Knowledge-Based Systems 147 (2018) 81–93

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Toward capturing heterogeneity for inferring diffusion networks: A mixed diffusion pattern model

Chang Su[a], Xiaohong Guan[a,b], Youtian Du[a,*], Xin Huang[a], Minhua Zhang[a]

[a] MOE Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, China
[b] Department of Automation, Tsinghua University, China

ARTICLE INFO

Article history:
Received 24 May 2017
Revised 16 January 2018
Accepted 8 February 2018
Available online 9 February 2018

Keywords:
Diffusion pattern
Diffusion network
Cascade model
Online social network (OSN)

ABSTRACT

Inferring diffusion network structure from observed cascades has attracted tremendous attention due to its utmost significance for many applications in online social network (OSN) analysis. Most previous studies assume that information diffuses with a uniform diffusion pattern. However, in OSNs, user interactions usually show different preferences and different speeds, and hence the diffusion processes are heterogeneous and show diverse diffusion patterns. It is difficult for traditional methods to capture the heterogeneity of information diffusion processes in OSNs. In this paper, we study the problem of inferring diffusion networks based on multiple latent diffusion patterns. To this end, we first analyze massive users' retweeting behaviors to investigate pairwise information transmissions. This analysis allows us to present a reasonable formulation of pattern-based pairwise information transmission probabilities to model the diffusion processes. Then, we incorporate multiple latent diffusion patterns into a probabilistic mixture model to infer diffusion network structures by fitting the observed cascades. We provide the estimation method of our proposed model based on Expectation Maximization (EM) algorithm. The results of experiments conducted on real OSN datasets demonstrate the superior performance of our model in inferring diffusion networks and show that our model can discover latent diffusion patterns effectively.

1. Introduction

Today, the Twitter[1]-like OSN sites provide fully flexible and free platform for information communication. Information diffusion study has become a fundamental issue of OSN analysis and provides many meaningful applications in multiple scenarios, such as viral marketing [1,2], rumor suppression [3,4], influence maximization [5,6], recommendation systems [7], political sentiment analysis [8] and social multimedia network structure analysis [9–11].

We think of the diffusion of information as a process taking place on a network which we call the diffusion network, where information propagates from node to node like an epidemic. But for some reasons, such as privacy limitation of OSN sites, we can only observe when a user adopts the information and cannot track the information diffusion pathway to determine who is infected by whom. In other words, edges of the diffusion networks are often unknown and need to be inferred. Therefore, to better understand information diffusion dynamics for forecasting, influencing and retarding infections, etc., inferring diffusion network structure from the observed diffusion cascades has become an important branch of information diffusion research. In this context, previous studies can be roughly grouped into two categories: works that focus on inferring the connectivity of diffusion networks [12,13] and those that attempt to infer not only diffusion networks but also pairwise transmission probabilities (i.e., the probability that information is transmitted from a user to another) [14–17]. Generally, these studies model the information diffusion processes in networks in a probabilistic view and then solve the inference problem by optimizing the likelihood functions to fit the observed diffusion cascades to obtain the diffusion networks (and pairwise transmission probabilities).

Existing methods normally model diffusion processes based on the assumption that information diffuses with a uniform diffusion pattern. In particular, it is assumed that the transmission probability from one node to another is fixed and that the transmission time delay follows a fixed distribution. Consequently, the generated diffusion results are *homogeneous*. Although traditional methods perform well for synthetic and traditional networks, their performance for OSNs, such as Twitter, are usually undesirable because studies on social theories [18–20] have found that people tend to

* Corresponding author.
E-mail addresses: changsu@stu.xjtu.edu.cn, csu@sei.xjtu.edu.cn (C. Su), duyt@mail.xjtu.edu.cn (Y. Du).

[1] http://twitter.com .

https://doi.org/10.1016/j.knosys.2018.02.017
0950-7051/

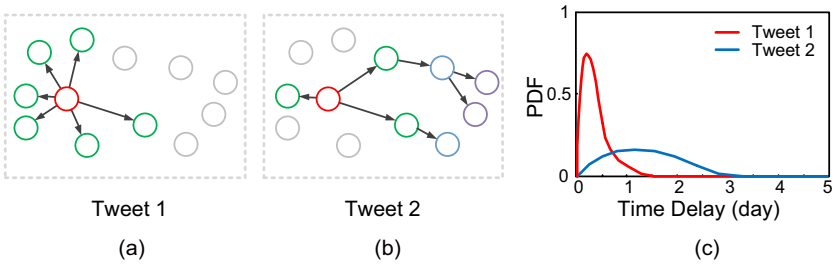

Fig. 1. A toy example of two diffusion processes with different diffusion patterns. (a) and (b) The spatial aspect: the red nodes denote posters who originally post tweets; the grey nodes are the users who ignore the tweets; and others denote the infectors that retweet the tweets, of which color denote their distance from the original poster. (c) The temporal aspect: the time-delay distributions of retweeting behaviors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

interact with different individuals with different speeds for different topics or during different time periods, and the information diffusion processes are actually *heterogeneous* and show diverse *diffusion patterns*.

