Neurocomputing 149 (2015) 1469–1480

Contents lists available at ScienceDirect

Neurocomputing

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

Event detection and popularity prediction in microblogging

Xiaoming Zhang [a,*], Xiaoming Chen [a], Yan Chen [a], Senzhang Wang [a], Zhoujun Li [a], Jiali Xia [b]

[a] State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China
[b] School of Software, Jiangxi University of Finance & Economics, Nanchang, China

ARTICLE INFO

Article history:
Received 1 September 2013
Received in revised form
6 August 2014
Accepted 22 August 2014
Available online 3 September 2014

Keywords:
Event detection
Popularity prediction
Burst words
Burst event

ABSTRACT

As one of the most influential social media platforms, microblogging is becoming increasingly popular in the last decades. Each day a large amount of events appear and spread in microblogging. The spreading of events and corresponding comments on them can greatly influence the public opinion. It is practical important to discover new emerging events in microblogging and predict their future popularity. Traditional event detection and information diffusion models cannot effectively handle our studied problem, because most existing methods focus only on event detection but ignore to predict their future trend. In this paper, we propose a new approach to detect burst novel events and predict their future popularity simultaneously. Specifically, we first detect events from online microblogging stream by utilizing multiple types of information, i.e., term frequency, and user's social relation. Meanwhile, the popularity of detected event is predicted through a proposed diffusion model which takes both the content and user information of the event into account. Extensive evaluations on two real-world datasets demonstrate the effectiveness of our approach on both event detection and their popularity prediction.

1. Introduction

Microblogging sites such as Twitter and SinaWeibo have become a popular way for users to share and disseminate information. Most microblogging services allow users to post short text message. For example, the length of content posted in SinaWeibo is limits to no more than 140 characters. As an important information sharing and consuming platform, microblogging sites usually hasa large volume of users, and generate huge number of contents every day. Take Twitter as an example, by July 2012 Twitter has over 500 million users, and the users collectively create over two billion tweets each week. With the huge volume of users and contents, microblogging provides us a desirable platform to study the spreading information from many perspectives. First, microblogging captures everything from mundanely daily routine of the masses to spectacular breaking events to other messages of significant historic impact. As a result, it is very difficult for users to capture new events which will be popular in the future. Thus, there is an urgent need to automatic detect event online. However, it is very challenging to effectively detect event in microblogging. The primary obstacle is that the language used in tweets is usually substantially different from that used in traditional text due to the length constraint of the tweets. Second, many researchers have shown that microblogging provides a probe into the sentiments and behaviors of entire popularity, enabling what has come to be known as computational social science. For example, content generated in social network sites can be used to predict the adoption of a new product or the result of an election [16]. As for the news events, it would be of interest to predict how popular they will be. Analyzing the spread of event would also be of interest in many domains. For example, online advertisers could use this analysis for efficiently targeted marketing campaigns. Public organizations can know how the society is influenced by the event and then determine how to reply to the public opinion.

In this paper, we study how to detect burst events from the stream of microblogging contents, such as “Hurricane Sandy hitting New York”, “Japan Earthquake in 2011” and “Beijing rainstorm in 2012”. Then, we further predict its popularity in the near future by modeling the spreading of the event. Burst event detection and information diffusion in social network are extensively studied For example, some works utilized a feature-pivot strategy to detect event in social media [11,19,43].These methods first detect burst words based on their frequency distributions on the time axis. Then, burst words are clustered, and each cluster represents a burst event. However, the definition of burst event in these works is only based on the novelty of event. The popularity of burst event in the near future is usually ignored. Moreover, as the length of micro-blog document is limited, the representation of micro-blog using terms vector faces the feature

* Corresponding author. Tel.: +86 10 82338247.
E-mail address: yolixs@buaa.edu.cn (X. Zhang).

http://dx.doi.org/10.1016/j.neucom.2014.08.045
0925-2312/

sparsity problem. This problem can be alleviated by exploiting other types of web resource. Usually, an event appears and spreads in multiple sources simultaneously, such as microblogging sites, blog sites, web forum and traditional news sites. Thus, these web resources can be used to enrich the information of micro-blog documents.

The information spread model has also been studied for many years. These works can be categorized into two groups. The first group focuses on the topology of social graph, investigating what topologies and what activation patterns facilitate efficient propagation of information. For example, the macro-level dynamics and characteristics of information diffusion are discussed [13,22], key factor that affect the adoption of behaviors is revealed [4,23] and contagion models are designed to simulate the diffusion process [36]. Some researches focus on the micro-level analysis of information spread in large scale social network, such as who got the gossip from whom or who infected whom [26]. The topology-based approaches usually make the following assumptions: (a) all the users have the same role in the spread process; (b) the spread path between users are known and the influence can only be transmitted over the edges of the underlying network; and (c) complete network data is available. However, in many scenarios, the underlying network is implicit or even unknown [40]. Furthermore, the topology-based approaches are not capable to analyze the content of the propagated information. They do not consider the difference among different users' reaction on the propagated information, either. For example, users who are interested in sports are more likely to be active in the spread of the event "Liu Xiang falls in Olympic game" than users who have little interest in sports. The other group of works mainly focuses on the analysis of the coarse-grained features of information content and network nodes [39], such as the average out-degree of node and the length of hashtag. They did not distinguish different activities and interest of users in event spread.

