\}_{k=1}^{\infty}$ on a countably infinite set, where the discrete probabilities are defined as follows:

$$\begin{aligned} v_k|\alpha &\sim Beta(1,\alpha) \\ \beta_k &= v_k \prod_{l=1}^{k-1}(1-v_l). \end{aligned} \tag{1}$$

The k^{th} weight is a random proportion v_k of the remaining stick after the previous$(k-1)$ weights have been defined. This stick-breaking construction is generally denoted by $\beta \sim GEM(\alpha)$ (GEM stands for Griffiths, Engen and McCloskey). A random draw $G \sim DP(\alpha, H)$ can be expressed as:

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \quad \theta_k|\alpha, H \sim H, \tag{2}$$

where δ_θ is a probability measure concentrated at θ.

A second perspective on the Dirichlet process is provided by the *Pólya urn scheme* (Blackwell and MacQueen, 1973). It refers to draws from G. Let $\theta_1, \theta_2, ...$ represent a sequence of independent and identically distributed (i.i.d.) random variables distributed according to G. Blackwell and MacQueen (1973) showed that the conditional distributions of θ_i given $\theta_1, ..., \theta_{i-1}$ have the following form:

$$\begin{aligned} &\theta_i|\theta_1, ..., \theta_{i-1}, \alpha, H \\ &\sim \sum_{j=1}^{i-1} \frac{j}{i-1+\alpha}\delta_{\theta_j} + \frac{\alpha}{i-1+\alpha}H. \end{aligned} \tag{3}$$

Eq.(3) shows that θ_i has positive probability of being equal to one of the previous draws. We use $\phi_1, ..., \phi_K$ to represent the distinct values taken on by $\theta_1, ..., \theta_{i-1}$, and Eq.(3) can be re-expressed as:

$$\begin{aligned} &\theta_i|\theta_1, ..., \theta_{i-1}, \alpha, H \\ &\sim \sum_{k=1}^{K} \frac{m_k}{i-1+\alpha}\delta_{\theta_{\phi_k}} + \frac{\alpha}{i-1+\alpha}H, \end{aligned} \tag{4}$$

where m_k is the number of values $\theta_{i'} = \phi_k$ for $1 \leq i' < i$.

2.1.2 Dirichlet Process Mixture Models

In nonparametric Bayesian statistics, DPs are commonly used as prior distributions for mixture

models with an unknown number of components. Let $F(\theta_i)$ denotes the distribution of the observation x_i given θ_i. We can get the observation x_i as follows:

$$\theta_i|G \sim G$$
$$x_i|\theta_i \sim F(\theta_i).$$

Given $G \sim DP(\alpha, H)$, each observation x_i from an exchangeable data set $\mathbf{x}$ is generated by first choosing a parameter $\theta_i \sim G$, and then sampling $x_i \sim F(\theta_i)$. This model is referred to as a Dirichlet process mixture model. This process is often described by a set $\mathbf{z}$ of independently sampled variables $z_i \sim Mult(\beta)$ indicating the component of the mixture $G(\theta)$ associated with each data point $x_i \sim F(\theta_{z_i})$. Then we can get:

$$z_i|\beta \sim Mult(\beta)$$
$$x_i|\{\theta_k\}_{k=1}^{\infty}, z_i \sim F(\theta_{z_i}).$$

2.2 DPMM Based Hashtag Recommendation

2.2.1 The Generation Process

Let D represent the number of microblogs in the given corpus. A microblog contains a bag of words denoted by $w_d = \{w_{d_1}, w_{d_2}, ..., w_{d_{N_d}}\}$, where N_d is the total number of terms in the microblog. A word is defined as an item from a vocabulary with W distinct words indexed by $w = \{w_1, w_2, ..., w_W\}$. Each microblog may have a number of hashtags denoted by $h_d = \{h_{d_1}, h_{d_2}, ..., h_{d_{M_d}}\}$. M_d is the number of hashtags of microblog d. Each hashtag is from the vocabulary with V distinct hashtags indexed by $h = \{h_1, h_2, ..., h_V\}$. Given an unlabeled data set, the task of hashtag recommendation is to discover a list of hashtags for each microblog.

In standard LDA, each document is viewed as a mixture of topics, and each topic has probabilities to generate words. A LDA is a generalization of a finite mixture model. Since DP is the extension of finite mixture models to the nonparametric setting, the appropriate tool for nonparametric topic models is HDP. However, both LDA and HDP are normally suitable for long documents. For microblogs, which have limited number of words, a single microblog is most likely to talk about a single topic. Hence, in this work, we regard that each microblog associates with only one topic. The set of documents are viewed as a mixture of infinite topics. And we use DPs as prior distributions for the mixture of infinite topics.

