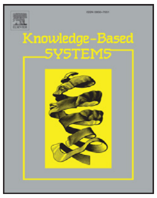
Knowledge-Based Systems 141 (2018) 211–221

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/© 2017 Elsevier B.V. All rights reserved.

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 maximization algorithm based on LT and IC models to overcome the difficulties in finding edge probabilities on Flickr and Flixter data sets. Their approach requires to have information of who shared other people's information which Twitter does not provide us. Twitter only gives the original tweet owner of a retweet even if re-tweeting user has seen the tweet via another person.

Lou et al. [20] proposed an approach to identify influencers that maximizes the information diffusion with an intuition that users that serve as *Structural Holes* are the influencers. *Structural Holes* are defined as the users who connect two or more communities in the network. In another recent work, Liu et al.[25] proposed an influence maximization algorithm ("diffusion-containment model") based on LT model and their model addresses to minimize the spread of opponent's spread while maximizing their own spread.

Another graph based approach for finding influencers in a social network is Google's well-known PageRank algorithm [2]. This algorithm was proposed to rank Web pages for sorting search results by assigning an importance score to each Web page. Kwak et al. [33] used this algorithm to compare the performance of PageRank scores with retweet rate and follower count. They identified that PageRank scores are highly parallel to follower count of users. On the other hand, retweet rate and PageRank scores do not yield to similar influencer sets.

All of these graph-based approaches treat their users identically if they are global influencers (not focus on any topic) and they don't consider any user specific information rather than their place in the network. Although these algorithms identify global authorities, the algorithms can be used on sub-networks that are formed by subset of users who focus on specific topic. Hence, they can be used to identify topical authorities.

2.3. Graph-based approaches with nodal features

This class of algorithms consider the topology of a network as well as user specific features. For instance, Weng et al. [3] proposed TwitterRank algorithm. Their algorithm considers topical information of tweets and topical similarity between each pair of users. They proposed an algorithm similar to PageRank which runs on topic specific networks and they used biased transition between users with respect to topical similarities between them. This algorithm will also be analyzed in detail in Section 4 as a baseline methodology.

Romero et al. [18] argued that, to be an authoritative user, users need to be both popular and active. Their definition of influence of a user involves both the size of the influenced audience, in addition to passivity of those users. They introduced an algorithm that determines influence and passivity of users. Their algorithm Influence-Passivity (IP), is similar to well-known HITS algorithm [34] which has been first proposed to determine influence score of web pages, similar to the PageRank algorithm. Sankar et al. [23] proposed a swarm intelligence algorithm that consider user, message, network and temporal features that mimic honey bees searching for food source. The *reputation* of users are calculated recursively where getting large number of retweets positively effected the outcome.

Riquelme et al. [35] prepared a thorough survey that summarizes researches on influence analysis conducted on Twitter. We can see from the research that, most of the methodologies either lack in measures that are used, or time complexity performance. Moreover, most of the related work are on finding global authorities rather than topical experts. Assuming one person is authoritative in all topics is not usually true as shown in a recent work [1]. Some other methods identify topical authorities by using topical information in posts of users [3,22].

Moreover, Haveliwala et al. [36] proposed "Topic-Sensitive PageRank" where they used PageRank with a topic-biased teleportation vector. They proved that their approach yields better influencer sets than PageRank algorithm.

Recently, Jendoubi et al. proposed two new influence maximization models for Twitter. Their approach introduces a new influence measure based on importance of the user in the network and popularity of her tweets. Their approach uses this measure in an influence maximization algorithm like Credit Distribution algorithm. However, evaluation of their algorithm considered retweet, follow and mention information where their influence measure is also based on the same information. We believe this biases the evaluation. Hence, in this work, we discard retweet rate, which is our evaluation measure is based on, from our experimental model.

3. Baseline and comparative methods

In this section, two state-of-the-art methodologies are explained as baseline methods and two different simple approaches are also explained as comparative methods.

Page-Rank [2] (denoted as PR) and TwitterRank [3] (denoted as TR) algorithms are selected as baseline methodologies to compare the performance of proposed methodology Personalized PageRank (denoted as PPR).

PR is first proposed to rank web pages for Google Search but it is a widely-used methodology to rank nodes in any network. It recursively calculates scores of nodes such that a node A contributes to its friends' scores by its score divided by its out degree. Thus, a high scoring user (influential) contributes more to its friends especially when it has fewer friends. Eq. (1) is the formulation that is used by PageRank algorithm. The first part of the equation is the initialization of scores ($PR^{(0)}$) which is set to 1 for each node. In the second part of the equation, F_v denotes the set of the nodes pointing to v, and N_u denotes outgoing degree of node u. This calculation is computed for each node for each iteration. Page et al. [2] showed that 50 iterations is sufficient for scores to converge in most of the cases.

$$\begin{aligned} PR^{(0)}(v) &= 1, \\ PR^{(i+1)}(v) &= \textstyle\sum_{u \epsilon F_v} PR^{(i)}(u)/N_u \end{aligned} \tag{1}$$

A damping factor is incorporated into this equation to deal with sinks, cycles or disconnected components in the graph. Damping factor is used to model a random surfer who jumps to any other web page rather than following links on current web page. Eq. (2) models the random surfer model for PageRank calculation. d is the damping factor which is usually set to 0.85.

$$\begin{aligned} PR^{(0)}(v) &= 1, \\ PR^{(i+1)}(v) &= (1-d) + d \textstyle\sum_{u \epsilon F_v} PR^{(i)}(u)/N_u \end{aligned} \tag{2}$$

In this research, PR algorithm is applied to topical networks which are sub-networks of the collected data set. These topical networks are formed after calculating topical information of tweets using Latent Dirichlet Allocation (LDA). Nodes in the topical networks are users who posted on related topic and edges show following relationship in the network.