In this paper, we propose an approach to detect event from micro-blog stream. To tackle the challenge of data sparsity, we propose to enrich the microblogging information related to the event by exploiting multi-source content, such as blog sites, web forum and traditional news sites. We extend the burst event to incorporate temporal aspect of timeliness. In other words, we not only detect burst novel events, but also predict how popular they will be in the near future. This presents an additional challenge to model the temporal characteristics of event in real-time microblogging stream. We need to predict how popular an event will be, as soon as this event is detected. Detecting events and accurately predicting their future trends can contribute to better providing users and organizations potentially valuable information. For example, organizations may be interested in tracking the events related to them, and users would like to be informed of new burst events which are fast gathering momentum in microblogging. Considering the unique characteristics of microblogging, we use the feature-pivot method to detect micro-blogging event. Particularly, we first propose to combine the social relations of users and the frequency distribution of words to detect burst words. Then, the burst words are clustered into groups, and each group of words can be considered to be related to a specific event. The two-state model proposed by Kleinberg [19] is revised to detect burst words whose weights are real values. To cluster burst words, we construct a words graph based on their co-occurrence in micro-blogs and other web resources. Then, strongly connected burst words are clustered, and each of the clusters represents a burst event.

To predict the popularity of a detected event, we model the spread of an event by combing the posts related to the events and the social relations of related users. Different from most information diffusion models, our approach can handle graphs with incomplete structure information. The proposed approach utilizes a linear spread prediction function to predict the future popularity of the events. The linear function combines all kinds of information available, such as the influence power and interest of users, and the historical popularity information of the event. The motivation of this function is based on the following observations. First, a more active user usually contributes more to the spread of an event. To measure the activity of users in the social network, we introduce the influence power of users. Second, different users may have different interest on different topics. For example, some users prefer entertainment related topics, while some prefer politics related topics. Users are more likely to participate in events which are about the topics they are interested in and share them with their friends. To model the interest of users, a topic model is proposed. By extracting the profile of users and all the tweets of the events, the model automatically discovers the latent topics of users and the event and represents them as two topic vectors. The similarity between the two vectors is used to denote user's interest in the event. Third, if an event is very popular in the past, it is more likely to be popular in the future. To combine all above factors, the popularity of an event is assumed to be a linear function of the volumes produced by different users infected in the past and the volume introduced by its historical popularity. The main contributions of the paper can be summarized as follows:

1. We not only detect burst events that are novel, but also to predict how hot the events will be in the near future. To detect burst event by combining term's occurrence information and users' social relation information, the two-state model is improved to deal with real value.
2. We proposed a spread model based on the analysis of both event content and users' profile. The major advantage of our model is that it distinguishes users' contributions according to user's influence power and interest in the predicted event, which is different from other approaches that use the same parametric form for all events and users.
3. We further evaluate our approach in two real-life datasets, and experiment results indicate the efficiency of our approach.

The remainder of this paper is organized as follows. In the next section, we introduce related works. We formally formulate the problem in Section 3, and we propose our event detection algorithm in Section 4. Section 5 describes how to predict the popularity of an event, and the experiments are described in Section 6. Finally, the paper is concluded in Section 7.

2. Related works

The enormous amount of contents generated by social network users in the last decade is creating new challenges and new research interests for data mining, social network analysis and other related community. In this section, we present an overview of those works which are related with our work, i.e., event detection and information spread model.

The first issue when dealing with text stream is the aggregation of them. Many works try to aggregate text documents using the event detection approach [3,31,44]. Existing event detection approaches can be broadly classified into two categories: document-pivot approaches and feature-pivot approaches. The former ones detect events by clustering documents based on the similarity between documents [25,30,41], while the latter ones detect which words refer to event [19,11]. Since messages in microblogging sites are constrained to be short text (up to 140 characters), the sparse vector representation affect the similarity measure between two micro-blogs and hence affect the performance of document-pivot methods. Therefore, most of the works detect event in social media using the feature-pivot approach. We mainly introduce the works based on feature-pivot approach in the following part of this section.

Among the feature-pivot approaches, many works mainly depend on the analysis of term frequency. Kleinberg proposes to detect burst