The main assumptions of our model are as follows. When user u publishes a microblog, he will first generate the content and then generate the hashtags. When constructing the content, he will select a topic based on the topic distribution. Then he will choose a bag of words one by one from the word distribution of the topic or from the background words that captures white noise. Hashtags will be chosen according to the following two situations. In the first situation, hashtags summarize the corresponding microblogs. Hashtags of a microblog can be generated from the content through the topic-specific alignment probability between words and hashtags. In the second situation, hashtag is used as a label of the microblog. We recommend the hashtags using the words in the microblog, which is based on the frequency of words regarded as this type of hashtag.

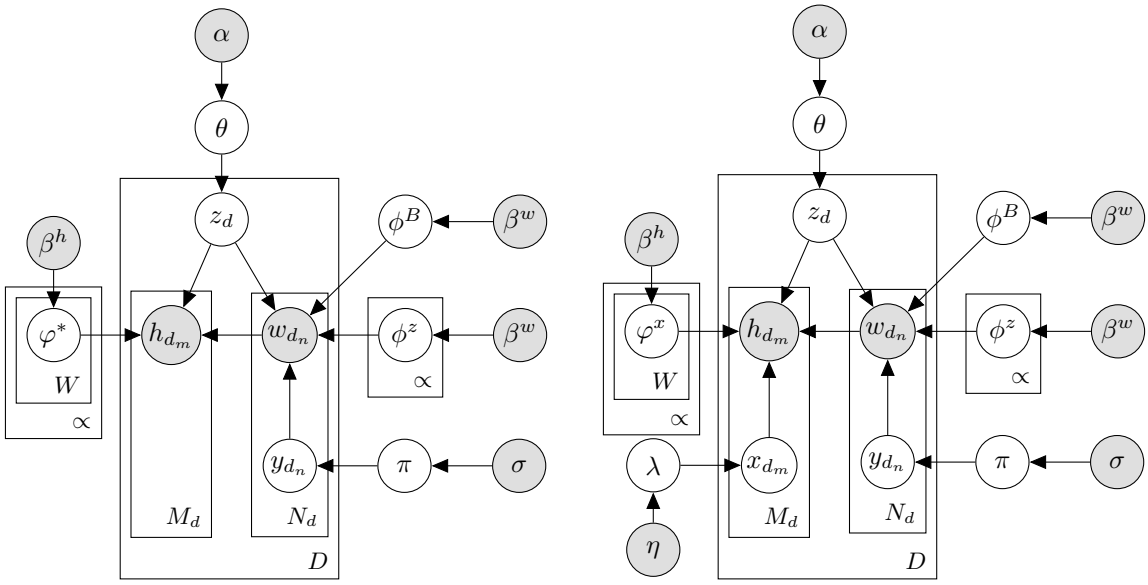
Let π be the probability of choosing a topic word or a background word, and we use $y_d = \{y_{d_n}\}_{n=1}^{N_d}$ to indicate a word to be a topic word or background word. θ denotes the topic distribution, and ϕ^k represents the word distribution for topic k. ϕ^B represents the word distribution for background words. We use x_{d_m} to represent the type of hashtag h_{d_m}, and use z_d to represent the topic of document d. Then each hashtag h_{d_m} is annotated according to the translation possibility $P(h_{d_m}|w_d, z_d, x_{d_m}, \varphi^{x_{d_m}})$, where $\varphi^{x_{d_m}}$ is the probability alignment table between words and hashtags. The generation process is as Algorithm 1.

Figure 1(a) shows the graphical representation of the generation process in Algorithm 1. Figure 1(b) is the graphical model which does not take the types of hashtags into consideration, where $\varphi^* \in \{\varphi^1, \varphi^2\}$. If $\varphi^* = \varphi^1$, the model is just considering the first situation. when $\varphi^* = \varphi^2$, only the second type of hashtag will be considered.

2.2.2 Learning

We use collapsed Gibbs sampling (Griffiths and Steyvers, 2004) to obtain samples of hidden variables assignment and to estimate the model parameters from these samples.

The sampling probability of being a topic/background word for the nth word in the microblog d can be calculated by the following



(a) CNHR (b) NHR*

Figure 1: The graphical representation of the proposed model. Shaded circles are observations or constants. Unshaded ones are hidden variables. *CNHR* represents the proposed hashtag recommendation method. *NHR** represents the model which does not take the types of hashtags into consideration.

