Information Sciences 422 (2018) 432–446

Contents lists available at ScienceDirect

Information Sciences

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

Forecasting participants of information diffusion on social networks with its applications

Cheng-Te Li [a], Yu-Jen Lin [b], Mi-Yen Yeh [b,*]

[a] Department of Statistics, National Cheng Kung University, Tainan, Taiwan
[b] Institute of Information Science, Academia Sinica, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 11 September 2016
Revised 30 May 2017
Accepted 11 September 2017
Available online 12 September 2017

Keywords:
Diffusion forecasting
Information diffusion
Social network
Diffusion participation
Social influence
Popularity prediction

ABSTRACT

Social networking services allow users to adopt and spread information via diffusion actions, e.g., share, retweet, and reply. Real applications such as viral marketing and trending topic detection rely on information diffusion. Given past items with diffusion records on a social network, this paper aims at forecasting who will participate in the diffusion of a new item c (we use hashtags in the paper) with its k earliest adopters, without using content and profile information, i.e., finding which users will adopt c in the future. We define the Diffusion Participation Forecasting (DPF) problem, which is challenging since all users except for early adopters can be the candidates, comparing to existing studies that predict which one-layer followers will adopt a new hashtag given past diffusion observations with content and profile info. To solve the DFP problem, we propose an Adoption-based Participation Ranking (APR) model, which aims to rank the actual participants in reality at higher positions. The first is to estimate the adoption probability of a new hashtag for each user while the second is a random walk-based model that incorporates nodes with higher adoption probability values and early adopters to generate the forecasted participants. Experiments conducted on Twitter exhibit that our model can significantly outperform several competing methods in terms of Precision and Recall. Moreover, we demonstrate that an accurate DPF can be applied for effective targeted marketing using influence maximization and boosting the accuracy of popularity prediction in social media.

1. Introduction

Information diffusion is the major production of social interactions among users in online social networking services such as Twitter and Facebook. With various diffusion actions, such as retweet, share, reply and rating, different items (e.g., hashtags, short texts, images, and URLs) can be propagated from one user to another, which we call *diffusion participants*, in a social network. Understanding information diffusion from online social data can enable many real-world applications, such as viral marketing [12], trending topic detection [32], and information summarization [26]. Predicting the future participants who will spread a certain item is one of the most essential steps to uncover the mechanism of information diffusion. Some of existing studies predict whether or not a user will adopt the same item that his/her followees had adopted earlier [8,21,36] given the item content and/or user profiles. However, we think it is insufficient to find the diffusion participants

* Corresponding author.
E-mail addresses: chengte@mail.ncku.edu.tw, reliefli@gmail.com (C.-T. Li), maxwellbiga@gmail.com (Y.-J. Lin), miyen@iis.sinica.edu.tw (M.-Y. Yeh).

http://dx.doi.org/10.1016/j.ins.2017.09.034

from only the followers of the *early adopters* who are the first users participating in the diffusion. In addition, in real scenarios, it is unrealistic to assume that the item content and user profiles are always accessible due to the privacy issue and the storage constraint. Therefore, we propose to forecast the diffusion participants of an item among followers and followers of followers (i.e., multi-hop away) from early adopters without using item content and user profile information.

