TOMOHA: TOpic MOdel-based HAshtag Recommendation on Twitter

Jieying She Lei Chen
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong SAR, PR China
{jshe,leichen}@cse.ust.hk

ABSTRACT

On Twitter, hashtags are used to summarize topics of the tweet content and to help to categorize and search tweets. However, hashtags are created in a free style and thus heterogeneous, increasing difficulty of their usage. We propose TOMOHA, a supervised TOpic MOdel-based solution for HAshtag recommendation on Twitter. We treat hashtags as labels of topics, and develop a supervised topic model to discover relationship among words, hashtags and topics of tweets. We also novelly add user following relationship into the model. We infer the probability that a hashtag will be contained in a new tweet, and recommend k most probable ones. We propose parallel computing and pruning techniques to speed up model training and recommendation process. Experiments show that our method can properly and efficiently recommend hashtags.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*data mining*; I.5.4 [**Pattern Recognition**]: Applications—*text processing*

Keywords

Twitter; hashtag recommendation; topic model

1. INTRODUCTION

On Twitter, a string in a tweet can be marked by a # symbol, the so called hashtag, to represent a topic of the content, which helps tweet search and allows users to join the discussions. However, the arbitrariness of hashtags can lead to mess, prohibiting prevalence of hashtags. Automatic hashtag recommendation is one solution.

Previous works focused on two directions. The first is to recommend hashtags based on content similarity, facing difficulties in storage and efficient retrieval of a large volume of tweets. Fewer focused on the abstracted topics of tweets, mainly adopting Latent Dirichlet Allocation (LDA). LDA is unsupervised, and thus needs efforts to associate tweets with hashtags, which was not accomplished by [1]. Also, LDA, developed for long documents, could fail in short text. [4] proposed an unsupervised model, Twitter-LDA, specific for Twitter, but did not touch any recommendation task.

In this paper, we propose a supervised TOpic MOdel-based HAshtag recommendation (TOMOHA) solution. We follow the assumptions of [4] that each tweet is about one local topic and there is a global background topic for the corpus. We further treat hashtags as labels of local topics, and novelly add following relationship into the model. Based on the trained model, we infer the possibility that a hashtag will be contained in a new tweet and recommend the most probable ones. Scale-up techniques are further proposed for both model training and hashtag recommendation.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577292.

2. PROPOSED MODELS

1. TOMOHA: We propose a supervised topic model modified from Twitter-LDA, which assumes that a tweet is solely about one of the T local topics since each tweet is of limited length. Each word is assumed to be either a local topic word or a background word that is prevalent in many tweets. We adopt the same assumptions, but further treat hashtags as labels of topics, associating each local topic with a hashtag distribution, and thus propose a supervised model. The hashtags in a tweet depend on the local topic.

The generative process is as follows. When user u writes a new tweet d, u first samples a local topic $z_{u,d} \sim \text{Multi}(\boldsymbol{\theta}_u)$. Then for each word $w_{u,d,n}$, u decides whether it is a background word or a local topic word based on Bernoulli($\boldsymbol{\pi}$). If it is a background word, u samples a word from Multi($\boldsymbol{\phi}_B$), and otherwise from Multi($\boldsymbol{\phi}_{z_{u,d}}$). Finally, u samples hashtags from Multi($\boldsymbol{\psi}_{z_{u,d}}$). $\boldsymbol{\theta}$, $\boldsymbol{\pi}$, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are assumed to follow symmetric Dirichlet distributions.

TOMOHA-follow. On Twitter, users can be influenced by their followees. Thus, we further propose a model where users may follow topics of their followees. More specifically, for a new tweet, u first decides whose topic to follow based on Multi($\boldsymbol{\eta}_u$). Each $\eta_{u,r}$ reflects how likely u follows r's topics, where $r \in F_u$ can be either one of u's followees or u himself/herself. After deciding to follow $f_{u,d}$, u then samples a topic $z_{u,d} \sim \text{Multi}(\boldsymbol{\theta}_{f_{u,d}})$ and the remaining process is the same. We call this model TOMOHA-follow.