Algorithm 1 The generation process of CNHR

Draw $\pi \sim Beta(\sigma)$, $\lambda \sim Beta(\eta)$
Draw background word distribution $\phi^B \sim Dir(\beta^w)$
Draw $\theta|\alpha \sim GEM(\alpha)$
for each microblog $d = 1, 2, ..., D$ **do**
  Draw $z_d \sim Mul(\theta)$
  Draw word distribution $\phi^{z_d} \sim Dir(\beta^w)$
  for each word $n = 1, ..., N_d$ **do**
    Draw $y_{d_n} \sim Ber(\pi)$
    if $y_{d_n} = 0$ **then**
      Draw a word w_{d_n} from the background-word distribution $w_{d_n} \sim Mul(\phi^B)$
    else
      Draw a word w_{d_n} from the topic-word distribution $w_{d_n} \sim Mul(\phi^{z_d})$
    end if
  end for
  for each hashtag $m = 1, ..., M_d$ **do**
    Draw $x_{d_m} \sim Ber(\lambda)$
    Draw $\varphi^{x_{d_m}} \sim Dir(\beta^h)$
    Draw a hashtag $h_{d_m} \sim P(h_{d_m}|w_d, z_d, x_{d_m}, \varphi^{x_{d_m}})$
  end for
end for

equation:

$$p(y_{d_n}|\mathbf{w}, \mathbf{h}, \mathbf{z}, \mathbf{y}_{\neg \mathbf{d_n}}, \sigma, \beta^w) \propto \frac{N_{\neg n,p} + \sigma}{N_{\neg n,(.)} + 2\sigma} \cdot \frac{N^{w_{d_n}}_{\neg n,l} + \beta^w}{N^{(.)}_{\neg n,l} + \beta^w W}, \quad (5)$$

where $l = B$ when $p = 0$ and $l = z_d$ when $p = 1$, $N_{\neg n,p}$ is a count of words that are assigned to background words and any topic respectively, $N^{w_{d_n}}_{\neg n,B}$ is the number of word w_{d_n} assigned to background words, $N^{w_{d_n}}_{\neg n,z_d}$ is the number of word w_{d_n} that are assigned to topic z_d. All counters are calculated with the current word w_{d_n} excluded.

We sample z_d for the microblog d using the following equation:

$$p(z_d|\mathbf{w}, \mathbf{h}, \mathbf{z}_{\neg \mathbf{d}}, \mathbf{y}, \mathbf{x}, \alpha, \beta^w, \beta^h) \propto p(z_d|\mathbf{z}_{\neg \mathbf{d}}, \alpha) \cdot p(\mathbf{w_d}|\mathbf{z}, \mathbf{w}_{\neg \mathbf{d}}, \mathbf{y}, \beta^w) \cdot p(\mathbf{h_d}|\mathbf{z}, \mathbf{w_d}, \mathbf{y}, \mathbf{x}, \beta^h). \quad (6)$$

We can also represent $p(z_d|\mathbf{z}_{\neg \mathbf{d}}, \alpha)$ with CRP as described in the previous section. Since $z_1, z_2, ...$ is a sequence of i.i.d random variables, they are exchangeable. Let us consider the dth variable z_d is the last observation, we can get the following expression:

$$p(z_d|\mathbf{z}_{\neg \mathbf{d}}, \alpha) \sim \sum_k^K \frac{N^k_{\neg d}}{N^{(.)}_{\neg d} - 1 + \alpha} \delta(z_d, k) + \frac{\alpha}{N^{(.)}_{\neg d} - 1 + \alpha} \delta(z_d, \bar{k}), \quad (7)$$

where k is an exist topic and $\bar{k}$ is a new topic, $N^k_{\neg d}$ is the number of microblogs assigned with topic k, $N^{(.)}_{\neg d}$ is the total number of microblogs, α is concentration parameter. All counters are

calculated with the current microblog d excluded.

If z_d equals an exist topic $z_d = k$, then we can calculate $p(\mathbf{w_d}|\mathbf{z}, \mathbf{w}_{\neg\mathbf{d}}, \mathbf{y}, \beta^w)$ by:

$$p(\mathbf{w_d}|\mathbf{z}, \mathbf{w}_{\neg\mathbf{d}}, \mathbf{y}, \beta^w) = \frac{\int_{\phi_k} \widetilde{f}(\mathbf{w_d}|\phi_k) \prod_{z_j=k, j\neq d} \widetilde{f}(\mathbf{w_j}|\phi_k) h(\phi_k) d\phi_k}{\int_{\phi_k} \prod_{z_j=k, j\neq d} \widetilde{f}(\mathbf{w_j}|\phi_k) h(\phi_k) d\phi_k}, \tag{8}$$

where $\widetilde{f}(\mathbf{w_d}|\phi_k) = \prod_{1\leq n\leq N_d, y_{d_n}=1} f(w_{d_n}|\phi_k)$. N_d is the number of words in microblog d. $f(w_{d_n}|\phi_k)$ is the density of word w_{d_n} given topic k. $\mathbf{w_d}$ are the words in microblog d. $h(\phi_k)$ is the density of base measure H.

