Expert Systems with Applications 41 (2014) 5115–5124

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Social network user influence sense-making and dynamics prediction

Jingxuan Li [a,c], Wei Peng [b], Tao Li [a,*], Tong Sun [b], Qianmu Li [c], Jian Xu [c]

[a] School of Computer Science, Florida International University, 11200 SW 8th St, Miami, FL, USA
[b] Xerox Innovation Group, 800 Phillips Rd, Webster, NY, USA
[c] Nanjing University of Science and Technology, Nanjing 210094, PR China

ARTICLE INFO

Keywords:
Social media
Influential users
Dynamic information diffusion
Continuous-Time Markov Process

ABSTRACT

Identifying influential users and predicting their "network impact" on social networks have attracted tremendous interest from both academia and industry. Various definitions of "influence" and many methods for calculating influence scores have been provided for different empirical purposes and they often lack the in-depth analysis of the "characteristics" of the output influence. In addition, most of the developed algorithms and tools are mainly dependent on the static network structure instead of the dynamic diffusion process over the network, and are thus essentially based on descriptive models instead of predictive models. Consequently, very few existing works consider the dynamic propagation of influence in continuous time due to infinite steps for simulation. In this paper, we provide an evaluation framework to systematically measure the "characteristics" of the influence from the following three dimensions: (i) *Monomorphism vs. Polymorphism*; (ii) *High Latency vs. Low Latency*; and (iii) *Information Inventor vs. Information Spreader*. We propose a dynamic information propagation model based on *Continuous-Time Markov Process* to predict the influence dynamics of social network users, where the nodes in the propagation sequences are the users, and the edges connect users who refer to the same *topic* contiguously on time. Finally we present a comprehensive empirical study on a large-scale twitter dataset to compare the influence metrics within our proposed evaluation framework. Experimental results validate our ideas and demonstrate the prediction performance of our proposed algorithms.

1. Introduction

1.1. Identifying influential users

Social network analysis has been gaining attention from different domains, including economics, anthropology, biology, social psychology, physics, etc.

The rapid growth of the online social network sites (e.g. Facebook, Twitter, LinkedIn, and Google+) and their publicly available data acquiring API has led the prosperity of social network analysis research these days. One of most popular topics of the social network analysis is identifying influential users and their "network impact". Knowing the influence of users and being able to predict it can be leveraged for many applications. The most famous application to researchers and marketers is viral marketing (Domingos & Richardson, 2001; Kempe, Kleinberg, & Tardos, 2003; Richardson & Domingos, 2002), which aims at targeting a group of influential users to maximize the marketing campaign ROI (Return of Investment). Other interesting applications include search (Adamic & Adar, 2005), expertise/tweets recommendation (Song, Tseng, Lin, & Sun, 2006, 2007, 2010), trust/information propagation (Gruhl, Guha, Liben-Nowell, & Tomkins, 2004; Golbeck & Hendler, 2006), and customer handling prioritization in social customer relationship management.

1.2. Limitations of current research efforts

There are two main limitations of current research efforts on identifying influential users in social network analysis: one is on the characteristics of influence, and the other is on the influence models and measures.

1.2.1. Characteristics of influence

Various definitions of "influence" and many methods for calculating influence scores have been provided for their own empirical purposes, or applications. Since they often lack the in-depth analysis of the "characteristics" of the output influence. it is difficult to adapt or choose them for other applications.

* Corresponding author. Tel.: +1 305 348 6036; fax: +1 305 348 3549.
E-mail addresses: jli003@cs.fiu.edu (J. Li), wei.peng@xerox.com (W. Peng), taoli@cs.fiu.edu (T. Li), tong.sun@xerox.com (T. Sun), qianmu@njust.edu.cn (Q. Li), dolphin.xu@njust.edu.cn (J. Xu).

http://dx.doi.org/10.1016/j.eswa.2014.02.038
0957-4174/

1.2.2. Influence models and measures

Currently most applications and tools compute user influence based on their static network properties, such as, the number of friends/followers in the social graph, the number of posted tweets/received retweets/mentions/replies in the activity graph, or users' centrality (e.g. PageRank, Betweeness-centrality, etc.).

A few works investigate adoption behaviors of social network users as the dynamic influence propagation [1] or diffusion process (Rogers, 2003). The adoption behaviors refer to some activities or topics (tweets, products, Hashtags, URLs, etc.) shared among users implicitly and explicitly such as users forwarding a message to their friends, recommending a product to others, joining some groups with the similar musical favor, and posting messages about the same topics, etc. According to the diffusion theory, the information cascades from social leaders to followers. In most diffusion models, propagators have certain probabilities to influence their receivers, and the receivers also have certain thresholds to be influenced. Finding the social leaders or the users who can maximize the influence coverage in the network is the major goal of most diffusion models.