For the second baseline methodology, TR algorithm was implemented as in [3]. TR is different than PR in terms of its transition between nodes. While PR's random surfer jumps to a random node arbitrarily, TR favors users who are similar to each other. This similarity is defined based on the topics that users posted on. Thus, a user-user matrix should be stored for each topic, which makes space complexity quadratic to user count.

Baseline models PR (that runs on topical networks) and TR are both topic specific as the proposed model PPR. Moreover, while PR considers only network properties, TR also considers nodal features such as topical similarity.

We also compared our proposed method with two different simple approaches. First one is selecting influencers randomly and

Table 1
Statistics about data sets.

	#users	#tweets	#links/follows
Training Set	186K	31M	16M
Test Set	181K	7M	15.6M

the second one is by selecting users who have the highest number of followers. Second approach is used by many of the marketers/companies to spread their brand information. It will be used to demonstrate that follower count is not the best way to identify influencers, as already proven by some researches [17–19].

4. Proposed methodology

In this section, we explain our data collection, and preprocessing steps as well as the topic modeling, and influence analysis methodologies.

4.1. Data collection

Prior to data collection, 20 Twitter users that are thought to be focusing on different topics such as politics, sports, TV, religion etc. are manually selected. Afterwards, friends and followers of these users are collected in a breadth first manner until sufficient number of users are obtained. Since Twitter API does not allow us to see information and tweets of protected profiles, they were eliminated. Then, 20 tweets of these users are retrieved and language code of these tweets are checked to identify the users whose tweets are in Turkish most of the time (80%). Users who don't post mostly in Turkish are also eliminated. At this point we had around 180K public user ids who post majorly in Turkish.

Using Twitter Streaming API, tweets related to these users (posted by them or mentioned about them) were collected between November, 4th 2015 and January 12, 2016. Afterwards, the data set was split into training and test sets where tweets before December 30, 2015 are separated as the training set and the rest as the test set. Table 1 demonstrates statistics about the collected data set. As seen from the table, 181K of 186K users also appeared in the test set. 38M tweets were divided to training and test sets as 31M and 7M respectively.

4.1.1. Preprocessing and topic modeling

After data collections, the next steps are the preprocessing and identification of topic(s) of tweets. Tweets, by nature, contain the highest amount of abbreviations, grammatical and orthographical errors [37]. The 140 character limit of Twitter makes people use abbreviations, and shorten words (write "u" instead of "you"). Thus, this type of text data needs thorough investigation and preprocessing. Moreover, since Turkish is an agglutinative language, a root word can get many forms with suffixes which makes Turkish text data harder to process. A recent NLP tool was used that is built by Yildiz et al. [38] using deep learning techniques. The tool was built for Turkish language to be used for morphological analysis. The root words for each token in each Tweet are identified using this tool. This process reduced the number of unique words in the entire data set significantly which improves the performance of topic modeling tools. After stemming, stop-words, punctuation, and mentions that have no meaning for topic modeling were also eliminated from the data set.

After data preparation, stemmed and cleaned tweets were used for topic modeling using Latent Dirichlet Allocation (LDA) tool called MALLET (MAchine Learning for LanguagE Toolkit) [39]. MALLET uses smoothed LDA for topic modeling as Blei et al. proposed in [40]. LDA uses a multinomial word distribution to represent semantically coherent topics. Since LDA is a bag-of-words approach, short text like tweets lowers the topic modeling performance. However, some pooling approaches [41–44] were proposed to enhance LDA performance on short text. In a recent work [44], we showed that pooling tweets for each user and each day gives better performance than other pooling techniques. There are also many works that emphasize usage of hashtags for detecting topics or sentiment analysis [44–46]. It is evidential that users tend to use hashtags to spread the word, and character limit of Twitter makes them very suitable for this job. However, when we started working on Turkish tweets, we realized that hashtag usage was not that common. Hence, topic modeling and preprocessing for influence analysis depends highly on language of tweets and the methodologies should be selected appropriately.

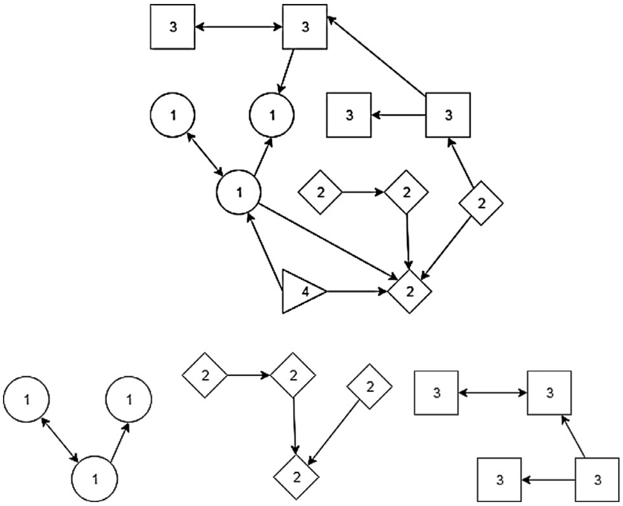


Fig. 1. Topical network construction example.

After processing tweets with LDA, topics were formed as explained in [44]. Three human experts analyzed the output of LDA, which is the word clusters representing each topic. Afterwards, they identified coherent topics which contain words that are semantically similar. Finally, they labeled the coherent topics with appropriate topic titles. After classifying tweets with topics, topical networks were formed. For each user, percentage of topical tweets are calculated and topical labels were assigned to users according to topical contents of their tweets. For each topic, a sub-network is generated where nodes represent the users who post on the specific topic and edges represent the following relationship. A certain threshold is also selected such as if percentage of topical tweets of a user is below this threshold, she would not added to the sub-network [44]. This process reduces the noise if a user has only few number of topical tweets of specific topic.

