Computers in Human Behavior 52 (2015) 184–189

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Measuring influence in online social network based on the user-content bipartite graph

Zhiguo Zhu [a,b,*], Jingqin Su [a], Liping Kong [c]

[a] Faculty of Management and Economics, Dalian University of Technology, 116023 Dalian, PR China
[b] School of Management Science and Engineering, Dongbei University of Finance and Economics, 116025 Dalian, PR China
[c] Vocational and Technical College, Dongbei University of Finance and Economics, 116025 Dalian, PR China

ARTICLE INFO

Article history:
Available online 15 June 2015

Keywords:
Online social network
Bipartite directed graph
Influence measurement
Markov model

ABSTRACT

With the rising of online social networks, influence has been a complex and subtle force to govern users' behaviors and relationship formation. Therefore, how to precisely identify and measure influence has been a hot research direction. Differentiating from existing researches, we are devoted to combining the status of users in the network and the contents generated from these users to synthetically measure the influence diffusion. In this paper, we firstly proposed a directed user-content bipartite graph model. Next, an iterative algorithm is designed to compute two scores: the users' Influence and boards' Reach. Finally, we conduct extensive experiments on the dataset extracted from the online community Pinterest. The experimental results verify our proposed model can discover most influential users and popular broads effectively and can also be expected to benefit various applications, e.g., viral marketing, personal recommendation, information retrieval, etc.

1. Introduction

The idea of studying a person's behavior in the context of his/her social connections is quite old as can be seen in Stanley Milligram's experiments which lead him to conclude that the average shortest path length between any two people in the world is about six (Milgram, 1967). However, nowadays online social networks have rapidly become very important hubs of social activity and conduits of information. Popular social sites such as Facebook, Pinterest and Twitter have undergone explosive growth and are turning into community spaces, where users interact with their friends and acquaintances. With the numbers of active users on these sites numbering in the millions or even tens of millions, how to identify and measure the influential users and topics becomes an important problem with applications in marketing (Kempe, Kleinberg, & Tardos, 2003), information dissemination (Gruhl & Liben-nowell, 2004; Leskovec et al., 2007), social relation visualization (Kim, Ji, & Park, 2014), and expertise discovery (Davitz, Yu, Basu, Gutelius, & Harris, 2007).

Given this widespread generation and consumption of content in online social networks, it is desirable to target one's messages to highly connected people who will propagate them further in the social network. In spite of the seemingly chaotic fashion with which all these interactions take place, certain topics manage to get an inordinate amount of attention, thus bubbling to the top in terms of popularity and contributing to new trends and to the public agenda of the community. There is considerable consensus on the fact that two aspects of information transmission seem to be important in determining the influence in social network.

One aspect is the popularity and status of given users in these social networks, which can be measured by the level of attention they receive in the form of followers who create links to their accounts to automatically receive the content they generate. This can also be generally viewed as the degree of user (Bonchi, Castillo, Gionis, & Jaimes, 2011; Mislove, 2009; Valente, 2010). The other aspect is the influence of contents that these users wield, which is determined by the actual propagation of these contents through the network. This influence is determined by many factors, such as the novelty and quality of these contents, and the frequency at which they generate these contents. Therefore, we are interested in combining the status of users in the network and the contents generated from these users to synthetically measure the influence diffusion existed in online social network.

In this paper, we employ the user-content bipartite graph to construct the model and then develop an iterative algorithm to quantify the influence in the network. With the example of Pinterest, which is one of the fastest growing social networks on

* Corresponding author at: Faculty of Management and Economics, Dalian University of Technology, No.2 Linggong Road, Ganjingzi District, Dalian, PR China. Mobile: +86 1594088193.
E-mail address: zhuzg0628@126.com (Z. Zhu).

http://dx.doi.org/10.1016/j.chb.2015.04.072
0747-5632/

the Internet and is becoming the focus of advertising companies and brands eager to exploit this vast new medium, we try to understand how the influence is determined by analyzing the propagation of 'pins' on Pinterest. In a word, our influence measure utilizes both the structural properties of the network and the contents the users published.

The rest of this article is organized as follows. The literature background and related work are investigated in Section 2. Section 3 formalizes the problem of measuring influence in mathematical terms and then proposes a directed bipartite graph model. Next, we design an iterative algorithm for the model to discover influential users and popular broads in Section 4. Section 5 presents experimental results that validate the effectiveness of our methodology. Last Section 6 concludes the paper and discusses the possible directions in the future research.