Fig. 1illustrates a toy example of two diffusion results with different diffusion patterns from the viewpoints of both spatial (the structure of diffusion traces) and temporal (the time-domain features) aspects. As shown in Fig. 1(a) and (b), Tweet 1 causes intense interactions between the original poster and its neighbors, while the propagation of Tweet 2 tends to involve users far away from the original poster. As shown in Fig. 1(c), the time delay distributions for the diffusion processes indicate that Tweet 1 spreads much faster than Tweet 2. Although the toy example clearly shows two typical diffusion patterns with distinct spatial and temporal characteristics, in practice, it is difficult to identify diverse diffusion patterns in OSNs because the user interactions and diffusion processes are extremely complicated. The diffusion patterns are usually deep hidden in the massive sophisticated diffusion data and cannot be explicitly observed. Therefore, *it is challenging to infer diffusion networks by fully considering diverse diffusion patterns*. To address this challenge, we need a general diffusion cascade model that can accurately infer diffusion networks by automatically distinguishing the underlying diffusion patterns from the heterogeneous and complicated diffusion processes.

To model diffusion processes, a key point is how to handle the pairwise transmission probabilities. Several traditional studies [14–16] consider the transmission probability between two nodes as a parameter to learn in the diffusion models. All such probabilities constitute an enormous parameter space, and abundant observed diffusion cascades are required to learn each transmission probability separately, which inevitably leads to over-fitting. Other studies [17,21] have represented the pairwise transmission probability by a uniform function of node attributes. This representation avoids over-fitting in parameter learning due to the efficaciously reduced parameter space. However, factors influencing pairwise transmission probabilities in OSNs are more complicated than those in traditional networks, and diverse diffusion patterns would lead to variations of pairwise information transmissions. Consequently, *it is challenging to formulate the pattern-based pairwise transmission probabilities effectively in information diffusion modeling*. As the information transmission depends on the retweeting behaviors, a thorough understanding of users retweeting behaviors is needed to address the challenge of formulating pattern-based pairwise transmission probabilities.

In this paper, we study the inference of diffusion networks based on multiple diffusion patterns in Twitter-like OSNs and address the above two challenges. First, by analyzing massive users' retweeting behavior data, we formulate pattern-based pairwise transmission probabilities by a function over users' structural properties. Next, we discuss how to incorporate multiple diffusion patterns into a mixture model to infer diffusion networks from the observed cascades. We also provide an Expectation Maximization (EM) algorithm to effectively estimate the parameters of our proposed model. Finally, we conduct extensive experiments on real OSN datasets to evaluate the proposed model in terms of inferring diffusion networks and pairwise transmission probabilities as well as to analyze the discovered basic diffusion patterns.

The main contributions of this paper are summarized as follows:

- We define the problem of inferring diffusion networks by utilizing diverse latent diffusion patterns. In this way, we exploit a new perspective on information diffusion modeling in OSNs and capture the heterogeneity of diffusion processes.
- We present a formulation of pattern-based transmission probabilities over structural properties of users' social connection by analyzing retweeting behaviors from approximately 400,000 real information diffusion results. This formulation effectively describes the factors that affect real heterogeneous diffusion processes as well as avoids over-fitting in model learning due to the reduced parameter space.
- By incorporating diverse diffusion patterns and the pattern-based transmission probability formulation, we propose a mixed diffusion pattern cascade model. In this model, we consider that diffusion of information is driven by multiple basic diffusion patterns, and the final diffusion process is the result of superposition of the patterns. To effectively estimate parameters of our model, a learning algorithm based on EM framework is introduced.
- Implemented on real OSN datasets, our proposed model discovers several basic diffusion patterns with distinct spatial, temporal and semantic characteristics, which can be used for various practical applications of OSN analysis.

This paper is organized as follows. Section 2 presents a review of related work. In Section 3, we implement retweeting behavior analysis on real OSN diffusion data. In Section 4, we introduce the details of our model and present an EM algorithm for parameter inference. We implement our model on real OSN datasets and report the experimental results in Section 5. Finally, we conclude this paper in Section 6.

2. Related works

Many studies have sought to infer diffusion networks from observed diffusion cascades [12–17,22] in recent decades. These methods can be roughly divided into two categories. In the first category, NetInfer [12] and its variant, MulTree [13], focus on inferring the structure connectivity of diffusion networks and formulate the problem of inferring diffusion networks as the maximization of a submodular function. The second category includes