Fig. 1 demonstrates an example global network and colors of the shapes represent the users who post on specific topics. Fig. 1b–d are sub-networks constructed using topical information from global network. Notice that inter-topical edges are dropped, and users may not belong to any topical network if their tweets do not contain any topical information. Moreover, a user can be in more than one network. This situation occurs when a user posts tweets related to more than one topic.

4.2. User modeling

For topical influence analysis, after topic modeling and construction of topical networks, the next step of PPR is modeling users with proposed user specific features. Several features, that are in the literature or that might be important to be influential on specific topics are identified. All of these features are calculated using simple parameters for each user that are summarized

Table 2
Parameters used for user modeling.

Param.	Explanation
p_u	Tweets of user u
p_u^t	Tweets of user u posted on topic t
$p_{u,\, rt}$	Retweeted tweets of user u
$p_{u,rt}^t$	Retweeted tweets of user u on topic t
rt_u	Retweets of user u (retweeted by u)
rt_u^t	Retweets of user u on topic t
d	Total number of days
d_u^t	Number of days user u posted on topic t
rt_{time}^p	Duration passed for first retweet of post p

in Table 2. All of the user features are calculated using these parameters that are computed on the training set. Proposed user features that will be employed for user modeling are:

1. Focus Rate: This feature calculates the intensity of tweets of a user for a specific topic. We believe that users who focus on fewer topics tend to be more influential than users who post on many topics. Moreover, users with fewer topical interest tend to have higher focus on those topics. The ratio of tweets for each user u on each topic t to the number of her total tweets is calculated as in Eq. (3). A similar feature is previously used in [22] as “Topical Signal”.

$$fr_u^t = \frac{|p_u^t|}{|p_u|} \tag{3}$$

2. Activeness: This feature is used to identify the effect of being active to be influential. Three different activeness features were calculated for each user on each topic as: (1) the number of active days (post anything related to a specific topic) for each user; (2) average number of tweets for each day on a specific topic; (3) the average number of tweets on a specific topic multiplied by the active days. Eq. (4) shows calculation of each *activeness* measure. For the best of our knowledge, activeness feature is being used for the first time on influence analysis .

$$ac1_u^t = \frac{d_u^t}{d} \qquad ac2_u^t = \frac{|p_u^t|}{d} \qquad ac3_u^t = \frac{|p_u^t| \cdot d_u^t}{d} \tag{4}$$

3. Authenticity: This feature indicates how original the users’ tweets are. It is calculated as the number of posts that are not retweets over the total number of tweets per user as in Eq. (5). Authenticity measures whether a user is usually posting her own words or re-tweeting others’. If most of the tweets of a user are originated from herself then the authenticity score of that user is higher. Intuitively, an influencer user is expected to be more authentic. A similar feature is previously used in [22] as “Signal Strength”.

$$au_u^t = \frac{|p_u^t| - |rt_u^t|}{|p_u^t|} \tag{5}$$

4. Speed of getting reaction: This is the average time passing after a person posts a tweet and the tweet gets its first retweet. The idea that is being tested here is that “if a user is getting reaction very fast most the time, she is likely to increase the diffusion of information and becomes an influencer”. For the best of our knowledge, *speed of getting reaction* feature is being used for the first time on influence analysis.

$$sp_u^t = \frac{\sum_{\forall p \in p_{u,rt}} rt_{time}^p}{|p_u^t|} \tag{6}$$

First three parameters ought to be positively correlated to influence while the last one, *speed of getting reaction*, ought to be negatively correlated.

4.3. Influence analysis

After topic and user modeling steps, the next step is the Influence Analysis for *PPR* in order to identify topical authorities in the data set. The nodal/user features, explained in Section 4.2, are incorporated to PR algorithm to achieve high information diffusion levels as a result of identification of highly influential users. Proposed algorithm is called Personalized PageRank (*PPR*) since features that are the outcomes of user modeling phase are used to alter PR algorithm. To see the individual effects of the selected user features different variations of the PPR algorithm are implemented such as PPR_{fr}, PPR_{ac1}, PPR_{ac2}, PPR_{ac3}, PPR_{au}, and finally PPR_{sp} where the subscript stands for the user feature. For example, PPR_{ac1} is the implementation of the *PPR* algorithm for the first activeness measure. PR calculation in Eq. (2) is reformulated where damping factor is used as user features varying for each user and topic. The proposed *PPR* is calculated as in Eq. (7) where w_u^t is specific to experiment (i.e: fr_u^t, $ac1_u^t$, $ac2_u^t$, $ac3_u^t$, au_u^t, and sp_u^t) and varying for each user and topic.

$$\begin{aligned} PPR^{(0)} &= 1, \\ PPR^{(i+1)}(u) &= (1 - w_u^t) + w_u^t \textstyle\sum_{v \in F_u} PPR(v)/N_v \end{aligned} \tag{7}$$

Since measure sp_u^t is negatively correlated to influence, $1 - sp_u^t$ is used in place of w_u^t in Eq. (7) for PPR_{sp} experiments.

PPR algorithm is applied to topical networks as explained in Section 4.1.1. Hence, *PPR* algorithm ran for each user feature and each topical network, which is 36 times for this research (6 measures by 6 topics).

The proposed algorithm is implemented using map and reduce functions of Spark framework. These mappers and reducers are similar to Hadoop Map-Reduce which is a software framework for distributed processing of large data sets. As a basic concept, Mappers get some data and emit key-value pairs from it. Reducers get these key-value pairs and combines the values for each key. This combination can be a simple summation or more complex functions.

