Computer Networks 75 (2014) 544–559

Contents lists available at ScienceDirect

Computer Networks

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

Whom to follow: Efficient followee selection for cascading outbreak detection on online social networks

Junzhou Zhao [a,*], John C.S. Lui [b], Don Towsley [c], Xiaohong Guan [a]

[a] MOEKLINNS Lab, Xi'an Jiaotong University, China
[b] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
[c] School of Computer Science, University of Massachusetts at Amherst, United States

ARTICLE INFO

Article history:
Received 30 November 2013
Received in revised form 8 July 2014
Accepted 11 August 2014
Available online 5 October 2014

Keywords:
Follow model
Followee selection
Outbreak detection
Submodularity

ABSTRACT

Online social networks (OSNs), such as Twitter and Sina Weibo, have become important platforms for generating and spreading information on the Internet. On these OSNs, the "follow model" has become a popular way to discover information; i.e., a user subscribes to content generated by others by following them as information sources. The content producers are called followees. Due to human beings' limited attention capacity and the constraints imposed by OSNs, a user can only follow a few followees. The question then arises: which subset of followees shall we follow so that we can discover the most information in an OSN in a timely fashion? To solve this problem, we present a randomized method that does not require complete OSN data and is well suited for third parties who do not own OSN data. Our method is based on the birthday paradox and is mathematically tractable for analysing its solution quality and computational efficiency. Moreover, we find that the power-law structure of real-world OSNs can further improve the solution quality of our method. Experiments conducted on two real datasets demonstrate that our method can create a good trade-off between solution quality and computational efficiency.

1. Introduction

As platforms for communicating with friends, updating status and sharing information, online social networks (OSNs), such as Twitter and Sina Weibo, have become extremely popular. These platforms provide users with near-real-time services that can be accessed across multiple devices at any Internet-enabled venue. Due to their large user bases and ubiquitous services, microblogs, where users act as sensors reporting events happening around them, become essential news sources, i.e., the so-called social media [1,2]. In fact, social media have attracted surveillance from conventional media outlets to discover breaking news and from governments to detect signals of riots. For example, the death of Osama bin Laden was first reported on Twitter rather than traditional news media [3], and, during the period of England riots, people used social media to organize [4].

The emergence of social media has changed the way we discover information. Traditional ways, such as information retrieval, rely on user-specified queries (e.g., keywords) to retrieve the information from indexed data [5–7]. However, specifying explicit queries might be difficult, as keywords for time-evolving and emerging information are highly dynamic [8] and unpredictable (e.g., the death of bin Laden [3]). In recent years, the follow model [9] has become a convenient way to discover information. A microblog user, say, Alice, obtains information mainly from her timeline, which comprises tweets generated by

* Corresponding author. Tel.: +86 29 8266 3330; fax: +86 29 8266 4603.
E-mail address: junzhouzhao@gmail.com (J. Zhao).

http://dx.doi.org/10.1016/j.comnet.2014.08.024
1389-1286/

users she follows, called her *followees*[1] in the follow model. Once a new tweet is posted by a user, it spreads to the user's followers and their followers iteratively, depending on whether users retweet the message. Finally, the tweet appears in Alice's timeline with a certain *probability* and *time delay*. Such probability and time delay mainly depend on which subset of followees Alice chooses to follow. By choosing different followees as her information sources, Alice can discover varied information with different time delay from the aggregated tweets in her timeline.

In the current information era, common goals that people want to achieve are to discover as much information as possible, i.e., maximize information coverage, and to obtain the information as soon as possible, i.e., minimize time delay. To achieve this, it seems that if Alice can follow all microblog users, then she will discover all information with zero time delay. However, due to human beings' limited attention capacity [11] and the constraints imposed by OSNs (e.g., a user on Twitter and Sina Weibo can follow at most 2000 users, in general [12,13]), Alice can only follow a few followees (or budgeted followees). Consequently, a problem arises: how to optimally choose these budgeted followees as information sources to maximize information coverage and minimize time delay?