Parallel training. We use parallel computing to speed up training. We follow [3] to develop a distributed algorithm. We randomly assign tweets to P processors and train the models locally. The processors are synchronized every a few iterations, during which global counts are aggregated from local counts, and then all local counts are synchronized by the global counts.

2. Hashtag Recommendation: After training, we recommend existing hashtags $\{h\}$ to new tweets. Given a new tweet d, the probability $p(h|d)$ that d will contain h is calculated as Equation (1). For TOMOHA-follow, $\theta_{u,t}$ is replaced by $(\sum_{r \in F_u} \eta_{u,r}\theta_{r,t})$. We recommend k most probable ones.

Table 1: Statistics of Two Datasets

Dataset	#train tweets	#test tweets	#hashtags	#users
Dataset 1	33,720	14,931	637	183
Dataset 2	1,206,894	877&42,922	31,435	170,428

Table 2: Measurements of Topic Models on Dataset 1

Measure-ment	T VS L	T-f VS L	T VS T-f	T bg VS local	T-f bg VS local
JS_{avg}	0.34	0.34	0.23	0.5	0.53
τ_{avg}	0.16	0.15	0.37	0.14	0.24

$$p(h|d) = \sum_t \psi_{t,h} \times (\theta_{u,t} \prod_{w_{u,d,n}} (\pi_0 \phi_{B,w_{u,d,n}} + \pi_1 \phi_{t,w_{u,d,n}})) \quad (1)$$

Pruning. In Equation (1), $\psi_{t,h}$ is independent of new tweets. Thus, we can preprocess a sorted list of size T for each h, whose elements are sorted in non-increasing order by the value of $\psi_{t,h}$. On the other hand, $ts_t = \theta_{u,t} \prod_{w_{u,d,n}} (\pi_0 \phi_{B,w_{u,d,n}} + \pi_1 \phi_{t,w_{u,d,n}})$ is independent of the hashtag candidates. Thus, we can pre-calculate ts_t over T topics for a new tweet before looping the hashtag candidates. We also calculate the sum of ts_t. When calculating $p(h|d)$, instead of looping all T topics, we stop once we find that h cannot be recommended even if we sum up the remaining topics.

3. EVALUATION

We collected tweets through Twitter's REST API, and use two sets of data sampled and preprocessed from the collected data for different purposes of evaluation. Dataset 1 is for case study and Dataset 2 is for more realistic effectiveness and efficiency study. Details are presented in Table 1. We set T as 200.

1. Effectiveness: **Topic models.** We use Jensen-Shannon divergence D_{JS} to measure the similarity between word distributions of topics, and Kendall's τ to measure their agreement. We follow [2] to calculate the average JS_{avg} of D_{JS} and the average τ_{avg} of τ over all pairs of similar topics trained by two models. The results are presented in Table 2, where T, T-f and L represents TOMOHA, TOMOHA-follow and LDA, and "bg VS local" means we compare background topic with the most similar local topic. It indicates that there is a larger gap between the topics inferred by TOMOHA models and those by LDA, and that the background topic is quite different from local topics.

For the details of models, we observe that TOMOHA can better distinguish topics, e.g. LDA mixes discussions of Beijing and breast cancer into one topic that TOMOHA can well separate. Also, TOMOHA can discover topics that are not covered by LDA, e.g. only TOMOHA detects London Olympic Games. Finally, the top-ranked hashtags with high probabilities in the local topics reflect the corresponding topics, e.g. $\#ted$ weights 0.73 for the topic of Ted talks, $\#driving$ and $\#news$ both weight 0.5 for the topic of car accidents, and $\#olympics$ and $\#london2012$ weight 0.18 and 0.16 for London Olympic Games.

