Knowledge-Based Systems 106 (2016) 196–205

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Suggest what to tag: Recommending more precise hashtags based on users' dynamic interests and streaming tweet content

Jia Li, Hua Xu*

State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 11 November 2015
Revised 21 May 2016
Accepted 23 May 2016
Available online 24 May 2016

Keywords:
Recommender systems
Social networks
Micro-blogging
Hashtag recommendation
Topic models
Online algorithms

ABSTRACT

Twitter is an online social networking microblogging service that allows registered users to broadcast 140-character messages called *tweets*. The service has gained worldwide popularity since it was created in March 2006, with more than 316 million monthly active users in June 2015 who posted 500 million tweets per day. As the number of available tweets grows, the problem of managing tweets becomes extremely difficult, which could lead to information overload. To avoid this problem, people use the hashtag symbol # before a relevant keyword or phrase in their tweets to categorize those tweets and help them show more easily in each Twitter search. Furthermore, hashtags can be used to collect public opinions on events and their ideas at the individual, community or even the world level. Incorporating hashtags to obtain better performance such as sentiment classification and breaking events detection also has attracted considerable research attention in recent years. However, there are very few tweets containing hashtags, which impedes the quality of search results and their further usage in various applications. Therefore, hashtag recommendation has become a particularly important research problem. In this paper, we first propose a novel model, namely online Twitter-User LDA to learn Twitter users' dynamic interests. Then considering the shortness, sparsity, and high volume of tweets, we introduce an effective method to discover the latent topics of streaming tweet content, which uses recently proposed incremental biterm topic model (IBTM). We finally design an automatic hashtag recommendation method called User-IBTM by combining the online Twitter-User LDA and IBTM. As shown in the experimental results on real world data from Twitter, our design method based on dynamic user interests and streaming tweet content significantly outperforms several other baseline methods and can suggest more precise hashtags.

1. Introduction

Twitter is an online social networking microblogging service that allows registered users to broadcast 140-character messages called *tweets*. More than posting personal lives and random thoughts, users can also follow and communicate with people based on common interests. Furthermore, Twitter has proved its worth as a breaking news platform in recent years. The shortness of tweets means that the first think people do at the scene of a developing public event or a news story is to tweet about it. In addition, lots of celebrities present on Twitter. You can follow your idols and even interact with them. Therefore, Twitter has experienced tremendous success in the past decade and has become one of the most important social network services. As of June 2015, on average 500 million tweets were posted per day, which corresponded to an average of 5700 tweets per second, by more than 300 million active Twitter users[1].

As the number of available tweets grows, the problem of managing tweets becomes extremely difficult, which could lead to information overload. To avoid this problem, an effective way called hashtag, the # symbol used before a relevant keyword or phrase, was created organically by Twitter users to categorize tweets. Not only can hashtags help users tag topics on the tweets, but they can also help users join a hashtag chat they are interested in. A hashtag chat is a way of organizing a conversation so that anyone on Twitter can follow, join, and contribute to it. For example, #edchat is the largest educational chat on Twitter. It has been a collaborative tool for educators to debate and evaluate solutions to various problems.

Hashtags have also attracted much attention in the research area. Laniado and Mika [19] introduce some metrics to describe the

* Corresponding author.
E-mail addresses: basicvbvc@gmail.com (J. Li), xuhua@tsinghua.edu.cn (H. Xu).

[1] https://about.twitter.com/company

http://dx.doi.org/10.1016/j.knosys.2016.05.047

characteristics of hashtags such as specificity, frequency, and stability over time. Davidov et al. [9] propose a supervised sentiment classification framework which combines hashtags and smileys as sentiment labels. This framework can avoid the need for labor intensive manual annotation and allow classification of diverse sentiment types of short texts. Cui et al. [8] utilize hashtags in Twitter as an indicator of events and study the properties of hashtags for breaking events detection. Meng et al. [24] integrate hashtags as weakly supervised information into topic modeling algorithms to obtain better interpretations and representations for calculating the similarity among them. Tsur and Rappoport [36] present an efficient approach based on a linear regression to predict the spread of a hashtag in a given time frame. Recently Sedhai and Sun [32] create HSpam14 dataset used for hashtag-oriented spam research in tweets.

Despite the important applications of hashtags, large portions of tweets do not contain hashtags. We find there are 32.4% of tweets in our dataset containing hashtags and Kywe et al. [18] report that only less than 20% of users add hashtags actively. Thus, a reliable hashtag recommendation system is needed to help users tag their new tweets with suitable hashtags.

To this end, many methods have been proposed to recommend hashtags for new tweets. An intuitive approach is to directly classify new tweets with annotated hashtags, for example [22] utilize a Naive Bayes model as the classifier. More methods attempt to solve this task based on traditional recommender systems [3,14,21,25,33]. [40] recommend hashtags based on a TF-IDF representation of the tweet. [18] suggest hashtags by combining hashtags of similar users and similar tweets. However, these methods may not fare well in practice since they all recommend existing hashtags to new tweets. The number of the collected hashtags has become a bottleneck. Indeed there are very few hashtags in existing tweets and most hashtags only have a very short life span according to [18]. Recently a handful of methods based on topic models have been proposed. For instance, [11] apply standard latent Dirichlet allocation (LDA) to automatically recommend keywords as hashtags. As we can see, although many studies approach hashtag recommendation in Twitter, the dynamic nature of hashtags is ignored.