Fig. 2 schematically explains the *PPR* calculation with mappers and reducers. Example network in Fig. 2a is used for *PPR* calculation as in Fig. 2b. In Spark, a master node distributes jobs to workers nodes automatically depending on the cluster configuration. For the example case in Fig. 2, map and reduce jobs are distributed to 3 worker nodes. Notice that node sizes of the sample network reflect the weighs (user measure values) of users i.e. *focus rate*. Layers for *PPR* calculation on Spark (Fig. 2b) can be explained as:

- Top layer demonstrates user weights and initialized ranks of the nodes. As explained in Eq. (7), initial ranks are set to *1*. Weights are user features that are obtained from user modelings. For instance, *focus rate* is used for PPR_{fr} experiment.
- Contributions that are received by users are calculated using Mappers in the second layer. Partial contributions are ranks that come from followers and add up to the total rank of the user. For instance, Nodes *D* and *E* contributes to Node *A*’s rank proportional to their friend count.
- Finally, new ranks are calculated with Reducers in the bottom layer. Reduce function for the *PPR* is:

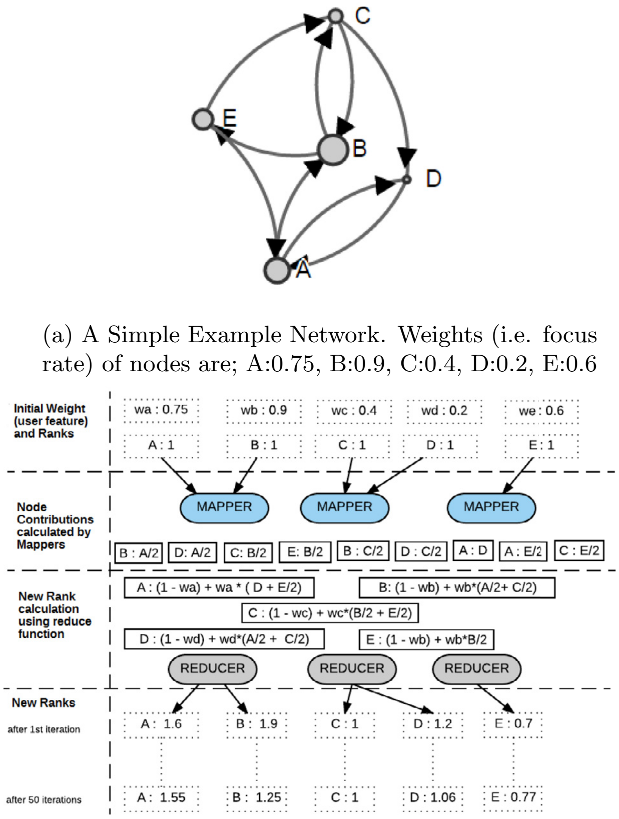
  $$(1 - w_u) + w_u * sumOfContibutions$$

  for each node as in Eq. (7). For instance, *sumOfContributions* for node *A* in the example is:

  $$rank(D) + rank(E)/2$$

  These steps are repeated 50 times as explained in Section 4.

Bottom layer demonstrates the Ranks of nodes after the 1st and 50th iterations. After 50 iterations, Node *A* has the highest rank and Node *E* has the lowest rank.



(a) A Simple Example Network. Weights (i.e. focus rate) of nodes are; A:0.75, B:0.9, C:0.4, D:0.2, E:0.6

(b) Rank calculation of above network using Map-Reduce Framework.

Fig. 2. Map-Reduce framework for PPR algorithm.

Weights (w) that are used in the example are; A:0.75, B:0.9, C:0.4, D:0.2, E:0.6. Notice that, even though Nodes B and D are receiving identical contributions from followers, B has a higher rank. This is because weight of Node B is higher for this specific example. Similarly, Node A has the highest rank even though it's weight is not the highest. This situation is a cause of high contribution of A's followers. Hence, our proposed *PPR* algorithm well balances weight and rank contributions in order find real influencers.

4.4. Complexity analysis

The *PPR* algorithm is implemented as a Scala script that runs Spark where the weight of each node is calculated with *map* and *reduce* functions on worker nodes as in Fig. 2b. In each iteration, jobs are split to worker nodes and the contributions of follower weights are summed up with *reduceByKey* function as in Eq. (7). This made our approach highly scalable and near real time. On the other hand, TwitterRank algorithm needs to hold the pairwise topical similarity between each user pair. Even the influence calculations can be split to worker nodes, the similarity values need to be held by each node. The file size of these pairwise similarities can easily reach to 10GBs with 50K users which is quadratic to user size. In *PPR*, the user specific features like *focus rate* is a single instance per user which makes the space complexity linear to user size. Hence, PPR algorithm is more efficient than TR algorithm in terms of space complexity. *PPR* is one of the first algorithms that has low computation and space requirements and achieves certain accuracy for influencer identification such as in a recent work by Lu et al. [24].

All of the algorithms ran on Spark environment on a 4-Core processor, 8GB RAM machine and HDFS is used to the store data such as networks and user features. With this much computational power, *PPR* calculations for 180K nodes and 16M edges took around 3 min for 50 iterations. This experiment is conducted to test the performance of global network that is collected for this research. Running time is proportionally faster for sub-networks.

