Knowledge-Based Systems 000 (2017) 1–11

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Predicting information diffusion probabilities in social networks: A Bayesian networks based approach

Devesh Varshney[a,b], Sandeep Kumar[a,*], Vineet Gupta[b]

[a] Department of Computer Science and Engineering, IIT Roorkee, India
[b] Adobe Research Labs, Bangalore, India

ARTICLE INFO

Article history:
Received 28 April 2017
Revised 28 June 2017
Accepted 1 July 2017
Available online xxx

Keywords:
Social network analysis
Information diffusion
Diffusion network
Bayesian network modeling
Diffusion probability

ABSTRACT

In past few years, social networking has significantly contributed to online presence of users. These social networks are hosts to a number of viral phenomena. This has fetched a lot of attention from various researchers and marketers all over the world. Major portion of the studies done in the field of information diffusion through social networks has focused on the problem of influence maximization. These methods demand the diffusion probabilities associated with the links in the social networks to be provided as inputs. However, the problem of computing these diffusion probabilities has not been as widely explored as the problem of influence maximization. In this paper, we tackle the problem of predicting the probabilities of diffusion of a message through the links of a social network. This paper presents a Bayesian network based approach for solving the aforesaid problem. In addition to the features related to the social network, this machine learning based Bayesian framework utilizes user interests and content similarity modeled using the latent topic information. We evaluate the proposed method using the data obtained from the well-known social network platform - Twitter.

1. Introduction

In present scenario, social networking and micro-blogging sites have become dynamic and widely used media for communication. Using these sites, people share information on various topics through which they express their likes and interests. As a result, social networking sites like Facebook[1] and Twitter[2] have shown a tremendous potential to make content viral, instantly, for example, Twitter during the United States presidential election in 2008 [1] and Facebook during the 2010 Arab spring [2]. Today, social networks are hosts to a number of viral phenomena: breaking news propagation, information dissemination during emergency, marketing campaigns, etc. [3,4]. This potential has caught the eyes of researchers as well as marketers, prompting them to focus on the word-of-mouth marketing strategy using these social platforms.

Studying and modeling the information propagation through social networks is important in making effective use of social platforms. It is not only helpful in understanding how information is diffused in the online social networks, but it can also be leveraged for solving a number of problems like influence maximization, personalized recommendation systems, trending topics detection, trust propagation, feed ranking in social networking sites, ad delivery, computing diffusion centrality measures in social networks [5–11].

Several researchers have studied the information diffusion through online social networks in the past. Most of the works in this area are based on the probabilistic models of information diffusion through networks, namely independent cascade (IC) and linear threshold (LT) [12]. These works assume the information diffusion probability (IDP) for each link in the network to be given as input. Some research works [7,13,14] have addressed the problem of predicting information diffusion, i.e., whether a user will retweet or not. On the other hand, the problem of computing IDP values has not been widely explored. Recently, some works [8,15–17] have reported the study of influence computation in social networks. Some of these works [8,13,16] use only network dynamics for creating solution models. However, the use of network dynamics alone cannot accurately capture the user interests and other relevant features [18]. Studies show that information dissemination processes are homophily-driven [19] and message propagation occurs more frequently between users having common interests [13]. Recently, Romero et al. [20] have proved that the textual

* Corresponding author at: Department of Computer Science and Engineering, IIT Roorkee, India.
E-mail addresses: devu.var@gmail.com (D. Varshney), sgargfec@iitr.ac.in, sandeepkumargarg@gmail.com (S. Kumar), vineetgupta10@gmail.com (V. Gupta).

[1] www.facebook.com
[2] www.twitter.com

http://dx.doi.org/10.1016/j.knosys.2017.07.003
0950-7051/

content plays a significant role in information propagation through social networks.

Taking these findings into consideration, we present a novel approach to predict the information diffusion probabilities of a message through various links in a social network. The approach uses textual content of messages, diffusion history of the network, and the network and user characteristics to build a Bayesian network model. The method exploits latent information extracted from the textual content to compute various features like user similarity, content similarity and user interests as described in Section 3.2. Given a social network and its diffusion history, the proposed approach finds the probabilities associated with the links of the social network for diffusion of a given message. Following are the contributions of this work:

- We analyze different factors which affect the IDPs and present a generic approach to compute these as features using the data available from online social networks.
- This work investigates the dependencies that exist among these features and how can we leverage this information to build a reliable model.
- This paper presents a Bayesian network based approach to compute the IDP values for different contents in a social network.
- We compare the previous studies in this field and demonstrate how latent topic information obtained from diffused message contents can be utilized to improve the state-of-the-art methods.
- The experimental evaluation of the proposed approach on real-world data demonstrates the effectiveness and efficiency of the method.

