Knowledge-Based Systems 141 (2018) 211–221

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Identifying topical influencers on twitter based on user behavior and network topology[☆]

Zeynep Zengin Alp[a,*], Şule Gündüz Öğüdücü[b]

[a] Institute of Science and Technology, Istanbul Technical University. Maslak, Istanbul, 34469, Turkey
[b] Department of Computer Engineering, Istanbul Technical University. Maslak, Istanbul, 34469, Turkey

ARTICLE INFO

Article history:
Received 26 January 2017
Revised 14 November 2017
Accepted 20 November 2017
Available online 21 November 2017

Keywords:
Social influence
Topical influence
Twitter
User modeling
PageRank
Topic modeling
LDA
Evaluation metrics
Viral marketing

ABSTRACT

Social media web sites have become major media platforms to share personal information, news, photos, videos and more. Users can even share live streams whenever they want to reach out to many other. This prevalent usage of social media attracted companies, data scientists, and researchers who are trying to infer meaningful information from this vast amount of data. Information diffusion and maximizing the spread of words is one of the most important focus for researchers working on social media. This information can serve many purposes such as; user or content recommendation, viral marketing, and user modeling. In this research, finding topical influential/authority users on Twitter is addressed. Since Twitter is a good platform to spread knowledge as a word of mouth approach and it has many more public profiles than protected ones, it is a target media for marketers. In this paper, we introduce a novel methodology, called *Personalized PageRank*, that integrates both the information obtained from network topology and the information obtained from user actions and activities in Twitter. The proposed approach aims to determine the topical influencers who are experts on a specific topic. Experimental results on a large dataset consisting of Turkish tweets show that using user specific features like topical focus rate, activeness, authenticity and speed of getting reaction on specific topics positively affects identifying influencers and lead to higher information diffusion. Algorithms are implemented on a distributed computing environment which makes high-cost graph processing more efficient.

1. Introduction

Social media has become a key media for sharing personal information, photos, videos, news and many more. Users spend several hours on social media daily.[1] This attracts attention of marketers, and businesses as a media to reach many people with no cost. Mostly researched areas of social media mining is influence analysis [1–5], and sentiment analysis [6–9]. Both areas can leverage marketing activities of companies. Similar to influence analysis, researchers also evaluate information diffusion models on social media to identify key users who increase diffusion of information. Recommender systems can leverage this information by recommending topical authorities to users who are interested in specific topics. Marketers can also use this information to spread their company information or campaigns on social media. Brand awareness, and campaign performance can increase drastically with the spread of information and using influential users is one of the most important ways to achieve this goal.

Influence has been broadly studied in areas like sociology, psychology, political science, medicine predating 1950s. Psychologists worked on influence analysis on subjects like effect of judgment, leadership [10,11], self-esteem [12]. Studying sociological and psychological aspects of influence help us better understand how certain trends spread more and faster than the others and who are the key influencers on these trends. Katz at al.[13] theorized that a few users called “influencers” can create a chain-reaction of influence that is based on word-of-mouth approach and reach to a very large scale of users. This idea is similar to viral marketing where companies try to reach as many customers as possible over social media with a low cost. In this work, we define social influence as the effect of users on others that results with sharing information, which is the most common definition of social influence [4,5,14]. Similarly, Aral et al. [15] recognizes social influence as a key factor in the propagation of ideas, behaviors, and economic outcomes in society.

☆ This research is partially funded by 2211 - TUBITAK BIDEB Ph.D. Scholarship Fund.

* Corresponding author.
E-mail addresses: zzalp@itu.edu.tr (Z. Zengin Alp), sgunduz@itu.edu.tr (Ş. Gündüz Öğüdücü).

[1] https://www.statista.com/statistics/270229/usage-duration-of-social-networks-by-country/.

https://doi.org/10.1016/j.knosys.2017.11.021
0950-7051/

Aral et al. [16] showed that influential users are influenced less from non-influential users. Thus, they claim that influential users who have influential friends should be addressed to obtain higher information diffusion levels. Hence, recursive calculation of influence depending on followers' influence may help us to identify influencers that are followed by other influencers. Moreover, someone could intuitively say that, most socially influencing person is the mostly-connected person (i.e. who has many followers). However, many researches proved that [17–19] this is not necessarily true. More complex analysis is required for this purpose. Thus, network structure and user specific information are both important features to detect the influence level of users. A recent study [20] on Twitter revealed that 1% of users who serve as influencers control the 25% of the information diffusion.

