Expert Systems with Applications 41 (2014) 4330–4336

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Identifying interesting Twitter contents using topical analysis

Min-Chul Yang, Hae-Chang Rim *

Department of Computer & Radio Communications Engineering, Korea University, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Twitter
Interesting content
Topic model
LDA
Social media

ABSTRACT

Social media platforms such as Twitter are becoming increasingly mainstream which provides valuable user-generated information by publishing and sharing contents. Identifying interesting and useful contents from large text-streams is a crucial issue in social media because many users struggle with information overload. Retweeting as a forwarding function plays an important role in information propagation where the retweet counts simply reflect a tweet's popularity. However, the main reason for retweets may be limited to personal interests and satisfactions. In this paper, we use a topic identification as a proxy to understand a large number of tweets and to score the interestingness of an individual tweet based on its latent topics. Our assumption is that fascinating topics generate contents that may be of potential interest to a wide audience. We propose a novel topic model called Trend Sensitive-Latent Dirichlet Allocation (TS-LDA) that can efficiently extract latent topics from contents by modeling temporal trends on Twitter over time. The experimental results on real world data from Twitter demonstrate that our proposed method outperforms several other baseline methods.

1. Introduction

With the rise of the Internet, blogs, and mobile devices, social media has also evolved into an information provider by publishing and sharing user-generated contents. By analyzing textual data which represents the thoughts and communication between users, it is possible to understand the public needs and concerns about what constitutes valuable information from an academic, marketing, and policy-making perspective.

Twitter (http://twitter.com) is one of the social media platforms that enables its users to generate and consume useful information about issues and trends from text streams in real-time. Twitter and its 500 million registered users produce over 340 million tweets, which are text-based messages of up to 140 characters, per day[1]. Also, users subscribe to other users in order to view their followers' relationships and timelines which show tweets in reverse chronological order. Although tweets may contain valuable information, many do not and are not interesting to users. A large number of tweets can overwhelm users when they check their Twitter timeline. Thus, finding and recommending tweets that are of potential interest to users from a large volume of tweets that is accumulated in real-time is a crucial but challenging task.

A simple but effective way to solve these problems is to use the number of retweets. A retweet is a function that allows a user to re-post another user's tweet and other information such as profile credit. Retweeting, similar to forwarding in email, is a key part of information sharing and propagation in Twitter. However, not all retweets are meaningful insights in information sharing. For instance, a tweet such as "blessed and grateful. thank you" from Justin Bieber, the most-followed person on Twitter, is interesting only to his followers and fans; the tweet actually has gotten a lot of retweets, but it contains mundane content and is not informative or useful. According to Boyd, Golder, and Lotan (2010), some users retweet to spread tweets to new audiences, using it as a recommendation or productive communication tool, whereas others retweet a message not because of its content, but only because they are asked to or because they regard retweeting as an act of friendship, loyalty, or homage towards the person who originally tweeted.

Basically, the more social links a user has, the better the chances the user's postings will spread on social media, but the propagation may be limited to the user's social network. On the other hand, the content that attracts large audiences can be easily propagated even if its author is not popular. In other words, to find tweets that are interesting to a large number of users, it is important to consider the content's popularity rather than the author's popularity. Hence, we need to use semantic analysis for short social text messages that usually include noisy and conversational contents. To solve these problems, we adapted the Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), a statistical and popular unsupervised model. This text clustering model has been widely applied to text mining problems, and does not require manually constructed training data. In a training step in LDA, we can collect a set of related words that co-occurred in similar documents, which is

* Corresponding author. Tel.: +82 2 3290 3195; fax: +82 2 929 7914.
E-mail addresses: mcyang@nlp.korea.ac.kr (M.-C. Yang), rim@nlp.korea.ac.kr (H.-C. Rim).

[1] http://en.wikipedia.org/wiki/Twitter

0957-4174/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.12.051

referred to as "topics" from the continuous text streams. Theoretically, the LDA models the two types of probability distributions as latent variables: the probability of words under each topic and the probability of topics under each document. Since these probabilities can indicate each topic's characteristic and quality, we can score the topic based on its generality and specificity. Given a set of test data, we can also obtain the topic distribution of the new tweet from a trained topic model. To analyze a large number of documents, we exploit an alternative analysis such as the latent topic-based analysis, rather than examine the documents closely as previous methods–that used surface textual features–have done.

In this study, we focus on the automatic method to identify tweets that may be of potential interest to a wide audience. We conduct LDA-based topical analysis and infer latent topics to understand the content of an individual tweets and to score their interestingness. Our contributions are summarized as follows:

- We propose a novel and unsupervised approach that identifies interesting contents for a wide audience in Twitter and filters uninteresting contents to prevent users from dealing with an overwhelming number of tweets.
- We model textual contents in Twitter as latent topic structures using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and propose Trend Sensitive-LDA (TS-LDA) that can reflects the trends over time for more effective analysis.
- Based on topic identification, we score the weight of topics as its relative importance. Finally, we select individual tweets that contain more important and interesting topics than other tweets.
- We use the Amazon Mechanical Turk (AMT)[2] platform to collect tweets labeled by various groups of people.
- We conduct extensive experiments on a real dataset crawled from Twitter. The results prove that our model is more effective than the existing models.

