Learning to Annotate Tweets with Crowd Wisdom

Wei Feng
Tsinghua University
Beijing, China
feng-w10@mails.tsinghua.edu.cn

Jianyong Wang
Tsinghua University
Beijing, China
jianyong@tsinghua.edu.cn

ABSTRACT

In Twitter, users can annotate tweets with hashtags to indicate the ongoing topics. Hashtags provide users a convenient way to categorize tweets. However, two problems remain unsolved during an annotation: (1) Users have no way to know whether some related hashtags have already been created. (2) Users have their own way to categorize tweets. Thus personalization is needed. To address the above problems, we develop a statistical model for Personalized Hashtag Recommendation. With millions of <tweet, hashtag> pairs being generated everyday, we are able to learn the complex mappings from tweets to hashtags with the wisdom of the crowd. Our model considers rich auxiliary information like URLs, locations, social relation, temporal characteristics of hashtag adoption, etc. We show our model successfully outperforms existing methods on real datasets crawled from Twitter.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering, Retrieval Models, Selection Process*

Keywords

Social Media; Recommender Systems

1. INTRODUCTION

Hashtags are words prefixed with "#" and are used to indicate the topics of tweets. For example, "#Election2012" can be used in tweets related to United States presidential election of 2012. Despite the great importance of hashtags, a few problems remain unsolved when a user wants to annotate a tweet: (1) Before creating a new hashtag, is there any way for the user to find out whether some related hashtags have already been created and widely used? (2) Even if all the related hashtags are known, can we only suggest those hashtags that the user would likely to use according to their personal preferences for categorizing tweets. (3) Unannotated tweets cannot be used by hashtag-based applications. To the best of our knowledge, this is the first paper to study Personalized Hashtag Recommendation at tweet level. Yang[2] has studied user-level hashtag recommendation, i.e., what hashtags a user may adopt in the future regardless of which tweet is being considered. But their work cannot be directly used to facilitate tweet annotation.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

2. FRAMEWORK

Different from traditional recommender systems which only deal with <*user*, *item*> pairs, personalized hashtag recommendation handles <*user*, *tweet*, *hashtag*> triples with rich auxiliary information. Our model combines linear discriminative models with latent factor models. Let D, U and H denote tweet set, user set and hashtag set, respectively. Given a tweet $d \in D$ composed by user $u \in U$ and a hashtag candidate $h \in H$, the ranking score r_{udh} for hashtag h is

$$r_{udh} = \boldsymbol{\theta}^T \mathbf{x} + Rel(u, h) + Rel(d, h) \quad (1)$$

where $\boldsymbol{\theta}^T \mathbf{x}$ measures the contribution from explicit features (such as hashtag features). $\boldsymbol{\theta}$ is the weight vector to be learned and $\mathbf{x}$ is the feature vector. $Rel(d, h)$ and $Rel(u, h)$ measure content-relevance and user-relevance, respectively. Now we discuss each component in detail.

2.1 Measuring Content Relevance

The most intuitive idea is to recommend content-relevant hashtags. Suppose the target tweet $d \in D$ contains $k^{(w)}$ words $\{w_1, w_2, ..., w_{k^{(w)}}\}$, $k^{(l)}$ links $\{l_1, l_2, ..., l_{k^{(l)}}\}$, $k^{(m)}$ mentions $\{m_1, m_2, ..., m_{k^{(m)}}\}$, the content-relevance score between tweet d and hashtag h is computed by

$$Rel(h, d) = (\sum_{i=1}^{k^{(w)}} \alpha_i^{(w)} \mathbf{w}_i^T + \sum_{i=1}^{k^{(l)}} \alpha_i^{(l)} \mathbf{l}_i^T + \sum_{i=1}^{k^{(m)}} \alpha_i^{(m)} \mathbf{m}_i^T)\mathbf{h} \quad (2)$$