If z_d is a new topic $z_d = \bar{k}$, then we can calculate $p(\mathbf{w_d}|z_d = \bar{k}, \mathbf{w}_{\neg\mathbf{d}}, \mathbf{y}, \beta^w)$ by:

$$p(\mathbf{w_d}|z_d = \bar{k}, \mathbf{w}_{\neg\mathbf{d}}, \mathbf{y}, \beta^w) = p(\mathbf{w_d}|\beta^w) = \int_{\phi_{\bar{k}}} \widetilde{p}(\mathbf{w_d}|\phi_{\bar{k}}) h(\phi_{\bar{k}}) d\phi_{\bar{k}}, \tag{9}$$

where $\widetilde{p}(\mathbf{w_d}|\phi_{\bar{k}}) = \prod_{1\leq n\leq N_d, y_{d_n}=1} p(w_{d_n}|\phi_{\bar{k}})$.

We can calculate the probabilities of generating hashtags from two situations as follows:

$$p(\mathbf{h_d}|\mathbf{z}, \mathbf{w_d}, \mathbf{y}, \mathbf{x}, \beta^h) = \begin{cases} \prod_{m=1}^{M_d} \sum_{n\in\widetilde{N}_d} \frac{M^{k,\neg d}_{w_{d_n}, h_{d_m}} + \beta^h}{M^{k,\neg d}_{w_{d_n},(.)} + \beta^h V} & x_{d_m} = 1 \\ \prod_{m=1}^{M_d} \sum_{n\in\widetilde{N}_d, w_{d_n}=h_{d_m}} \frac{M^{k,\neg d}_{w_{d_n},2} + \beta^h}{M^{k,\neg d}_{w_{d_n},(.)} + 2\beta^h} & x_{d_m} = 2, \end{cases} \tag{10}$$

where $\widetilde{N}_d$ represent the index set of topic words($y = 1$) in the microblog d, $M^{k,\neg d}_{w_{d_n}, h_{d_m}}$ is the number of occurrences that word w_{d_n} is translated to hashtag h_{d_m} given topic k, $M^{k,\neg d}_{w_{d_n},(.)}$ is the total number of occurrences that word w_{d_n} is under topic k, $M^{k,\neg d}_{w_{d_n},2}$ is the number of word w_{d_n} recommended as the second type of hashtag given topic k. All counters with $\neg d$ are calculated with the current microblog w_d excluded.

We sample the index variable x_{d_m} for mth hashtag in the microblog d as follows:

$$p(x_{d_m}|\mathbf{z}, \mathbf{w_d}, \mathbf{y}, \mathbf{x}_{\neg\mathbf{d_m}}, \mathbf{h}_d, \beta^h) \propto \sum_{n=1}^{\widetilde{N}_d} \varphi^{x_{d_m}}_{h_{d_m}, z_d, w_{d_n}} \frac{N^{\neg d_m}_{x_{d_m}} + \eta}{N^{\neg d_m}_{(.)} + 2\eta}, \tag{11}$$

where $N^{\neg d_m}_{x_{d_m}}$ is the number of hashtags that is generated by the type x_{d_m}, $N^{\neg d_m}_{(.)}$ is total number of hashtags, the counters with $\neg d_m$ are calculated with the current hashtag excluded.

After enough sampling iterations to burn in the Markov chain, φ^1 and φ^2 are estimated as follows:

$$\varphi^1_{h,k,w} = \frac{M^k_{w,h} + \beta^h}{M^k_{w,(.)} + \beta^h V}, \varphi^2_{h,k,w} = \frac{M^k_{w,2} + \beta^h}{M^k_{w,(.)} + 2\beta^h}, \tag{12}$$

The potential size of the probability alignment φ^1 between hashtag and word is $W \cdot V \cdot K$. The data sparsity may pose a more serious problem in estimating φ^1 than the topic-free word alignment case. We use interpolation smoothing technique for φ^1. In this paper, we employ smoothing as follows:

$$\varphi^{1*}_{h,k,w} = \gamma\varphi^1_{h,k,w} + (1-\gamma)P(h|w), \tag{13}$$

where $\varphi^{1*}_{h,k,w}$ is the smoothed topical alignment probabilities, $\varphi^1_{h,k,w}$ is the original topical alignment probabilities, $P(h|w)$ is topic-free word alignment probability. In this work, we obtain $P(h|w)$ by exploring IBM model-1 (Brown et al., 1993). γ is trade-off of two probabilities ranging from 0 to 1. When $\gamma = 0$, $\varphi^{1*}_{h,k,w}$ reduces to topic-free word alignment probability, and when $\gamma = 1$, there will be no smoothing in $\varphi^{1*}_{h,k,w}$.