Hashtag recommendation. We recommend k most probable hashtags. We define Hit-1 and Hit-all rates, which are the percentages of test tweets that are recommended at least one or all original hashtags respectively. We implement three other algorithms for comparison: LDA, LDA-follow (L-F) and TFIDF (S). LDA is adapted to include hashtags, LDA-follow adopts the idea of TOMOHA-follow, and TFIDF recommends hashtags based on content similarity.

Table 3 presents the results on Dataset 1. We can observe that Hit-1 and Hit-all rates drop reasonably when we return a list of top-5 hashtags rather than a top-10 list. LDA and LDA-follow perform much worse than the other algorithms in both Hit-1 and Hit-all rates, while the other three achieve very close results. Results on Dataset 2 are similar and we do not present them for brevity.

Table 3: Hit Rates of Different Methods on Dataset 1

k	Measurement	T	T-f	L	L-f	S
5	Hit-1 rate	0.82	0.81	0.54	0.36	0.81
5	Hit-all rate	0.75	0.74	0.40	0.27	0.76
10	Hit-1 rate	0.89	0.88	0.75	0.54	0.87
10	Hit-all rate	0.84	0.83	0.63	0.45	0.83

Table 4: Parallel Training Performance with Different Settings

#Pro-cessors	#Iters be-tween Syncs	TOMOHA sec (Hit-1/Hit-all)	TOMOHA-follow sec (Hit-1/Hit-all)
1	100	8347 (0.888 / 0.839)	7866 (0.882 / 0.832)
2	5	5902 (0.889 / 0.836)	5975 (0.885 / 0.833)
2	10	5083 (0.884 / 0.832)	4656 (0.882 / 0.834)
4	5	4216 (0.885 / 0.837)	4482 (0.887 / 0.837)

TOMOHA-follow performs slightly worse. We find that many users follow their own topics most often. On Dataset 1, 138 users are most likely to follow their own topics, while on Dataset 2 there are 80,773 such users. It indicates that the role that following relationship plays in the model needs more future exploration.

Since the choice of hashtags is subjective, the original hashtags are not necessarily the ground truth. For example, our models recommend $\#innovation$ to a tweet containing $\#creativity$, and recommend $\#socialmedia$ to the tweet "Facebook search helps mother find her kidnapped children: #examiner". It indicates that our models fail to return the original hashtags in some cases but do recommend related ones.

TOMOHA and TOMOHA-follow are much more advantageous in efficiency. The average time taken by TOMOHA and TOMOHA-follow to recommend top-10 hashtags on Dataset 2 is 0.11s, while that by TFIDF is 6.5s. It indicates that our solution is much more advantageous in real-time application on Twitter.

2. Efficiency Improvements: **Parallel training.** We use Dataset 1 for this part of experiments since Dataset 2 is too large for non-parallel model training. In Table 4, we present the four settings we use and also the parallel training time for TOMOHA and TOMOHA-follow. The values in the parentheses are the Hit-1 and Hit-all rates ($k = 10$). We can observe that the training time decreases with increasing number of processors, and the Hit-1 and Hit-all rates of parallel-trained models do not drop. The results indicate that our parallel training algorithm is promising.

Pruning. On Dataset 1, the average recommendation time taken by TOMOHA and TOMOHA-follow is around 0.03s slower than the non-pruning ones. This is due to the additional cost of comparison in the pruning process. However, on Dataset 2, TOMOHA and TOMOHA-follow (around 0.1s) is much faster than the non-pruning ones (around 2.5s). The results indicate that our pruning technique is promising for the large volume of data on Twitter.

4. REFERENCES

[1] F. Godin, V. Slavkovikj, W. De Neve, B. Schrauwen, and R. Van de Walle. Using topic models for twitter hashtag recommendation. In *WWW'13*.

[2] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *SOMA'10*.

[3] D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. *JMLR'09*.

[4] 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'11*.