In an earlier research [1], it has been shown that, people tend to be influential authorities on specific topics such as sports, economy, politics, rather than being global authorities. This result leads us to explore more on topical authorities. Our proposed approach considers both features related to user activities such as **focus rate, activeness, authenticity**, and **speed of getting reaction**, and network features, such as following information. In this paper, a novel representation of user related features is also introduced. These user specific (nodal) features are incorporated into network features and a modified version of PageRank algorithm called Personalized PageRank, is proposed. This proposed algorithm is applied in a distributed manner in order to efficiently analyze topical influence of users.

The outcomes of Personalized PageRank are evaluated with two different approaches. First, by calculating potential spread in our test set, it is empirically evaluated whether the information diffusion is high. The information diffusion is estimated with a measure called *spread score* and compared with baseline methodologies. The spread score is calculated by normalizing the number of retweets for a user over the number of tweets of that user. This measure is simple to calculate and exploits how information spreads from a given user. This empirical evaluation technique is different than the ones used in many researches which calculate precision of user recommendation, or do manual evaluation by using prior knowledge of domain experts. Since recommending users on Twitter and testing the results is not applicable in real life, this evaluation technique usually abides hypothetical. Retweet information is also used in other researches [21] to evaluate performance of influence analysis. However, this research uses retweet information in their experiments as a feature of users which may lead to biased results.

Secondly, a user study has been conducted by asking volunteers if they think our identified influential users are real influencers. This user study is conducted as a survey where sample tweets of influencers are demonstrated to users and asked their opinions.

Contributions of this research are many folds:

- The proposed approach integrates user related features and network features in order to identify topical influencers which yields a better performance than the proposed baseline methods.
- The spread score, along with other state-of-the art evaluation measures, are empirically tested in order to verify the ranking results of spread score.
- A relatively large data-set is collected and used in the experiments compared to many recent works [3,22–25].
- Last but not least, the proposed methodological approach can be easily extended to test further user and network features on different datasets.

Besides these contributions, we also present a technical improvement in terms of computational performance by applying distributed parallel processing methods for graph based algorithms.

The rest of the paper is organized as follows: In Section 2 influence analysis works in the literature will be briefly explained. In Section 3, baseline methodologies that will be used to compare our proposed method will be described. In Section 4 some details about our data collection, preprocessing, topic modeling, and influence analysis methodologies will be given. Afterwards, results of our experiments will be demonstrated in Section 5. Finally, we will briefly summarize our inferences and explain future directions in Section 6.

2. Related work

Influential user identification can be categorized into three groups based on the methods they applied: (1) non-graph based approaches; (2) graph-based approaches; and (3) graph-based approaches with nodal features.

2.1. Non-graph-based approaches

Non-graph-based approaches do not take network information into account such as who is following whom. For instance, Cha et al. [19] recently analyzed the correlation of influence to follower count, retweet count and mention count in Twitter network and identified that follower count is not a key indicator of influence. They defined social influence in a different way by de-emphasizing the role of influencers and determining key factors of influence as interpersonal relationship among ordinary users and the readiness of society to adopt an idea. They demonstrated that retweet and mentions are better indicators of being influencer meaning that influencers are users who increases information diffusion. This definition is also similar to our definition of influencers. Similarly, Pal et al.[22] identified several user metrics such as topic of interest, originality of tweets, and they used Gaussian clustering and ranking algorithms to identify authorities in Twitter.

These approaches do not consider the effect of network topology on influence, specifically user following relationship. Since influence and information diffusion are entwined together, and diffusion of information is highly related to network topology, this information is very important for influence analysis. Yang et al. [26] also showed that user roles in the network is a crucial information for influence analysis.

2.2. Graph-based approaches

Graph-based approaches are able to model diverse types of information obtained from network topology. These methods can also be categorized into two sub-groups such as diffusion models and influence models. Although these fields are two separate areas of research, they can be both used to identify influencers since we define influence as the information diffusion and influencers as the users who contribute most on the diffusion of information. There are some classical information diffusion models such as LT (Linear Threshold) [27] and IC (Independent Cascades) [28] models that try to identify diffusion of information using the network topology. Both models classify nodes in the network as active and inactive depending on the information exposure. These are threshold based models such that when a user becomes active, it activates its followers with a probability. If this probability is above a certain threshold, the follower becomes active too.

Assigning these probabilities and thresholds, which is an NP-Hard problem [29], is another challenge as addressed in many researches [4,30–32]. One of the recent approaches based on LT and IC models proposed by Long et al. [32] is called J-Min-Seed algorithm. This work addresses to minimize the number of seeds and try to activate (influence) at least some predetermined number of users. Goyal et al. [4] also recently proposed an influence