4.5. Evaluation

To measure the performance of *PPR* algorithm, an evaluation measure is calculated based on retweet rate of influencers to identify their potential of spreading the information. Since retweet information is not used in our experiments, it is a valid measure for testing purposes. For each topical network, *PPR* algorithm outputs (top 25 influencers) is used. Normalized retweet rate for each user is calculated and the results are summed to calculate the overall information diffusion estimate of top 25 influencers. Fig. 3 shows the spread totals of PPR_{au} experiment on different topics for different number of influencer users. It can be seen from the Figure that there is usually a threshold for information diffusion for 25 users. Thus, we used top 25 users as influencers for all experiments. Eq. (8) demonstrates how potential information diffusion is calculated.

$$spread(t) = \sum_{u\epsilon\{inf_t\}} \frac{|p^t_{u,tr}|}{|p^t_u|} \sum_{p\epsilon\, p^t_{u,tr}} |retweets_p| \tag{8}$$

In the equation, $p^t_{u,tr}$ and p^t_u parameters are used as explained in Table 2. For each retweeted tweet p of $p^t_{u,tr}$, we summed up the number of retweets and normalized this sum with the retweet rate of the user. This formula gives us the potential spread of posts of influential users. All of the variables in Eq. (8) are calculated using test data set.

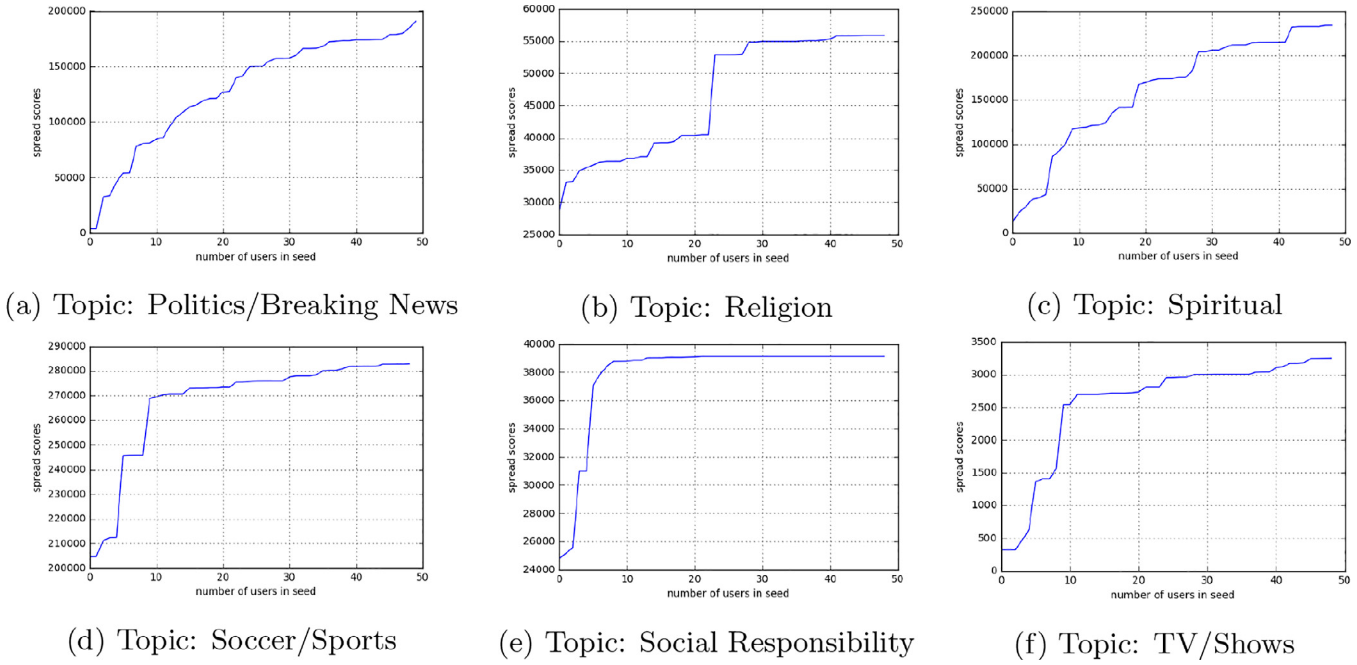
Higher spread scores indicate that the identified influencer users have more potential to spread the information which is the desired outcome. As an example, let's say two users A and B have same number of retweets which is 30, and same number of tweets which is 10. If A has 3 retweeted tweets and B has only one retweeted tweet, their spread score would be 9 and 3 for users A and B respectively. Hence, spread score also considers both potential of getting retweet and retweet count. Retweet rate has not been used in such a way previously for the best of our knowledge.

Secondly, we conducted a human survey by showing tweets of influencers and asked the volunteers if they think the owner of the tweets are “highly influencer”, “influencer”, “not an influencer”, or volunteers have another option to say “no idea”. 3 topics and 4 users from each experiment that don't overlap are selected to evaluate. In order not to make the survey unreasonably long, we could not have been able to evaluate all of the influencers and all of the topics.

5. Experimental results

This section gives experimental results for the methodologies explained in the previous section, such as topic modeling, user modeling and influence analysis. Results of two different evaluation methodologies are also demonstrated, (1) potential information diffusion that the identified influencers induced is calculated by Eq. (8) and (2) conducted a survey where volunteer opinions are asked if the identified influencers are real influencers by showing some random tweets of them.

For topic modeling, a previously proposed approach was used to improve LDA performance [44] that combines tweets by author and date as a document. 20 word clusters were obtained where human experts identified 6 of them as coherent and labeled these



(a) Topic: Politics/Breaking News

(b) Topic: Religion

(c) Topic: Spiritual

(d) Topic: Soccer/Sports

(e) Topic: Social Responsibility

(f) Topic: TV/Shows

Fig. 3. Spread scores for different number of seed users of experiment PPR_{au}.

Table 3
6 Topics generated by topic modeling using LDA.

