Neurocomputing 144 (2014) 463–470

Contents lists available at ScienceDirect

Neurocomputing

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

Predicting the topic influence trends in social media with multiple models☆

Yi Han [a,b,*], Binxing Fang [b,c], Yan Jia [b]

[a] Peking University, China
[b] National University of Defense Technology, China
[c] Beijing University of Posts and Telecommunications, China

ARTICLE INFO

Article history:
Received 18 October 2013
Received in revised form
18 February 2014
Accepted 11 March 2014
Communicated by M. Wang
Available online 19 June 2014

Keywords:
Social network
Prediction
Time series
Multiple models

ABSTRACT

Online social networks, such as twitter and facebook, are continuously generating the new contents and relationships. To fully understand the spread of topics, there are some essential but remaining open questions. Why are some seemingly ordinary topics attracting? Is it due to the attractiveness of the content itself, or some external factors, such as network structure, time or event location, play a larger role in the dissemination of information? Analyzing the influence and spread of upcoming contents is an interesting and useful research direction, and has brilliant perspective on web advertising and spam detection. In this paper, a novel time series model for predicting the topics social influence has been introduced. In this model, the existing user-generated contents are summarized with a set of valued sequences, and a hybrid model consisting of topical, social and geographic attributes has been adopted for predicting influence trends of newly coming contents. The empirical study conducted on large real data sets indicates that our model is interesting and meaningful, and our methods are effective and efficient in practice.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

With the rapid growth of the social media, the information spreads around the world with surprising speed and intensity. Users of some online social media, such as twitter and facebook, are continuously generating new contents, forming new interactions, and updating their status over time. With the evolving of the network structure, some contents generate and spread very fast. In twitter, some interesting posts (tweets) are usually retweeted thousands of times, and their social influence is also boosted with the reposting activities.

There are some interesting questions people may ask. Why are there some topics obviously more influential than other? What kind of topics could be attractive? Is there a method to predict or estimate the influence of a specified topic? Is the topological substructure of a social network related with the evolution of topics? Answering these questions is essential for understanding the mechanism of evolving social networks.

When a specified topic becomes a hot spot, we consider there are two possible reasons. The topic itself is interesting, or some external factors amplify its influence.

Social media has been studied extensively from variable angles such as degree distribution analysis [1,2], community extraction [3] and pattern discovery [4]. In this paper, we analyze the relationship between hot topics in social media, and formulate the information flows on different topics as a set of sequences. In order to predict the influence trend of a specific topic, an early prediction method with multiple factors is adopted. Moreover, we introduced a novel supervised learning method which considers topical, social and geographic properties of information flows in social networks.

To the best of our knowledge, there is no previous study on social media taking into account the early prediction and geographic properties. We made the following contributions.

First, we introduce a novel time series model to represent the continuously generated content in social media. The fact that user-created contents on a specified topic spread in a specific social network can be modeled by a sequence of vertices which participate the interaction.

Second, we propose a similarity measure among network time series. In a given time period, the continuously generated content about a given topic can be represented as a time series. For different topics, the distance among time series, which can be

☆The research was supported by National Basic Research Program of China (973 Program, No. 2013CB329600), National Natural Science Foundation of China (Nos. 91124002 and 61372191), and China Postdoctoral Science Foundation Program (Nos. 2012M520114 and 2013T60037).

* Corresponding author.
E-mail address: yihan@pku.edu.cn (Y. Han).

http://dx.doi.org/10.1016/j.neucom.2014.03.054
0925-2312/© 2014 Elsevier B.V. All rights reserved.

regarded as the similarity, is measured by content similarity with social and geographic attributes.

Third, we propose a novel prediction model on social media. The influence trends of the topics, as the target value, can be predicted effectively. The influence of upcoming content will be estimated by analyzing the individuals and related existing content.

Last, we conduct systematic experiments on two real data sets. The experimental results indicate that our model is useful and interesting, and our methods are effective.

Our solution can be applied as two different types of applications, topic tracing (Fig. 1(a)) and potential topic discovery (Fig. 1(b)).

In Fig. 1, both are based on the supervised classification model, in which the data before the current time stamp can be used as training data to generate the classifier, and the newly coming content can be used as testing data to tune the classifier. For a given topic, its spreading trace, which can be represented as a series of vertices, keywords and time stamps, can be extracted easily. The classifier which has been retrieved from the network can be applied for predicting the future state of the time series. Fig. 1(a) shows the example. Another important application is the influential topic prediction. The system automatically estimates the future impact of candidate time series, and outputs the topics which are potentially influential (shown in Fig. 1(b)).

