Modeling and predicting retweeting dynamics via a mixture process

Jinhua Gao, Huawei Shen, Shenghua Liu and Xueqi Cheng
gaojinhua@software.ict.ac.cn, {shenhuawei, liushenghua, cxq}@ict.ac.cn
CAS Key Laboratory of Network Data Science and Technology, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, China

ABSTRACT

Modeling and predicting retweeting dynamics in social media has important implications to an array of applications. Existing models either fail to model the triggering effect of retweeting dynamics, e.g., the model based on reinforced Poisson process, or are hard to be trained using only the retweeting dynamics of individual tweet, e.g., the model based on self-exciting Hawkes process. In this paper, motivated by the observation that each retweeting dynamics is generally dominated by a handful of key nodes that separately trigger a high number of retweets, we propose a mixture process to model and predict retweeting dynamics, with each subprocess capturing the retweeting dynamics initiated by a key node. Experiments demonstrate that the proposed model outperforms the state-of-the-art model.

Keywords

retweeting dynamics; popularity prediction; mixture process

1. INTRODUCTION

Different from traditional media where people access information released generally by a single source, social media platforms, e.g., Facebook, Twitter, and Sina Weibo, offer a decentralized style of information spreading, i.e., information spreads among users via retweeting or forwarding. Consequently, popularity of individual information follows a highly-asymmetric distribution, and it is a challenging problem to predict the popularity of information. Retweeting dynamics prediction, i.e., predicting how popular a piece of information will become, attracts much research attention in the last decade. Existing methods make predictions either by exploring relevant factors and applying standard regression/classification methods, or by fitting retweeting dynamics using certain class of functions [1].

Recently, researchers begin to directly model the process that a piece of information gains its popularity. Shen et al. employed reinforced Poisson process (RPP) to model the arriving process of paper citations [2], and this model is subsequently extended to model the retweeting dynamics of microblog [3]. RPP assumes that information spreading is an arriving process of attention to a particular piece of information, simply using the count of retweets as the sole indicator of triggering effect of all retweets rather than modeling the triggering effect of each retweet separately. Self-exciting Hawkes process (SEHP) is then proposed to explicitly model the triggering effect of each retweet, with the arriving rate of retweets being determined by the aggregation of the triggering effect of all the previous retweets [4, 5, 6]. Although SEHP has the flexibility to capture the triggering effect of each retweet, it is impractical to learn the strength of triggering effect for each retweet when only the retweeting dynamics of individual tweet is available. Therefore, we still lack an effective method to predict retweeting dynamics in social media.

In this paper, we propose a mixture process to model and predict retweeting dynamics. Our model is motivated by the observation that the retweeting process of tweets could be generally characterized by a diffusion tree with only a handful of *key* nodes, each triggering a high number of retweets (Fig. 1A). Thus, the whole process of retweeting dynamics could be divided into several subprocesses, each of which is well modeled via a RPP model (Fig. 1B), and the whole process is taken as the aggregation of these subprocesses (Fig. 1C). The model based on such a mixture process has higher flexibility than RPP model, and could be much easily trained than SEHP model. Experiments demonstrate that the proposed model outperforms the state-of-the-art model at predicting retweeting dynamics.

2. MODEL AND VALIDATION

Retweeting dynamics of a tweet during a time period $[0, T]$ is denoted by a set of retweeting moments $\{t_i\}(1 \leq i \leq n)$, where n represents the total number of retweets up to time T. Without loss of generality, we have $0 = t_0 \leq t_1 \leq \cdots \leq t_i \leq \cdots \leq t_n \leq T$. We model the retweeting dynamics as a mixture process of RPP characterized by the rate function

$$x(t) = \sum_{l=1}^{k} \lambda_l f(t - \tau_l; \theta_l) c(t), \tag{1}$$

where k is the number of subprocesses, λ_l captures the strength of triggering effect initiated by the *source* retweeter of the l-th subprocess, τ_l is the moment when the l-th subprocess begins, $f(t; \theta_l)$ is the relaxation function that characterizes

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889389.

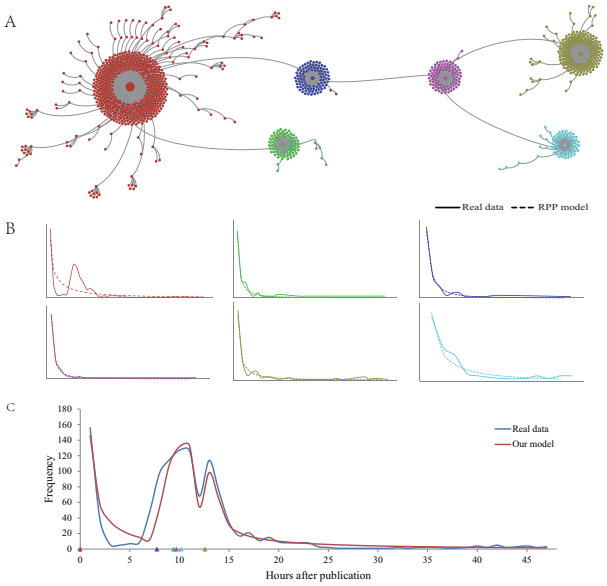

Figure 1: Retweeting dynamics of a message in Sina Weibo. (A) Di usion tree of the retweeting process; (B) Subprocesses, each being initiated by a key node; (C) Retweeting dynamics of the message.

up to time t. In this paper, we use log-normal function as the relaxation function as in [2].

The probability of the i-th retweet arrives at t_i, given the $(i-1)$-th retweet arrives at t_{i-1}, can be written as

$$p_1(t_i|t_{i-1}) = x(t_i)\exp\left(-\int_{t_{i-1}}^{t_i} x(t)dt\right), \tag{2}$$

and the probability of no retweet between $[t_n, T]$ is

$$p_0(T|t_n) = \exp\left(-\int_{t_n}^{T} x(t)dt\right). \tag{3}$$

Therefore, the likelihood of the retweeting dynamics $\{t_i\}_{i=1}^n$ during $[0, T]$ is

$$\mathcal{L} = p_0(T|t_n)\prod_{i=1}^{n} p_1(t_i|t_{i-1}). \tag{4}$$

The parameters (i.e., λ_l, θ_l, τ_l) are estimated by maximizing the log likelihood, and the retweeting dynamics of message could be predicted as

$$c(t) = n * \exp\left(\int_T^t \sum_{l=1}^{k} \lambda_l f(s-\tau_l;\theta_l)ds\right). \tag{5}$$

To evaluate the effectiveness of the proposed model, we compare it with the RPP model [2] on a dataset crawled from Sina Weibo, the largest microblogging mipl1(t)1(a)re,1(p). 5