Politics/News	Spiritual	Religion	Social responsibility	Soccer/Sports	TV/TV Shows
akp[a]	love	allah	need for	galatasaray[b]	watch
hdp[a]	beautiful	nmuhammed	blood	soccergame	show
chp[a]	#loveis	Rabbi	rh	ngoal	scene
syria	happy	pray	#blood	team	trailer
#brkngnews	hearth	hz	#urgent	beşiktaş[b]	#kirazmevsimi
martyr	peace	muslim	#thrombocyte	fenerbahçe[b]	#kiralkaşk[c]
paris	poem	religion	#collectingbooks	soccer	#güneşinkızları[c]
terror	lover	belief	#3decemberdisabledday	league	#poyrazkarayel[c]
france	#whatislove	heaven	awareness	fan	#pantenealtınkelebek[c]
#lastminute	#loveisactually	islam	#helptheoneinneed	champion	#muhteşemygüzyılkösem[c]

[a] A Turkish political party name
[b] A Turkish soccer team
[c] A Turkish TV Show

clusters as “Politics/Breaking News”, “Spiritual”, “Religion”, “Social Responsibility”, “Soccer/Sports”, and “TV/TV Shows”. Since the data collection period coincided with the political election in Turkey, “Politics” and “Breaking News” topical words were highly overlapping, thus a single topic is formed for both topics. Table 3 demonstrates selected topics and top words in the topics.[2]

After topic modeling is completed, topical networks are constructed by forming graphs where nodes represent users who post on the specific topic and edges represent the following relationships where edge direction is from follower to friend. Users whose topical tweet ratio is less than 25% is discarded from topical networks in order to avoid users who rarely post on the topic as a noise cancellation process. After topical network construction, a user might end up in zero, one or many topical networks (See example Fig. 1).

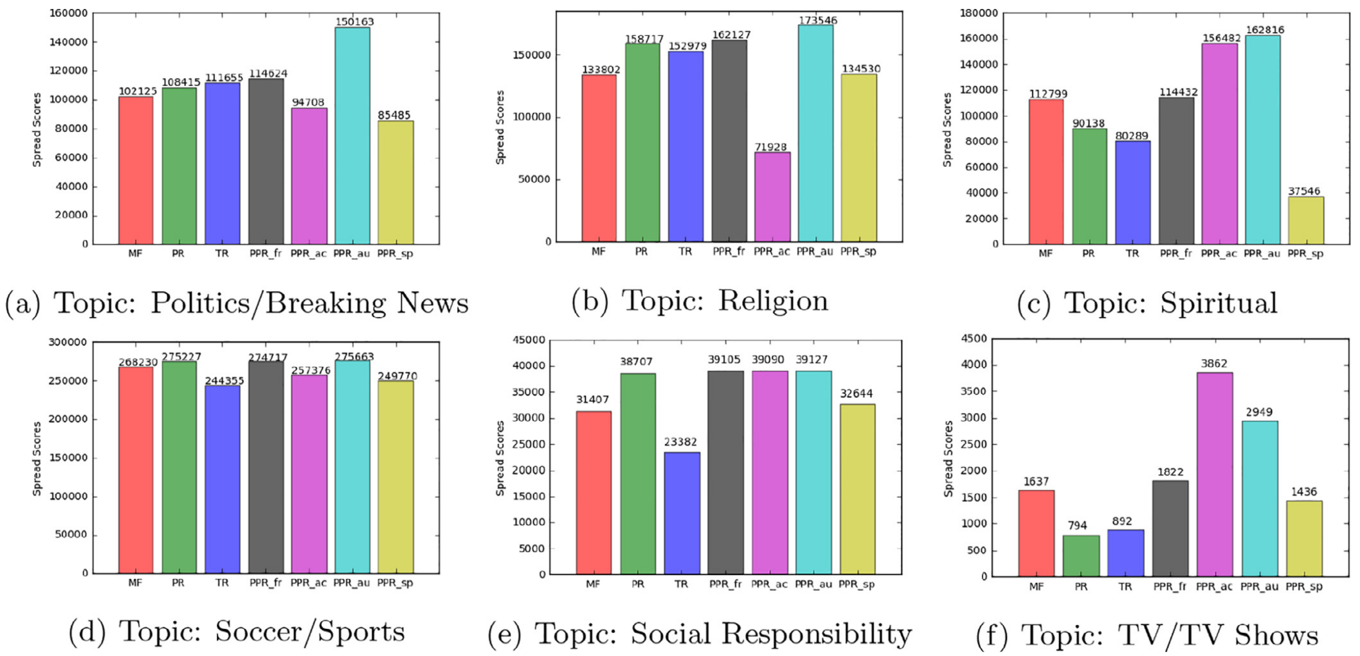
Table 4 demonstrates statistical information of each topical network.[3] As shown in the table, although entire network is very large and more dense than sub-networks, average number of tweets and retweets is not respectively large. This shows that our sub-networks are filtering out users that are not active and not passing information a lot. This makes sub-networks more valuable and computationally feasible. Table 4 also demonstrates that “TV/Shows” topic has larger average tweet rate than the other topics; however, its retweet rate is smaller. Thus, people are less willing to pass information on this topic. Table 4 also shows that retweet rate is very large for “Soccer/Sports” and “Politics/Breaking News” topics. Hence, people want to pass these type of information to others more than the other topics.

Table 4
Statistics about networks.

	Nodes	Links	Avg tweets	Avg retweets
Entire Network	185,898	16,5M	168	106
Politics/brkgNews	33,462	2,5M	394	456
Spiritual	57,877	2,3M	309	320
Religion	15,092	286K	312	350
Social Resp.	1292	6K	383	270
Soccer/Sports	14,335	266K	276	560
TV/shows	3516	42K	445	208

For influential user analysis, 10 experiments were conducted including baseline methodologies, random influencer selection, most followed user (denoted as MF) selection, and *PPR* with 6 differ-

[2] Topic words are translated into English if possible in order to be displayed in the paper.
[3] Data set is available upon request with limitations of Twitter privacy policy.