2.2.3 Hashtag Recommendation

Suppose given an unlabeled dataset, we firstly discover the topic and determine topic/background words for each microblog. The collapsed Gibbs sampling is also applied for inference. The process is almost same as previous section described the model learning. The different is that there are no hashtags in the unlabeled dataset. Hence, when sampling z_d for the microblog d, we use the following equation:

$$p(z_d|\mathbf{w}, \mathbf{z}_{\neg\mathbf{d}}, \mathbf{y}, \alpha, \beta^w) \propto p(z_d|\mathbf{z}_{\neg\mathbf{d}}, \alpha) \cdot p(\mathbf{w_d}|\mathbf{z}, \mathbf{w}_{\neg\mathbf{d}}, \mathbf{y}, \beta^w). \tag{14}$$

Since there are no differences between the word alignments with each hashtags for a new topic in the unlabeled dataset, after the hidden variables of topic/background words and the topic of each microblog become stable, we only need to estimate the distribution of topics exist in the training dataset. Then we can estimate the distribution of topics for the microblog d in the unlabeled data by:

$$\chi_{dk} = \frac{p(k)p(w_{d_1}|k)p(w_{d_2}|k)...p(w_{d_{N_d}}|k)}{Z}, \tag{15}$$

where $p(w_{d_n}|k) = \frac{N_k^{w_{d_n}}+\beta}{N_k^{(\cdot)}+W\beta}$ and $N_k^{w_{d_n}}$ is a count of words w_{d_n} that are assigned to topic k in the corpus. And $p(k) = \frac{N_k}{N_{(\cdot)}+\alpha}$ is regarded as a prior for topic distribution, where Z is the normalized factor. With topic distribution χ and topic-specific word alignment table φ^*, we can rank hashtags for the microblog d in the unlabeled data through the following equation:

$$p(h_{d_m}|w_d, \chi_d, \varphi^*) \propto \sum_{z_d=1}^{K}\sum_{n=1}^{N_d}\sum_{x=1}^{C} p(z_d|\chi_d)\cdot p(w_{d_n}|w_d)\cdot p(x_{d_m}) \cdot p(h_{d_m}|w_{d_n}, z_d, x_{d_m}, \varphi^{x_{d_m}*}), \quad (16)$$

where C is the number of hashtag types. $p(w_{d_n}|w_d)$ is the weight of the word w_{d_n} in the microblog content w_d, which can be estimated by the IDF score of the word, $p(x_{d_m})$ is the probability of hashtag belong to the type x_{d_m}, we can estimate it with Eq.(11). Based on the ranking scores, we can suggest the top-ranked hashtags for each microblog.

3 Experiments

3.1 Data Collection

We use a dataset collected from Sina Weibo[1], which provides the Twitter-like service and is one of the most popular one in China, to evaluate the proposed approach and alternative methods. The original data set contains 282.2 million microblogs posted by around 1.1 million users. These microblogs were obtained by starting from a set of seed users and their follower/followee relations. We extract the microblogs posted with hashtags between Jan. 2012 and July 2013. Finally, 1,118,792 microblogs posted are selected for this work. The unique number of hashtags in the corpus is 305,227. We randomly select 100K as training data, 10K as development data, and 10K as test set. The hashtags marked in the original microblogs are considered as the golden standards.

3.2 Experiment Configurations

We use precision (P), recall (R), and F1-score (F_1) to evaluate the performance. Precision is calculated based on the percentage of "hashtags truly assigned" among "hashtags assigned by system". Recall is calculated based on the "hashtags truly assigned" among "hashtags manually assigned". F_1 is the harmonic mean of precision and recall. We do 500 iterations of Gibbs sampling to train the model. For optimizing the hyperparmeters of the proposed method and alternative methods, we use development data set to do it. In this work, the scale parameter α is set to $Gamma(5, 0.5)$. The other settings of hyperparameters are as follows: $\beta^w = 0.1$, $\beta^h = 0.1$, $\eta = 0.01$, and $\sigma = 0.01$. The smoothing factor γ in Eq.(13) is set to 0.8. For estimating the translation probability without topical information, we use GIZA++ 1.07 (Och and Ney, 2003) to do it.