Some drawbacks of existing social network influence models based on either static networks or the "influence maximization" diffusion process are: (1) The static influence scores are not actionable for users. For example, marketers do not know what will be the difference if targeting users with influence scores of 30 or 80. (2) Most existing models are descriptive models rather than predictive models. For example, the number of friends or the centrality score of a given user describes his/her underlying network connectivity. The number of tweets that a user posted or get retweeted indicates the trust/insterest that his/her followers have on his/her tweets. All these measures/models are descriptive and very few models are able to predict users' future influence. (3) Existing "influence maximization" diffusion process is often modeled by discrete-time models such as Independent Cascade Model or Linear Threshold Model. Because the real world diffusion process is continuous-time, it is difficult to define an appropriate time step t for discrete-time models.

1.3. Content of the paper

The aforesaid limitations motivate our study on social network user influence and dynamics prediction in this paper. In particular, to address the first limitation, we take an initial step to introduce three dimensions of influence: (i). *Monomorphism vs. Polymorphism*; (ii). *High Latency vs. Low Latency*; and (iii). *Information Inventor vs. Information Spreader*, for understanding the characteristics of influential users calculated from various methods. These three dimensions provide an evaluation framework to systematically measure the influence.

To address the second limitation, we propose a dynamic information diffusion model based on the *Continuous-Time Markov Process* (CTMP) to predict the influence dynamics of social network users. CTMP assumes that the number of activations from a given node is following an exponential distribution over the time. This can be often seen in the real-world data (Kwak, Lee, Park, & Moon, 2010). Fig. 1 shows that the average number of topic adoptions decreases exponentially over the time. Hashtags receive more adoptions compared with URLs, and the number of Hashtag adoptions decreases more slowly. Furthermore, transition rates q are calculated and treated as the transition probabilities (or activation probability) of the embedded Markov chain in CTMP. Then the transition probability $P(t)$ can be computed from q, given any time t. In this paper, the nodes in the propagation sequences are the users, and the edges connect users who refer to the same topic contiguously on time. Topics here particularly refer to Hashtags (expressed as # followed by a word) and short URLs (e.g. bit.ly, TinyURL, etc.) on twitter, which is one of the most popular microblog services, was launched since July 13, 2006. Hashtags and URLs are both unique identifiers tagging distinct tweets with certain "topic" labels. We regard the temporal sequences of Hashtags and URLs as the diffusion paths, where the topics are reposted subsequently. Although retweeting is not included in our paper as a diffusion approach, it is implicitly considered because the retweets would usually contain the same Hashtags and URLs as in the original tweets. Our experimental results on a large-scale twitter dataset show that our proposed diffusion model outperforms other influence models for viral marketing. It also demonstrates a promising prediction performance on estimating the number of influenced users within a given time.

1.4. Paper contribution and organization

A preliminary study of the work has appeared at the 15th Asia–Pacific Web Conference in 2013 (Li, Peng, Li, & Sun, 2013). In that conference paper, the study focuses on the proposed influence model – IDM-CTMP, and shows its advantages over two baselines, which are not necessarily continuous-time models. In this journal submission, (1) we propose three "dimensions" of users' influence in the social network to help others understand different aspects of influence; (2) we conducted comprehensive experiment to systematically measure users' influence and compare different influence models over three proposed dimensions; (3) two heuristic continuous-time influence models are defined as baselines to further show the advantages of our proposed model.

In summary, the contributions of this paper are listed below.

1. We introduce three dimensions on application perspectives and provide an evaluation framework to systematically measure the influence and compare different influence models (See Section 5.3).
2. Comprehensive experiments are conducted on various extracted networks (mentions, retweets, replies), as well as temporal propagation paths from the large-scale twitter data (See Section 5).
3. Two heuristic influence models considering the topic diffusion in continuous time are defined as baselines (See Section 3) to highlight the strengths of our proposed dynamic information diffusion model based on the *Continuous-Time Markov Process*.

The remainder of this paper is organized as follows. Before discussing about any influence models, we propose three dimensions of social influence in Section 2. After, in Section 3, we first give the definition of the temporal influence network, introduce some existing influence models, and propose two heuristic dynamic influence models. In Section 4, we propose an information diffusion model based on the *Continuous-Time Markov Process*. Experimental results are demonstrated in Section 5. In particular, we discuss the three dimensions of influence and present a comprehensive empirical study on a large-scale twitter dataset to compare the influence metrics (including both the dynamic influence metrics and well-known static influence metrics) within our proposed evaluation framework in Section 5.3. We evaluate the prediction power of our proposed information diffusion model in Section 5.4. Related work on influence modeling is reviewed in Section 6. Finally Section 7 concludes the paper.

2. Three dimensions of influence

Everyone is talking about how to identify users with high influence, because it is believed that influential users can help with

[1] In this paper, we use "information/influence propagation", "information/influence diffusion", and "information cascade", interchangeably to represent the same concept.