(a) Topic: Politics/Breaking News

(b) Topic: Religion

(c) Topic: Spiritual

(d) Topic: Soccer/Sports

(e) Topic: Social Responsibility

(f) Topic: TV/TV Shows

Fig. 4. Top 25 users' total spread scores comparisons.

ent user specific features. Each algorithm ran for each topic and top influencers were identified. Top 25 of them are selected as the influencer set. Afterwards, potential spread is calculated for each topic and experiment. Total potential spread of different "activeness" calculations (*ac1, ac2*, and *ac3*) yielded very close information diffusion. For simplicity, we will demonstrate results of only PPR_{ac1} related to activeness. Fig. 4 gives potential spread scores of influencers for each experiment and topic. It can be seen from the figure that at least one *PPR* spread score is higher than the baseline methods in all of the cases. Especially, PPR_{au} is the best performing experiment in all cases except one, "TV/TV Shows". PPR_{ac} performs better for "TV/TV Shows" topic. This shows that active users tend to be more influential for this specific topic. Hence, we can conclude that although incorporating user specific user features improved the influence analysis performance for all cases, different features can affect differently depending on the topic. On the other hand, "authenticity" measure improves performance of influential user identification for all topics. Another interesting result is that, speed of getting reaction experiments (PPR_{sp}) performs worse than baseline algorithms for almost all topics. One would think that if a user is getting retweeted very fast, she would be an influencer, which is not the case for our experiments. Another conclusion that can be derived from results that, mostly followed users are not essentially the most influential people. In all cases *PPR* experiments outperform the *MF* experiment. We also experimented on selecting users randomly as influencers. However, results are not demonstrated since the spread scores for this algorithm were close to zero for all cases.

For "Politics/Breaking News" topic, PPR_{au} experiment outperforms all other experiments with large difference. For topic "Religion", authenticity feature still affects positively more than others but other experiments also give very close scores including baseline methods. Only "Activeness" feature does not affect influence positively for this topic. We can conclude that, activeness is not very important for topic "Religion". For "Spiritual" topic, PPR_{sp} experiment under-performs baseline methods, where all other *PPR* experiments outperform the baseline methods. Thus, "speed of getting reaction" is not a key feature for "Spiritual" topic. However, "activeness" and "authenticity" is very important. For "Soccer/Sports" topic, all experiments perform almost equally where PPR_{au} is slightly better performing experiment. This is similar for "social Responsibility" topic too. In summary, at least one *PPR* experiment outperforms baseline methods in all topics, which indicates that using user specific features positively affect identifying topical influential users to increase information diffusion. However, user feature effects can be different for each topic. This shows that topic specific features need to be identified by thorough analysis for each topic. On the other hand, "Authenticity" of users positively affect being influential in all cases, which could be identified as the best performing feature for this data set.

As can be seen in Fig. 4, the spread score ranges are different for each topic. This is due to the different average retweet rate of sub-networks. Since "Soccer/Sport" sub-network has the highest retweet rate (See Table 1), its spread scores are higher than the other topics. TV/TV Shows sub-network has the lowest retweet rate, hence its spread scores are lower. Thus, retweet rate of the network has direct effect on the spread results.

Evaluation of the identified influencers is a challenging task since it is a very subjective matter and many different approaches have been used in the literature. Some of the mostly used evaluation approaches are; manual evaluation of identified users [19,34], conducting human surveys [22,36], calculating node activation based on IC model [4,29–32], calculating retweet or mention rate of identified influencers [19,47], calculating in-degree of the influencers in the social network [3], calculating PageRank scores of influencer users [3,36].

Two of these evaluation methodologies are used as separate experiments to compare with proposed methodology, such as, PageRank (as PR experiment), in-degree (as MF experiment). Thus, these techniques have not been used for evaluation in order not to bias the results. Since "spread" is based on retweet rate, it is already used for evaluation. Next, we will compare "activation count" of state-of-the-art IC model and "spread" score. After that, results of human survey will also be demonstrated.

To prove "spread" is a suitable measure to use, an empirical comparison is made between "spread" and "activation count" of IC methodology [28]. IC model is a widely used and studied methodology, where many researches [4,29–32] ground their evaluation