Since hashtag recommendation task can also be modeled as a classification problem, we compare the proposed model with the following alternative methods:

- **Naive Bayes (NB)**: We formulate hashtag recommendation as a binary classification task and apply NB to model the posterior probability of each hashtag given a microblog.
- **Support Vector Machine (SVM)**: Similar to Naive Bayes, each hashtag can be regarded as one label and we use SVM to classify these microblogs.
- **Translation model (IBM-1)**: IBM model 1 is directly applied to obtain the alignment probability between the word and the hashtag (Liu et al., 2011).
- **Topical translation model (TTM)**: Ding et al. (2013) proposed the TTM for hashtag extraction. We implemented and extended their method for evaluating on the corpus constructed in this work. The number of topics in TTM is set to 20, and α is set to $50/K$. The hyperparameters used in TTM are also selected based on the development data set.

3.3 Experimental Results

Table 1 shows the comparisons of the proposed method with the state-of-the-art methods on the constructed evaluation dataset. "*CNHR*" denotes the method proposed in this paper. "*NHR1*" is a degenerate variation of CNHR, in which we consider all the hashtags are generated from distribution φ^1. "*NHR2*" is a model in which we consider all the hashtags are generated from

[1] http://www.weibo.com

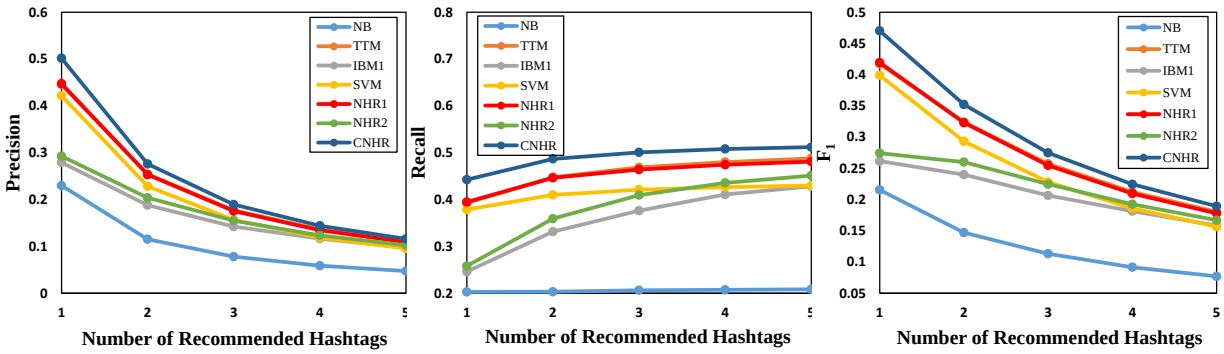


Figure 2: Precision, Recall and F_1 with recommended Hashtags range from 1 to 5

Table 1: Evaluation results of different methods on the evaluation collection.

Methods	Precision	Recall	F_1
NB	0.230	0.203	0.215
SVM	0.426	0.376	0.399
IBM1	0.279	0.246	0.261
TTM	0.445	0.393	0.417
NHR1	0.448	0.395	0.419
NHR2	0.293	0.258	0.274
CNHR	**0.502**	**0.442**	**0.470**

distribution φ^2. From the results, we can observe that discriminative methods achieve worse results than generative methods. We think that the large number of hashtags is one of the main reasons of the low performances.

From the results shown in Table 1, we also observe that the proposed method can achieve significantly better performance than existing methods. The relative improvement of proposed CNHR over TTM is around 12.7% in F_1. And we can see that the performances of TTM are similar as the results of NHR1. Since TTM and NHR1 are similar with each other except that TTM is based on LDA and NHR1 is adapted from DPMM. The results demonstrate the advantage of using DPMM over LDA. It does not need prior knowledge about number of topics. Comparing the results of the method CNHR with the methods NHR1 and NHR2 which do not take the types of hashtags into consideration, we can see that the proposed method benefits a lot from incorporating the types of hashtags.

Figure 2 shows the Precision, Recall, and F_1 curves of NB, IBM1, SVM, TTM, NHR1, NHR2 and CNHR on the test data. Each point of a curve represents the extraction of a different number of hashtags ranging from 1 to 5 respectively. In curves, the curve that is the highest of the graph indicates the best performance. Based on the results, we can observe that the performance of CNHR is the highest in all the curves. This indicates that the proposed method was significantly better than the other methods.

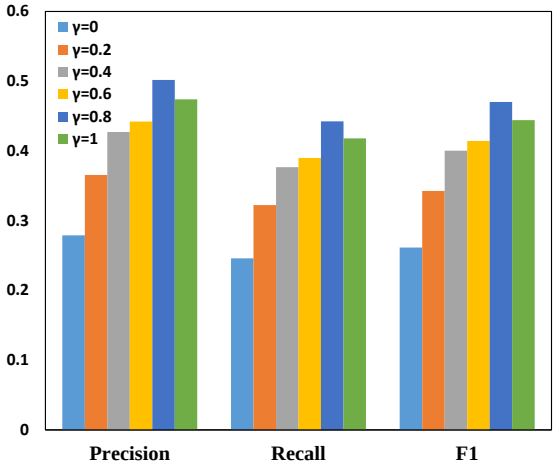
Table 2: The influence of the number of topics K of TTM.