The above problem is challenging. Selecting a subset of items from a population to maximize some specified utility function is a classical combinational optimization problem that has been studied for decades, e.g., the set cover problem [14], knapsack problem [15], influence maximization problem [16,17], and sensor placement problem [18,19]. These problems have been proven to be NP-hard, and we can only obtain suboptimal solutions using approximate algorithms. However, as we will see in Section 3, these algorithms cannot be applied to our problem and they do not scale to handle modern large-scale OSNs which have hundreds of millions of users.

Another challenge we face is that we are constrained to solving the problem from the perspective of a third party. A third party does not own OSN data. OSN companies own users' data, but there is a lack of cooperation from OSN companies due to user privacy and business secrecy concerns. Thus, third parties can only use the public APIs to crawl the data. However, OSN companies usually impose barriers to limit large-scale crawling by third parties [20] and restrict the request rate of APIs. For example, Twitter and Sina Weibo allow a user to issue at most 350 and 150 requests per hour, respectively [21,22]. As a result, it is practically impossible for third parties to crawl the complete data, and one has to consider the query cost (i.e., the number of API calls) while achieving the goal of selecting budgeted followees. We would like to note that most of the existing works [18,19,23] have ignored this second challenge and assumed that the complete data are available in advance. This assumption limits the practical application of existing methods, and the goal of this work is to fill this research gap.

[1] The Oxford dictionary defines a followee as a person who is being tracked on a social media website or application [10].

In this work, we present a framework to select a subset of users as followees to maximize the information coverage and minimize the time delay from incomplete data obtained via graph sampling methods. Our method guarantees both solution quality and computational efficiency (under the worst-case situation) that enable it to be used in large-scale OSNs. (Note that the proposed approach does not replace but instead supplements existing methods, and we elaborate on this point in Section 8). The basic idea behind our method is based upon the *birthday paradox*, which states that with more than 50% chance, there will be a birthday match among a handful of 23 people, and, for merely 70 people, the chance of matching increases to 99.9%. In our scenario, the randomized greedy algorithm, which is the main component of our framework, chooses one user in each iteration from a set of user samples, which is substantially smaller than the population, and the user samples contain at least one optimal followee with high probability according to the birthday paradox. Due to the significant reduction of search space, we achieve a major speedup in obtaining the solution. The quality of the final solution can be proven to be lower bounded when user samples are chosen uniformly at random in each iteration, which actually is the worst-case situation because we do not use any strategy in sampling (see Section 5).

Moreover, we find that if we bias user samples toward high degree nodes (i.e., users) in the network using graph sampling methods such as random walk [24], we need fewer user samples than when using uniform sampling, thereby improving efficiency. We present an in-depth analysis in Section 6 and reveal another important finding. For power-law networks, information cascades are not uniformly dispersed among nodes, but rather, high degree nodes are more likely to be infected by an information diffusion process than low degree nodes. Therefore, if the sampling is biased toward high degree nodes, we achieve a higher probability of detecting information cascades. This is related to the generalized birthday paradox, i.e., when people's birthdates are not uniformly distributed, the probability of matching increases [25]. Our numerical solutions in Section 6 and experiments in Section 7 both demonstrate the finding.

The rest of the paper is structured as follows. We review the related literature in Section 2 and formulate the problem in Section 3. Then, we motivate a randomized method through empirical observations in Section 4. The detailed analysis of our method is given in Section 5. In Section 6, we study how the power-law structure of real-world networks can benefit this method. We conduct experiments on real datasets to validate the method in Section 7 and conclude in Section 8.

2. Related work

Both Twitter and Sina Weibo provide the "whom-to-follow" services to recommend "interesting persons" to users [26]. This function is related to a large body of research on link prediction [27]. However, algorithms in link prediction are mainly based on common friends, shared interests, and