The rest of the paper is organized as follows. We review the related work in Section 2, and formulate the problem in Section 3. We discuss the similarity measure and prediction methods in Section 4. A systematic empirical study conducted on real data sets is reported in Section 5. Section 6 concludes the paper.

2. Related work

Our work is highly related to the previous studies on spreading dynamic models, classification and prediction on social media, and social influence estimation. In this section, we review some representative work briefly.

2.1. Spreading dynamic models

Epidemic propagation model is a successful mathematical model for which has a long history, in which, the statues of individuals can be summarized into 3 categories. S (susceptible) indicates the individual is in a healthy state and has a probability of being infected by someone. I (infected) indicates the individual has been inflected and has a probability of infecting others or being recovered. In a network, if two vertices are connected then they are considered to have contact, Thus, if one vertex is infected by a virus and the other is susceptible, then with a certain probability the latter may become infected as time goes on. R (recovered) indicates the individual cannot be infected anymore. Different combinations of above states lead to different models, such as SIR [5] and SIS [6] models. Epidemic propagation model is a straightforward model for describing the information diffusion process. However, the events on social media are affected by lots of external factors, like emergencies, breaking news, social spammers, which cannot be characterized by epidemic propagation model well.

2.2. Classification and prediction on social media

Supervised learning [7] on social networks is a central subject in graph data processing. Some previous studies utilized a certain number of subgraphs as training set. In training set, the target values, which can usually be vertex properties, are available. The goal is to derive the target values of the remaining part of the graph. In some large-scale social networks, a central task is to classify unlabeled nodes given a limited number of labeled nodes. For example, the social service provider manually labels a small number of people who responded to a certain advertisement as positive nodes, and people who did not respond as negative nodes. Based on these labeled nodes, other people's response can be predicted. Graph classification tasks can also be unsupervised. Unsupervised methods classify graphs into a certain number of categories by similarity [8,9]. An interesting direction on evolving social network is the link prediction. The appearance of new links indicates new interactions between vertices. Given a snapshot of a social network at time t and a small number Δt, the objective is to predict the edges that will be added to the network in the time interval $(t,t+\Delta t)$. Murata considered that proximities between nodes can be estimated by using both graph proximity measures and the weights of existing links, and a prediction method based on weighted proximity measure has been introduced in [10]. Leskovec introduced a method of predicting the edge sign. The main goal is to utilize the graph structure and vertex label to infer the hidden sign labeled on edges [11]. A logistic regression classifier has been used to combine the evidence from individual features. Backstrom in facebook developed an algorithm that combines the information from the network structure with node and edge level attributes to guide a random walk on the graph [12]. The scoring function was learned in order to assign weights to edges, and walkers are regarded more likely to visit the nodes to which new links will be created.

2.3. Social influence estimation

Several well-known link-based ranking algorithms, including PageRank [13] and HITS [14], have been designed for ranking entities in social networks in past decades. PageRank [13] measures the importance of a vertex v by considering how collectively others pointing to directly or indirectly. For each vertex, the amount of ranking contribution from a neighbor is decided by

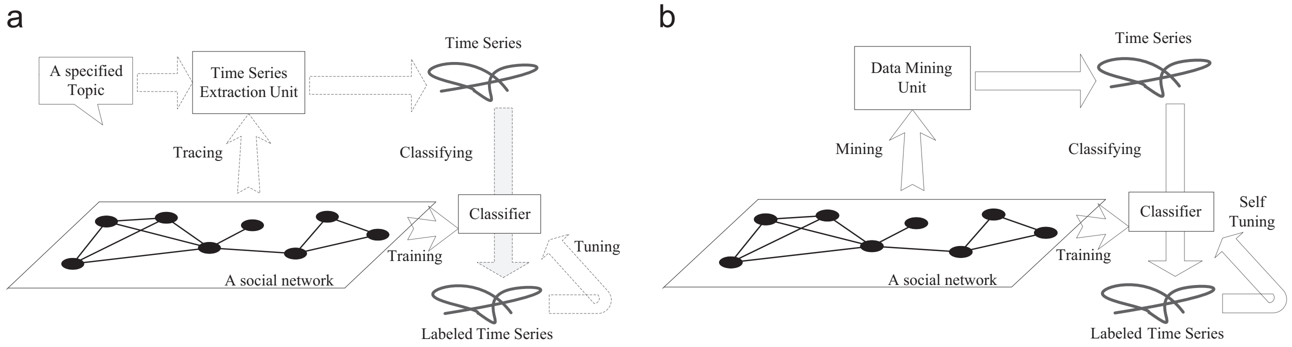

Fig. 1. Two different types of applications. (a) Topic tracing. (b) Potential topic discovery.

Download English Version:

https://daneshyari.com/en/article/412218

Download Persian Version:

https://daneshyari.com/article/412218

Daneshyari.com