K	Precision	Recall	F_1
10	0.441	0.389	0.413
20	**0.445**	**0.393**	**0.417**
30	0.432	0.381	0.405
40	0.413	0.364	0.387
50	0.391	0.345	0.367

In TTM, the number of topics K is also crucial factor. Table 2 shows the impact of the number of topics. From the table, we can observe that TTM obtains the best performance when K is set to 20. And performance decreases with more number of topics. We think that data sparsity may be one of the main reasons. With much more topic number, the data sparsity problem will be more serious when estimating topic-specific translation probability. We compare our method with the best performance of TTM.

From the description of the proposed model, we can know that there is a smooth parameter γ in the proposed method CNHR. To evaluate the impact of it, Figure 3 shows the influence of the translation probability smoothing parameter γ. When γ is set to 0.0, it means that the topical information is omitted. Comparing the results of $\gamma = 0.0$ and other values, we can observe that the topical information can benefit this task.

Figure 3: The influence of the smoothing parameter γ of CNHR.



When γ is set to 1.0, it represents the method without smoothing. The results indicate that it is necessary to address the sparsity problem through smoothing.

4 Related Works

Due to the usefulness of tag recommendation, many methods have been proposed from different perspectives (Heymann et al., 2008; Krestel et al., 2009; Rendle et al., 2009; Liu et al., 2012; Ding et al., 2013). Heymann et al. (Heymann et al., 2008) investigated the tag recommendation problem using the data collected from social bookmarking system. They introduced an entropy-based metric to capture the generality of a particular tag. In (Song et al., 2008), a Poisson Mixture Model based method is introduced to achieve the tag recommendation task. Krestel et al. (Krestel et al., 2009) introduced a Latent Dirichlet Allocation to elicit a shared topical structure from the collaborative tagging effort of multiple users for recommending tags. Ding et al. (2013) proposed to use translation process to model this task.

Based on the the observation that similar web pages tend to have the same tags, Lu et al. (2009) proposed a method taking both tag information and page content into account to achieve the task. They extended the translation based method and introduced a topic-specific translation model to process the various meanings of words in different topics. In (Tariq et al., 2013), discriminative-term-weights were used to establish topic-term relationships, of which users' perception were learned to suggest suitable hashtags for users. To handle the vocabulary problem in keyphrase extraction task, Liu et al. proposed a topical word trigger model, which treated the keyphrase extraction problem as a translation process with latent topics (Liu et al., 2012).

Most of the works mentioned above are based on textual information. Besides these methods, personalized methods for different recommendation tasks have also been paid lots of attentions (Liang et al., 2007; Shepitsen et al., 2008; Garg and Weber, 2008; Li et al., 2010; Liang et al., 2010; Rendle and Schmidt-Thieme, 2010; Huang et al., 2012). Shepitsen et al. (2008) proposed to use hierarchical agglomerative clustering to take into account personalized navigation context in cluster selection. In (Garg and Weber, 2008), the problem of personalized, interactive tag recommendation was also studied based on the statistics of the tags co-occurrence. Liang et al. (2010) proposed to the multiple relationships among users, items and tags to find the semantic meaning of each tag for each user individually and used this information for personalized item recommendation.

From the brief descriptions given above, we can observe that most of the previous works on hashtag suggestion did not take the types of hashtags into consideration. In this work, we propose to incorporate it into the generative methods.

5 Conclusions

In this paper, we study the problem of hashtag recommendation for microblogs. Since existing translation model based methods for this task regard all the hashtags generated from the same distribution, we propose a novel method which incorporates different type of hashtags have different distribution into the topical translation model for hashtag recommendation task. To evaluate the proposed method, we also construct a dataset from real world microblogging services. The results of experiments on the constructed dataset demonstrate that the proposed method outperforms state-of-the-art methods that do not consider these aspects.

6 Acknowledgement

The authors wish to thank the anonymous reviewers for their helpful comments. This work was par-

tially funded by National Natural Science Foundation of China (No. 61473092 and 61472088), the National High Technology Research and Development Program of China (No. 2015AA011802), and Shanghai Science and Technology Development Funds (13dz226020013511504300).

References

A.Bandyopadhyay, M. Mitra, and P. Majumder. 2011. Query expansion for microblog retrieval. In *Proceedings of The Twentieth Text REtrieval Conference*, TREC 2011.