The rest of the paper is organized as follows. In Section 2, we examine previous works related to this paper. In Section 3, we describe how to measure the interestingness of an individual tweet using topical analysis. In Section 4, we present the experimental results and analysis. Finally, Section 6 concludes the paper.

2. Related work

Many studies have provided insights into social media. Kwak, Lee, Park, and Moon (2010) firstly studied Twitter's structure by investigating various Twitter features. Recently, many works have focused mainly on analyzing or obtaining valuable information, such as influential users and posts on Twitter, from a large amount of social data. The most existing approaches (Castillo, Mendoza, & Poblete, 2011; Duan, Jiang, Qin, Zhou, & Shum, 2010; Hong, Dan, & Davison, 2011; Uysal & Croft, 2011) proposed to regard retweet counts as a measure of popularity, influence, and interestingness, and presented classifiers that predicted whether and how often new tweets will be retweeted in the future. They exploited various features of Twitter, such as textual data, author's metadata, and propagation information. Although the overall retweet count indicates a tweet's popularity, this may apply only to the followers of the tweet's author.

Twitter not only has textual data but also has linking data, such as follow and retweet links, which enable us to construct a network structure. The link-based approaches (Romero, Galuba, Asur, & Huberman, 2011; Yang, Lee, Lee, & Rim, 2012) applied a variant of the link analysis algorithm to a designed link structure in order to find interesting messages. However, the link structure requires a large volume of linking data to be analyzed and constructed and cannot be updated effectively when new documents stream in. Alonso, Carson, Gerster, Ji, and Nabar (2010) used crowdsourcing to categorize a set of tweets as interesting or uninteresting and reported that the presence of a URL link is a single, highly effective feature for selecting interesting tweets with more than 80% accuracy. This simple rule, however, may incorrectly categorize an uninteresting tweet (i.e., an uninteresting tweet contains links to meaningless pictures, videos, and advertisements) as interesting. Lauw, Ntoulas, and Kenthapadi (2010) suggested several features to identify interesting tweets but did not experimentally validate them. For user recommendation, Armentano, Godoy, and Amandi (2012) examined the topology of followers/followees network and identified the relevant users using social relation factors. Armentano, Godoy, and Amandi (2013) conducted not only topology-based profiling but also content-based profiling to find semantically similar users.

In social media, semantic analysis and topic modeling are widely used to understand textual data and can facilitate many applications such as user interest modeling (Pennacchiotti & Gurumurthy, 2011), sentiment analysis (Lin & He, 2009), content filtering (Duan & Zeng, 2013; Martinez-Romo & Araujo, 2013), and event tracking (Lee, 2012). Zhao et al. (2011) analyzed the topical differences between Twitter and traditional media using Twitter-LDA for investigating short messages. Wang and McCallum (2006) and Kawamae (2011) conducted topic modeling of temporally-sequenced documents in Twitter and tried to model the topics continuously over time. However, in our approach TS-LDA regards the mixtures of latent topics as a trend of its publishing time and is designed to learn changes in topic distributions, while other works focus on learning topic shifts based on word distributions. Chen, Nairn, Nelson, Bernstein, and Chi (2010) focused on recommending URLs posted in tweets using various combinations of topic relevance and social graph information. The Labeled-LDA (Ramage, Dumais, & Liebling, 2010) modeled a tweet using its labeled information, and then built the probability distribution vector of latent topics to represent the tweet's content. Based on similarity between the topic vectors, the researchers tried to find tweets that are similar to the ones which are already annotated "interesting." In terms of topic inference, our model is based on the model by Ramage et al. (2010) but is an unsupervised learning method with relative importance of latent topics.

3. Identifying interesting contents on Twitter

We treat the task of finding interesting tweets as a ranking problem where the goal is to obtain a scoring function that gives higher scores to interesting tweets in a given set of tweets. Our strategy is to focus on re-ranking the most-retweeted tweets according to their interestingness. We first introduce the two major concepts used in this paper.

Definition 1. Interesting in social media means that the content may be of potential interest to not only the authors and their followers but a wider audience. On the other hand, **uninteresting** means that the content is only interesting to the authors and their friends due to personal interests.

Definition 2. Interestingness indicates the size of the tweet's audience. Specifically, it is measured by the number of users that may be interested in the tweet. Also, an **interesting tweet** refers to a post whose interestingness is larger than a specific threshold.

Note that these definitions are followed by Alonso et al. (2010) and Lauw et al. (2010). Also, the survey on interestingness measures (Geng & Hamilton, 2006) reported that *Generality/Coverage*

[2] http://mturk.com