where (1) $\mathbf{w}_i$, $\mathbf{l}_i$, $\mathbf{m}_i$, $\mathbf{h}$ represent the latent factors for term w_i, link l_i, mention m_i, and the candidate hashtag h, respectively. (2) $\alpha_i^{(w)}$, $\alpha_i^{(l)}$, and $\alpha_i^{(m)}$ are weights of each latent vectors and they meet $\sum_{k=1}^{k=k^{(*)}} \alpha_i^{(*)} = 1$. (3) $\sum_{i=1}^{k^{(w)}} \alpha_i^{(w)} \mathbf{w}_i^T$, $\sum_{i=1}^{k^{(l)}} \alpha_i^{(l)} \mathbf{l}_i^T$, and $\sum_{i=1}^{k^{(m)}} \alpha_i^{(m)} \mathbf{m}_i^T$ represent weighted average of terms, links, and mentions, respectively.

Choices of $\boldsymbol{\alpha}^{(*)}$. $\alpha_i^{(w)}$ is defined to be TF-IDF(w_i) / ($\sum_{k=1}^{k^{(w)}}$ TF-IDF(w_k)). In this way, we can punish common words and promote informative words. $\alpha_i^{(l)}$ and $\alpha^{(m)}$ are defined to be the reciprocal of $k^{(l)}$ and $k^{(m)}$, respectively.

Term-Hashtag Affinity, which is the ratio of the number of times that hashtag h and term t co-occurred and the total number of times that hashtag h co-occurred with all terms.

2.2 Measuring User Relevance

Suppose user u has $k^{(f)}$ friends $\{u_1, u_2, ..., u_{k^{(f)}}\}$, and $k^{(p)}$ locations $\{p_1, p_2, ..., p_{k^{(p)}}\}$. The user-relevance score be-

Table 1: Dataset Statistics

Dataset	#User	#Social Relation	#Tweet	#Hashtag	#Links	#Mention	#Location
Month-Week	91,896	1,092,634	1,889,186	43,678	76,559	105,246	15,454
Week-Day	56,968	584,018	465,373	20,137	23,931	15,108	10,647

tween user u and hashtag h is

$$Rel(h,u) = [\beta \mathbf{u}^T + (1-\beta)\sum_{i=1}^{k^{(f)}} \alpha_i^{(f)} \mathbf{u}_{f_i}^T + \sum_{i=1}^{k^{(p)}} \alpha_i^{(p)} \mathbf{p}_i^T]\mathbf{h} \quad (3)$$

where (1) $\mathbf{u}$, $\mathbf{u}_{f_i}$, $\mathbf{p}_i$, $\mathbf{h}$ represent the latent factors for user u, her/his i-th friend u_{f_i}, location p_i, and the candidate hashtag h, respectively. (2) $\alpha_i^{(f)}$ and $\alpha_i^{(p)}$ are weights of the corresponding latent vectors and they meet $\sum_{k=1}^{k=k^{(*)}} \alpha_i^{(*)} = 1$. (3) $\beta\mathbf{u}^T + (1-\beta)\sum_{i=1}^{k^{(f)}} \mathbf{u}_{f_i}^T$ combines u's personal preference with her/his friends. $\beta \in [0,1]$ controls the biases.

Choices of $\boldsymbol{\alpha}^{(*)}$. $\alpha_i^{(f)}$ is defined as RT_COUNT(u, u_{f_i}) / $(\sum_{k=1}^{k^{(f)}}$RT_COUNT$(u, u_{f_k}))$, where RT_COUNT represents the times of u retweeting u_{f_i}. u is considered to trust u_{f_i} more if she/he retweets u_{f_i} more. Locations are set to be equally weighted since users have only one or two locations.

2.3 Incorporating Hashtag Features

Character Length. We find that hashtags of length 3 to 10 are more preferred.