In this work, using Twitter as an example, we study the diffusion participation of hashtags among users. Note that participating in the diffusion of a hashtag is equivalent to adopting such hashtag. We will use *participants* and *adopters* interchangeably throughout this paper. Given a newly emerging hashtag in a social network along with its early adopters, we aim at forecasting the future diffusion participants of such hashtag among users who can be directly or indirectly connected with early adopters, using no textual content and user profiles. The diffusion participation forecasting problem is challenging due to the following three reasons. The first is *data sparsity*. From the retweeting records of users in Twitter, we find that most users participate in very limited diffusions, i.e., they adopt very few hashtags. Only a small set of users frequently and actively participate in various kinds of diffusions. Such limited diffusion records lead to severe data sparsity, and thus make us difficult to learn the behaviors of diffusion participation for most users. The second is forecasting the participation of multi-hop followers from the early adopters. Some of existing studies [2,25] divide the diffusion records of a certain hashtag into training and testing sets, use the training set to learn the predictive model, and predict users' adoption on the testing set for the same hashtag. Other studies [3,36] train the predictive models using the diffusion records of past hashtags, and predict the adoption of a new hashtag for "one-hop" followers of early adopters. Our goal is to learn the behaviors of diffusion participation from users' adoption of past hashtags, and forecast the users' adoption for a new hashtag. It is apparently difficult since we cannot directly learn the users' habits on the new hashtag. The third is forecasting without using item content or user profiles. We assume the content information (e.g., texts and images) is not always accessible. Therefore, the users' behaviors of diffusion participation can be learned based on only historical (e.g., the participation of past diffusions) and social information (e.g., social connections, and temporal clues). Such setting is much challenging compared with past models [8,37] that highly depend on content and profile info.

We develop an Adoption-based Participation Ranking (APR) model to find the diffusion participants based on the early adopters. The technical goal is to obtain a *participation probability* for each user, and use such probability to generate a ranking list of users such that the actual participants can be placed at highest positions. Our main idea is that a user v has higher probability to adopt hashtag c if (a) v prefers hashtag c or has higher willingness to adopt c, and (b) the diffusion of hashtag c can easily *reach* v. For (a), we construct a predictive model to estimate an *adoption probability* for each user v that models the preference or willingness of v to adopt hashtag c. To have an effective predictive model, six categories of features are proposed to characterize the hidden correlation between the early adopters and each target user. For (b), we devise a random-walk ranking model to generate a *participation probability* for each target v, based on both the derived adoption probability of v and the early adopters. Users with higher participation probability values are considered as the most potential participants who will adopt c.

Experiments conducted on a Twitter dataset disclose three insights. First, the proposed APR model can significantly outperform other competing methods in terms of Precision, Recall, and F1-measure. Second, the estimation of users' adoption probabilities, which model the behaviors and willingness of diffusion participation, plays a significant role for having accurate forecasting results. Third, jointly using early adopters and other users with higher estimated adoption probability values can lead to satisfying performance. Furthermore, we aim at applying the forecasted diffusion participants to the tasks of influence maximization and popularity prediction. The applications' evaluation exhibits two promising results. First, considering the forecasted diffusion participants as the seed candidates of influence propagation can lead to effective targeted marketing. This is achieved by identifying influential seeds and maximizing the influence spread with respect to the given hashtag. Second, treating the forecasted participants as additional adopters of information diffusion can significantly boost the performance of tweet popularity prediction. With such results, the power of the proposed diffusion participation forecasting is further proven.

Here we summarize the contributions of this paper as follows.

- We propose a novel information diffusion research problem, Diffusion Participation Forecasting (DPF), whose setting is more realistic and essentially different from existing studies in three folds: (1) forecasting the participants in multiple hops away from the early adopters, (2) making the prediction of new items from past diffusions, and (3) setting on no diffusion content and no user profiles.
- We develop an Adoption-based Participation Ranking (APR) model. Based on only the early adopters and the social network structure, APR jointly learns the adoption probability values of nodes and estimates the reachability scores of nodes to generate a ranking list as the forecasting result. In addition, we develop a robust collection of predictive feature sets (i.e., Diffusive, Social, Fringe, Temporal, First-Adopter, and Target) to learn the information adoption.
- Experiments conducted on a real Twitter dataset demonstrates APR can outperform baseline and state-of-the-art competitors, and demonstrate the effectiveness of the developed feature sets. The learned adoption probability of each node is also proven to be useful in boosting the performance.
- The proposed APR enables two applications, Forecasting-enhanced Influence Maximization and Adopter-expanded Popularity Prediction, in which the early adopters of information diffusion can provide more realistic settings for targeted marketing and tweet popularity prediction. Empirical studies show the forecasted diffusion participants generated by APR can boost the quality of such two applications.