In this paper we propose an automatic hashtag recommendation method. A single tweet typically has several attributes such as content, author, and post time. Therefore, considering both the streaming tweet content and dynamic author interests, our method suggests more precise hashtags.

Not as what most previous research did, we do not just recommend existing hashtags to tweets. Most hashtags have very short life spans according to [18]. The strong timeliness of hashtags hinders the performance of those methods that only suggest existing hashtags in practical use. We use topic models to discover the latent topics of each single tweet and then select suitable keywords for hashtag recommendation. However, due to the informal writing style and the 140-character length constraint, tweets are noisy, short and sparse. Standard latent Dirichlet allocation (LDA) proposed by Blei et al. [2] suffers from the data sparsity problem when documents are extremely short [35]. Therefore, we try applying biterm topic model (BTM) [38] in our method. We further consider the high volume of tweets. It's not practical to use static topic models like BTM to discover latent topics of streaming tweet content, for tweets are ordered and the topic-word distributions dynamically change over time. In addition, it is impractical for most computing conditions to sample the large whole corpus repeatedly. We finally apply a natural extension of BTM—incremental biterm topic model (IBTM) [7] using online algorithms to discover more accurate topics of streaming tweet content. To suggest more precise hashtags, we focus on creating a personalized hashtag recommender system by incorporating user interests. Considering the temporal sequence of tweets, we have the hypothesis that user interests dynamically change over time. However, previous research such as Weng et al. [37] and Chen et al. [6] in modeling user interests did not address this issue explicitly, as most of them simply built a "bag-of-words" document containing his/her posted tweets for each user. We extend this model by using a computationally inexpensive online algorithm, namely incremental Gibbs sampler [5], which immediately updates estimations of the topics as each tweet is observed. That is to say, we propose the online Twitter-User LDA. Finally, we propose our method—User-IBTM to recommend more precise hashtags.

The remainder of this paper is organized as follows. Related work is reviewed and our contributions are described in Section 2. The details of our proposed method is revealed in Section 3. The key components are two models: online Twitter-User LDA and IBTM. Section 4 presents our analysis of hashtags and the experimental results of our design method and other three baselines. We end the paper with conclusions and thoughts for future work.

2. Related work and our contributions

In this section, we review the existing methods for hashtag recommendation in Twitter and previous research on topic models for short texts.

Hashtag recommendation in Twitter: Basically, these methods can be mainly classified into two classes. One class of them focuses on exploiting the similarity between tweets. Mazzia and Juett [22] apply a naive Bayes model to classify tweets with hashtags. They allow their algorithm to produce a ranked list of the top-20 recommended hashtags. Zangerle et al. [40] compare three approaches to recommend suitable hashtags based on a TF-IDF representation of the tweet. Kywe et al. [18] suggest hashtags by combining hashtags of similar users and similar tweets. The TF-IDF method is used again in computing similar tweets. Otsuka et al. [27] propose the HF-IHU ranking scheme, which is a variation of TF-IDF, that considers hashtag relevancy. Sedhai and Sun [31] formulate recommending hashtags as a learning to rank problem and use RankSVM to rank the candidate hashtags. However, their methods only deal with tweets containing hyperlinks. In a word, the main disadvantage of them is that they all recommend existing hashtags to new tweets. There are very few hashtags in existing tweets and most hashtags only have a very short life span. It's time consuming to collect suitable hashtags and impractical to deal with high volume of tweets.

The other class focuses on using conventional topic models to discover topic distribution of each single tweet. Godin et al. [11] first apply standard LDA to automatically recommend keywords as hashtags. However, because of the expensive computation, static LDA can not apply to high volume streaming tweets. Static LDA also can not capture the dynamic change of topics. Ding et al. [10] novelly assume that hashtags and the content of the tweet are talking about the same themes but written in different languages. They treat hashtag suggestion as a translation process from content to hashtags and use a topic translation model (TTM) to solve this problem. However, their method, like LDA, cannot capture dynamic change of topics. Considering trending effects, Lu and Lee [20] combine Topics-over-Time (TOT) and Mixed Membership Model (MMM) to recommend hashtags. Like LDA, this model has to repeatedly train over the whole corpus if new tweets arrive. She and Chen [34] present a supervised topic model-based solution for hashtag recommendation. They treat hashtags as labels of topics, and discover relationship among words, hashtags and topics of tweets. Obviously, the scarcity of hashtags impedes the performance of this method.

As we can see, although a few very recent studies approach hashtag recommendation in Twitter, these methods either use a static topic model that is not designed for streaming tweets or