Expected Frequency by Time Decay. Suppose a hashtag is used for N times in total. Let N_t denote the expected frequency at day t. According to the power-law distribution, the expected frequency at day t is $N_t = Nt^{-\lambda}$, where λ control the speed of decay and is fitted to 1.65 according to our data. Since we cannot know the real N, we replace N with the highest frequency N_0 of the target hashtag.

Time Span since Last Occurrence. This feature is used to filter out the out-dated hashtags and promote the currently used hashtags.

Uptrend. Uptrend measures whether a hashtag will grow or decend in the future. It is defined as $N(t_n)/N(t_{n-1})$, where t_n and t_{n-1} are two consecutive sampling time stamps. The interval is a day in this paper.

Frequency of Last Day of Occurrence. This feature represents whether the hashtag is popular according to the newest data.

2.4 Learning Parameters

We model our task as a binary classification problem. Suppose the ranking score is $\widehat{r}_{udp}$, the loss function is

$$loss = (\overline{r}_{udh}-1)\log(1-\widehat{r}_{udh}) - \overline{r}_{udh}\log(\widehat{r}_{udh}) + regularization \quad (4)$$

where (1) $\widehat{r}_{udh}$=sigmoid(r_{udh}), r_{udh} is the ranking score defined by Equation 1, and sigmoid$(x) = 1 \,/\, (1+e^{-x})$ maps r_{udh} to the range of (0, 1). (2) *regularization* term is defined as L2-norm regularization on all parameters. If $\widehat{r}_{udh}$ is close to the real label $\overline{r}_{udh}$, the loss is close to 0. We adopt stochastic gradient descent to minimize the loss function.

3. EXPERIMENTAL STUDY

Our datasets are crawled from Twitter using REST API. Two subsets are used: (1) **February vs The First Week of March, 2012** (denoted by 'M-W'), which is used to test whether we can use the data from the past month to predict for the next week. (2) **The Last Week of February vs The First Day of March, 2012** (denoted by 'W-D'), which is used to test whether we can predict for the next day using data from last week. The basic statistics of the datasets are shown in Table 1. We use Mean Average Precision (MAP) to measure the performance.

Table 2: Comparison of Different Models in MAP

Dataset	GraphRec	Content	User	Hybrid	Hybrid+
W-D	0.135	0.154	0.272	0.325	**0.355**
M-W	0.142	0.163	0.211	0.233	**0.264**

All the following models are evaluated: (1) **GraphRec**. Recently, Khabiri[1] proposed a general hashtag recommendation method based on the content of the tweet. To the best of our knowledge, this is the most relevant work. (2) **Content-based**. This model only considers content information discussed in Section 2.1. (3) **User-based**. This model only considers user information discussed in Section 2.2. (4) **Hybrid**. This model is a combination of Content-based and User-based without hashtag features. (4)**Hybrid+**. Based on Hybrid, this model further incorporates hashtag specific features discussed in Section 2.3. This is our final model.

The overall results are shown in Table 2. We have the following observations: (1) Content-based is slightly better than GraphRec mainly because Content-based makes use of two more indicators, i.e., web links and mentions. (2) user-based is surprisingly better than content-based. Notice that the test set only contains one day tweets and users only use 1.2 hashtag on average in this particular day. Since interest drifting is unlikely to happen in such a short time, recommending hashtags most preferred by the user or her/his neighbors is still a good strategy. (3) hybrid has a better performance than both Content-based and user-based. This indicates that recommended hashtags should be both user-relevant and content-relevant. (4) The performance of hybrid+ is further improved by considering temporal patterns of hashtag adoption. (5) The performance on 'W-D' is generally better than that on 'M-W'. This indicates predicting for the next day is easier than predicting for the next week since users interests may drift over time.

4. ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China under Grant No. 61272088.

5. REFERENCES

[1] E. Khabiri, J. Caverlee, and K. Y. Kamath. Predicting semantic annotations on the real-time web. In *HT*, pages 219–228, 2012.

[2] 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, 2012.