Charles E Antoniak et al. 1974. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, 2(6):1152–1174.

S. Asur and B.A. Huberman. 2010. Predicting the future with social media. In *WI-IAT'10*, volume 1, pages 492–499.

Hila Becker, Mor Naaman, and Luis Gravano. 2010. Learning similarity metrics for event identification in social media. In *Proceedings of WSDM '10*.

Adam Bermingham and Alan F. Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of CIKM'10*.

David Blackwell and James B MacQueen. 1973. Ferguson distributions via pólya urn schemes. *The annals of statistics*, pages 353–355.

D.M. Blei and M.I. Jordan. 2003. Modeling annotated data. In *Proceedings of SIGIR*, pages 127–134.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of COLING '10*.

Zhuoye Ding, Xipeng Qiu, Qi Zhang, and Xuanjing Huang. 2013. Learning topical translation model for microblog hashtag suggestion. In *Proceedings of IJCAI 2013*.

Miles Efron. 2010. Hashtag retrieval in a microblogging environment. In *Proceedings of SIGIR '10*.

Thomas S Ferguson. 1983. Bayesian density estimation by mixtures of normal distributions. *Recent advances in statistics*, 24:287–302.

Nikhil Garg and Ingmar Weber. 2008. Personalized, interactive tag recommendation for flickr. In *Proceedings of RecSys '08*.

Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. 2013. Using topic models for twitter hashtag recommendation. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 593–596, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*.

Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, and Erel Uziel. 2010. Social media recommendation based on people and tags. In *Proceedings of SIGIR '10*.

Ido Guy, Uri Avraham, David Carmel, Sigalit Ur, Michal Jacovi, and Inbal Ronen. 2013. Mining expertise and interests from social media. In *Proceedings of WWW '13*.

Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. 2008. Social tag prediction. In *Proceedings of SIGIR '08*.

Wenyi Huang, Saurabh Kataria, Cornelia Caragea, Prasenjit Mitra, C Lee Giles, and Lior Rokach. 2012. Recommending citations: translating papers into references. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1910–1914. ACM.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of ACL 2011*, Portland, Oregon, USA.

Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. 2009. Latent dirichlet allocation for tag recommendation. In *Proceedings of RecSys '09*.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM.

Ting-Peng Liang, Hung-Jen Lai, and Yi-Cheng Ku. 2007. Personalized content recommendation and user satisfaction: Theoretical synthesis and empirical findings. *Journal of Management Information Systems*, 23(3):45–70.

Huizhi Liang, Yue Xu, Yuefeng Li, Richi Nayak, and Xiaohui Tao. 2010. Connecting users and items with weighted tags for personalized item recommendations. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 51–60. ACM.

Zhiyuan Liu, Xinxiong Chen, and Maosong Sun. 2011. A simple word trigger method for social tag suggestion. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1577–1588. Association for Computational Linguistics.

Zhiyuan Liu, Chen Liang, and Maosong Sun. 2012. Topical word trigger model for keyphrase extraction. In *Proceedings of COLING*.

Yu-Ta Lu, Shoou-I Yu, Tsung-Chieh Chang, and Jane Yung-jen Hsu. 2009. A content-based method to enhance tag recommendation. In *Proceedings of IJCAI'09*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Tsutomu Ohkura, Yoji Kiyota, and Hiroshi Nakagawa. 2006. Browsing system for weblog articles based on automated folksonomy. *Workshop on the Weblogging Ecosystem Aggregation Analysis and Dynamics at WWW*.

Takanobu Otsuka, Takuya Yoshimura, and Takayuki Ito. 2012. Evaluation of the reputation network using realistic distance between facebook data. In *Proceedings of WI-IAT '12*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Steffen Rendle and Lars Schmidt-Thieme. 2010. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 81–90. ACM.

Steffen Rendle, Leandro Balby Marinho, Alexandros Nanopoulos, and Lars Schmidt-Thieme. 2009. Learning optimal ranking with tensor factorization for tag recommendation. In *Proceedings of KDD '09*.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of WWW '10*.

Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, and Robin Burke. 2008. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, pages 259–266, New York, NY, USA. ACM.

Yang Song, Ziming Zhuang, Huajing Li, Qiankun Zhao, Jia Li, Wang-Chien Lee, and C. Lee Giles. 2008. Real-time automatic tag recommendation. In *Proceedings of SIGIR '08*.

Amara Tariq, Asim Karim, Fernando Gomez, and Hassan Foroosh. 2013. Exploiting topical perceptions over multi-lingual text for hashtag suggestion on twitter. In *FLAIRS Conference*.

Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of CIKM '11*.