2. Related work

Since the advent of online social networks, especially in the past decade or so, there has been a lot of interest in measuring influence and modeling information flows in this new platform. Kleinberg et al. have studied the problem for quite some time and many of the results and concepts from their research are explained in Crandall, Cosley, Huttenlocher, Kleinberg, and Suri (2008), Backstrom, Huttenlocher, Kleinberg, and Lan (2006). Using a dataset of Wikipedia edits, Kleinberg et al. showed that influence (action in a user is triggered by one of his/her friends' recent actions) and selection (people prefer friends who share similar interests and hence they perform similar actions) often mutually enhance each other. Holme and Newman (2006) later improved upon this Kleinberg's study and suggested a probabilistic model for modeling this interaction. Another study that addresses the problem of interplay of selection and influence is presented in Christakis and Fowler (2007) where Fowler and Christakis study the spread of obesity in a network of around 12,000 people over a period of three decades. Kumar, Mahdian, and Anagnostopoulos (2008) have devised simple tests to identify social influence at play in an online community by tests such as the Shuffle Test and the Edge reversal test (also used by Fowler and Christakis).

While it assumes that influence exists as a real phenomenon in online social network, questions have been raised on whether there is evidence of genuine influence in real social network data. Anagnostopoulos, Kumar, and Mahdian (2008) have developed techniques for showing that influence may not be genuine: while there is substantial social correlation in tagging behavior, it cannot be attributed to influence. Another work highlighting the importance of separating influence-based contagion from homophily-driven diffusion is Aral (2010), where it is observed that the former can be overestimated if not measured correctly. Moreover, the strength of the different factors affecting the propagation of a piece of information may vary depending on what type of information (e.g., news, or discussion topic) is being propagated (Aral, Muchnik, & Sundararajan, 2009).

On the other hand, many researchers have designed some algorithms for quantitatively calculating the influence existed in online social networks. "Centrality" (Newman, 2010) is a fairly well studied concept in the context of networks and most measures of influence are more refined versions of one or more of these centrality measures. "PageRank", developed by Page and Brin (1998), has been one of the most influential measures developed in this area. Although PageRank is primarily used to rank sources of information in information networks (such as ranking web pages on the WWW), the algorithm has inspired various modifications to be applied specifically to social networks such as SimRank (Widom & Glen, 2002), Topic sensitive PageRank (Haveliwala, 2003) and TwitterRank (He, Weng, Jiang, & Lim, 2010). Jiawei Han's group (Han, Ming, & Danilevsky, 2011; Han, Sun, & Yu, 2009; Sun, Han, Zhao, Yin, & Cheng, 2009) has published several papers in the area of ranking nodes in heterogeneous networks, which after some adaptations, might be used to measure influence. Tang, Sun, Wang, and Yang (2009) introduce the novel problem of topic-based social influence analysis. They propose a topical-affinity propagation approach to describe the problem. Zhu (2013) build a model of information diffusion oriented for viral marketing and propose a dynamic algorithm of discovering the influential users in the process of information diffusion. Another popular class of algorithms has been improvements on the "HITS" algorithm proposed around 1998 by Kleinberg (1998). Kleinberg himself proposed a possible application of HITS to identify "important" people in a social network (Kleinberg, 1999). He suggested that HITS can be used to calculate "standing, impact or influence" of a node in such a network. This idea was developed further by Romero, Galuba, Asur, and Huberman (2011a) who came up with the IP algorithm and tested it on a Twitter dataset.

In this paper, we consider that the influence of a user thus depends not only on the size of the influenced audiences, but also on their generated contents. So, our will proposed a model based on an iterative algorithm to quantify the influence in the network, which is inspired by HITS (Kleinberg, 1998) and the Influence Passivity algorithm (Romero et al., 2011a). However, there is an important different thought: the prior algorithms treated networks as homogeneous whereas we consider heterogeneous networks.

3. Problem statement and model construction

3.1. Background about the community of Pinterest

Pinterest is a pinboard-style content sharing platform that allows users to exhibit collections of images or videos. To better present the proposed model, we briefly describe key terminologies in Pinterest below.

- Pin/Repin: Each image/video is called as a pin, and the act of posting a pin is referred to as pinning. If a posted pin is shared by someone, the shared pin is called as a repin. Users who posts and shares (i.e., repins) a pin are the original pinner and repinner, respectively.
- Like/Comment: Similar to Facebook, a user can push a like button for a pin that she likes and leave a comment on a pin.
- Board/Category: A board is a collection of pins organized by a user. Each board belongs to one of the categories in Pinterest.
- Following/Follower: Like many social platforms, the relation between two users in Pinterest is not symmetric. The fact that user A follows user B does not necessarily mean B follows A. If A follows B, A can see the updated news (e.g., the act of posting a new pin) of B.

3.2. A bipartite directed graph model

In this section, we formalize our problem of measuring influence in mathematical terms. We define precisely what do we mean by various terms, what data we have and what do we attempt to calculate by our algorithm.

We propose a bipartite directed graph model $G(V, E, W)$ based on the dataset described above with the following properties:

(1) The vertex set V has two types of vertices representing users and boards (representing contents) respectively. Mathematically $V = U \cup B$. $U = \{u_1, u_2, \ldots, u_m\}$ is the set of vertices representing users, in which m is the number of users in this online community and $B = \{b_1, b_2, \ldots, b_n\}$ is