Physica A 447 (2016) 247–260

Contents lists available at ScienceDirect

Physica A

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

Mining the key predictors for event outbreaks in social networks

Chengqi Yi [a], Yuanyuan Bao [b,c], Yibo Xue [b,c,*]

[a] School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China
[b] Research Institute of Information Technology, Tsinghua University, Beijing 100084, China
[c] Tsinghua National Lab for Information Science and Technology, Tsinghua University, Beijing 100084, China

HIGHLIGHTS

- We propose a new definition for event outbreaks, based on the structure of social networks.
- We analyze outbreak events by segmenting them into 20 temporally equal parts.
- We uncover those features relevant to accurate predictions at each stage of an outbreak event.
- Five specific features are especially relevant to accurate predictions of event outbreaks.

ARTICLE INFO

Article history:
Received 4 July 2015
Received in revised form 28 October 2015
Available online 21 December 2015

Keywords:
Social network
Outbreak prediction
Information dissemination
Predictors
Data-driven

ABSTRACT

It will be beneficial to devise a method to predict a so-called event outbreak. Existing works mainly focus on exploring effective methods for improving the accuracy of predictions, while ignoring the underlying causes: What makes event go viral? What factors that significantly influence the prediction of an event outbreak in social networks? In this paper, we proposed a novel definition for an event outbreak, taking into account the structural changes to a network during the propagation of content. In addition, we investigated features that were sensitive to predicting an event outbreak. In order to investigate the universality of these features at different stages of an event, we split the entire lifecycle of an event into 20 equal segments according to the proportion of the propagation time. We extracted 44 features, including features related to content, users, structure, and time, from each segment of the event. Based on these features, we proposed a prediction method using supervised classification algorithms to predict event outbreaks. Experimental results indicate that, as time goes by, our method is highly accurate, with a precision rate ranging from 79% to 97% and a recall rate ranging from 74% to 97%. In addition, after applying a feature-selection algorithm, the top five selected features can considerably improve the accuracy of the prediction. Data-driven experimental results show that the entropy of the eigenvector centrality, the entropy of the PageRank, the standard deviation of the betweenness centrality, the proportion of re-shares without content, and the average path length are the key predictors for an event outbreak. Our findings are especially useful for further exploring the intrinsic characteristics of outbreak prediction.

* Correspondence to: Research Institute of Information Technology, Tsinghua University, FIT Building 3-418, China. Tel.: +86 10 62772393.
E-mail address: yiboxue@tsinghua.edu.cn (Y. Xue).

http://dx.doi.org/10.1016/j.physa.2015.12.019
0378-4371/

1. Introduction

In recent years, social networks such as Facebook, Twitter, and Sina Weibo [1] is becoming an indispensable part of our lives. With the rapid development of social-networking websites, posting and re-sharing content through social networks is becoming important mechanisms for communication. Facebook and Twitter reported more than 1.415 billion and 288 million monthly active users, respectively, as of March 2015 [2]. Akin to a hybrid of Twitter[1] and Facebook[2], Sina Weibo[3] is a Twitter-like social forum in China, and it is becoming one of the most popular Chinese microblogging websites, with more than 176 million monthly active users as of April 2015. These numbers are expected to grow significantly over the next few years [3]. By 2018, it is estimated that there will be approximately 2.44 billion social networks users around the globe, up from 1.79 billion in 2014 [4]. Additionally, social networks allow hundreds of millions of Internet users to publish, discuss, and share various kinds of information, such as news, product recommendations, political opinions, ideas, interests. Social networks provide access to a vast source of information on an unprecedented scale.

As content is re-shared by users, massive events derived from the content can spread with the potential to reach many people—resulting in a so-called event outbreak. It will be beneficial to devise a method that can predict event outbreaks. Such a method has obvious applications: targeted marketing, viral advertising, emergency management, and even anti-terrorist campaigns. Existing works mainly focus on exploring effective methods for improving the accuracy of outbreak prediction. However, these works tend to ignore the underlying causes: *features that may significantly influence the prediction of an event outbreak in a social network*. In this paper, we focus on these features in an attempt to mine the key predictors for event outbreaks in social networks.

In short, we conduct our investigation in the following manner:

(1) We transformed the problem of predicting event outbreaks into a binary classification problem, and we proposed a novel definition for event outbreaks that takes into account structural changes to the network during the propagation of content.
(2) In order to investigate the universality of features at different stages of an event, we split the entire lifecycle of the event into 20 equal segments according to the proportion of the propagation time. Furthermore, we extracted 44 features related to content, users, structure, and time from each segment of the event.
(3) We attempt to harness real-world information-dissemination datasets based on Sina Weibo.
(4) In order to select the key predictors for event outbreaks, we applied a feature-selection algorithm to determine the top five features that can considerably improve the accuracy of the prediction.

We found that, as time goes by, our method is increasingly accurate, with a precision rate ranging from 79% to 97% and a recall rate ranging from 74% to 97%, after using several popular supervised classification algorithms. Data-driven experimental results show that the entropy of the eigenvector centrality, the entropy of the PageRank, the standard deviation of the betweenness centrality, the proportion of no-content re-shares, and the average path length are the key predictors for event outbreaks.

The rest of this paper is organized as follows. In Section 2, we present related work. In Section 3, we introduce a definition for event outbreaks and the feature-extraction approach. In Section 4, we describe the data-collection approach and the dataset. In Section 5, we discuss our data-driven experiments and show the experimental results. In Section 6, we conclude the paper and discuss future research directions.

2. Related work

In recent years, information diffusion, also known as information cascade, has drawn considerable attention from many fields of research, and a variety of methods and models have been proposed to capture information diffusion in social networks [5–8]. Some researchers focus on building effective models to explain the general process of information diffusion. These models are useful for simulating the flow of information in social networks [9–11]. However, because the process is complex – and because of the uncertainty – these models cannot be directly applied to detect or predict an outbreak. Therefore, pioneering research has focused on studying outbreak detection in several ways. The important aspects to outbreak detection have been discussed, and many valuable results have been obtained. Leskovec et al. presented a general methodology for near-optimal sensor placement in Ref. [12]. Kumar et al. analyzed the bursty evolution of a blog network, and discovered its likeness to the structure of a social community. They analyzed the formation of micro-communities over time and applied this analysis to online social networks [13]. Yao et al. proposed an approach to detect bursty tagging events. Their approach captures the relations within groups of correlated tags that are either bursty or associated with a bursty tag co-occurrence [14]. Mathioudakis et al. presented TwitterMonitor, a system for detecting trends over the Twitter stream. Their system identifies emerging topics on Twitter in real-time and provides meaningful analysis by generating accurate descriptions of each topic [15]. Prakash et al. proposed an efficient method, called Netsleuth, for the well-known susceptible–infected virus-propagation model and investigated how to identify the nodes from which the cascade started

[1] https://twitter.com/.
[2] https://www.facebook.com/.
[3] http://weibo.com/.