In this paper, we study the literature on information diffusion process from sociology and computer-science background and then perform experimental study to get better insights in order to answer the following research questions:

- **RQ1**: *What factors related to network settings, user characteristics, information source and message content affect the IDP values? How one can extract information out of the data available from online social networks to quantify these factors as features?*
- **RQ2**: *Whether these features are independent or there exist some dependencies among these and how one can formulate these dependencies?*
- **RQ3**: *How to design a model which can leverage these features and dependencies to reliably predict the IDP values in a social network for a given message?*

The rest of this paper is organized as follows. Section 2 provides background and discusses relevant works studying information diffusion. In Section 3, we present an approach for the computation of IDPs. Section 4 presents the experiments performed to evaluate the proposed method and the comparative analysis of the proposed approach with existing approaches. Section 5 presents the discussion of the experimental study. Finally, Section 6 concludes the paper with discussion of the proposed approach and scope for future works.

2. Related work

A number of researchers in the past have studied the process of information diffusion through social networks. These studies can be largely categorized into two parts. First, a significant amount of these works focus on addressing the problem of *Influence Maximization*. The goal of *Influence Maximization* is to find a seed set of users who can trigger cascades for maximizing the spread of an idea or opinion in the social network. For the first time, Domingos and Richardson [5,21] addressed this problem. They proposed a probabilistic model using Markov random fields. Later on, Kempe et al. [12] modeled the same as discrete optimization problem and proved its NP-hardness for two basic diffusion models namely–independent cascade and linear threshold. They also proposed a greedy approximation algorithm for solving the same. Most of the works studying the problem of *Influence Maximization* assume the IDPs of the links are given as inputs. In this work, instead we address the problem of predicting these probabilities. Second, some of these works aim to study the prediction of information diffusion, which includes determining whether a link will be active or not and what IDPs to assign to such links in social networks. Fei et al. [13] used a multi-task learning approach to predict a user's response (like or comment) to a post of her friend. They used features representing content similarity and user interests in their approach. Lin et. al [22]. proposed a probabilistic model TIDE (text-based information diffusion and evolution), to track the evolution of a topic with time in social communities and its diffusion paths. Their model extract features from text of posts and captures implicit features using Gaussian random field. But, this model ignores the features related to social connections.

Zhu et al. [15] studied the retweeting behavior of users and presented a logistic regression model to predict the retweeting probability of the incoming tweet by the target user. They used features related to the network such as the numbers of friends, followers, mutual friends, mutual followers, mutual mentions, mutual retweets, and the status count of the tweet author to capture the relationship among users. The model also takes into account the timing of tweets. They modeled content influence using URLs, mentions, and hash-tags in the tweet. The work captured topic similarity between the incoming tweet and the tweets of the user as the cosine similarity of their term frequency vectors. However, term frequency vectors are very sparse due to diverse use of vocabulary of interacting users.

Kuo et al. [14,23] have addressed the diffusion prediction on novel topic problem to predict both cross-topic-observed and unobserved diffusions. They used the latent information present in the posts to model a signature for each of the topics defined in the topic set and to model the user preferences towards these topics. These methods consider the feature related to the topic of the message but not the features related to the content of that message. Thus, the diffusion is specific to a topic irrespective of what is exact content of the message. However, we argue that the content of a message is also significant in the process of diffusion, thus we consider the message content for feature modeling. Varshney et al. [24] also addressed the problem of detecting the links which are active in the propagation of a given message through a social network. The works discussed above do not provide the probability of diffusion of the message through the link as the models used are classification based and not the probabilistic ones.

Similar to our work, the works in [8,16] also explored the problem of finding information diffusion probabilities. Saito et. al [16]. presented a method for predicting information diffusion probabilities for independent cascade (IC) model. This work defined the likelihood of multiple diffusion episodes and applied the Expectation Maximization (EM) algorithm to solve it. But this method is not scalable to huge datasets because EM algorithm has to update the diffusion probability of each link in each iteration. Goyal et al. [8] proposed models for learning the probabilities of influence between users with the focus on linear threshold (LT) model. They proposed various models like static model of probabilities, continuous time model and discrete time model. They also presented a technique for predicting the time by which a user is expected to perform the action. Due to the assumptions taken in their method such as a user performs an action at most once and the influence graph is DAG, their method is not suitable for tweet-retweet networks. Both of these works [8,16] ignore the content of posts to

Please cite this article as: D. Varshney et al., Predicting information diffusion probabilities in social networks: A Bayesian networks based approach, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.07.003