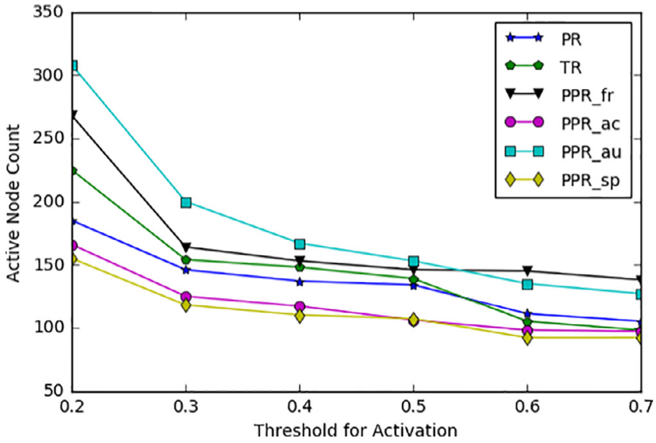


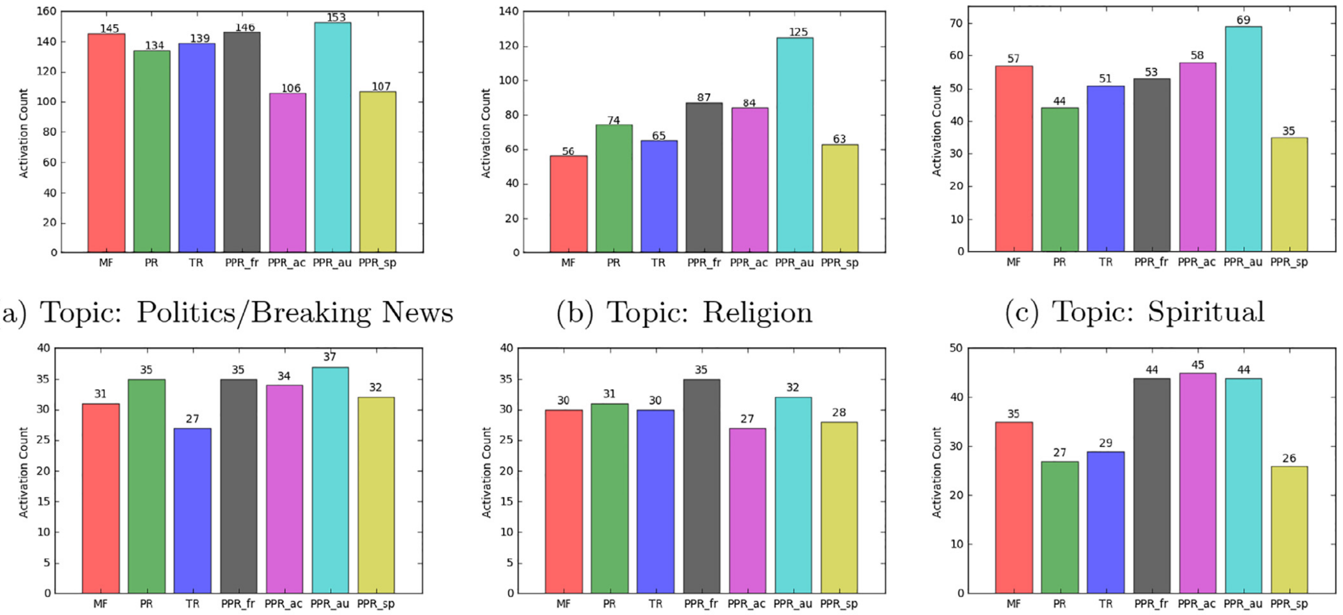
Fig. 5. Active node count based on IC model with different thresholds for Politics/BreakingNews topic.

methodology on. In IC model, influence is defined as how a single node influences it's neighbors. Each edge between nodes contains influence probability between nodes. In the IC model, first step is to initialize some nodes as active and calculate total activation at each step with these edge probabilities and an activation threshold. Iteration stops when no activation occurs in previous iteration. Assigning these edge probabilities and the threshold is another challenge, as discussed in Section 4.

In this evaluation technique, edge probabilities are calculated as retweet rate between each pair of users, i.e. if a user *A* retweets 5 of user *B*'s 10 topical tweets, the edge probability for *B* to activate *A* is set to 0.5. If the activation threshold of the network is below this value, when *B* is activated, it also activates *A*. Fig. 5 demonstrates Active node count for Politics/Breaking news network for different activation threshold values. When two empirical evaluation methodologies are compared ("Spread Score" and "Activation Count") different experiment performances are very similar (See Figs. 4a and 5). Fig. 4a gives spread score results for Politics/BreakingNews topic (same topic with Fig. 5). When we compare the two figures, PPR_{au} is the best performing and PPR_{sp} is the worst performing algorithm. Order of other algorithms are also similar in two figures.

For deeper comparison of "spread" and "active node count", Fig. 6 shows performance of activation on different experiments on all topics. For simplicity, a single threshold value is selected, which is 0.5. Fig. 6 also depicts that performance of experiments with spread scores in Fig. 3 are parallel to "activation count". This figure also shows that "spread" and "activation" calculations give similar results. Since spread is easier to calculate, it can be used for influence evaluation on social networks.

Findings of our experiments were also evaluated over a survey where several tweets of influencers were demonstrated to volunteers and they have been asked if they think the owners of tweets are influential in the specific topic. Specifically, volunteers asked if they think the owner of the tweets are "Highly Influential", "Influential", "Not Influential" or volunteers can select "Not Sure". Since time to complete the survey should be reasonable, we limited the number of topics to 3 ("Social Responsibility", "Soccer/Sports", "Politics/Breaking News"), influential user count from each topic to 4, and number of tweets to be shown by each influential to 6. We selected top 4 influencers from each set, while avoiding users who occur in most of the sets. After results were collected, influence scores for each influential user were calculated with weights 3, 2, 1, and 0 for each choice "Highly Influential", "Influential", "Not Sure", and "Not Influential" respectively. Finally, the average of user influence for each topic and experiment is calculated. 47 volunteers completed the survey. Fig. 7 demonstrates the influence scores that are calculated by the outcome of user survey. Although the best performing experiments are different from our findings in Fig. 4, still at least one *PPR* based experiment outperforms the baseline methodologies. An interesting outcome of the survey was that authenticity is not as important for user decisions as the outcome of calculated "Spread" scores. On the contrary, PPR_{sp} performs best for "Social Responsibility" topic while it is the worst experiment with "spread" evaluation. PPR_{ac} is the best performing experiment for "Politics/BreakingNews" and "Soccer/Sports" topics with survey results. Even so, it is impor-



(a) Topic: Politics/Breaking News (b) Topic: Religion (c) Topic: Spiritual

(d) Topic: Soccer/Sports (e) Topic: Social Responsibility (f) Topic: TV/TV Shows

Fig. 6. Top 25 users' total activation count comparisons of each experiment for each topic.

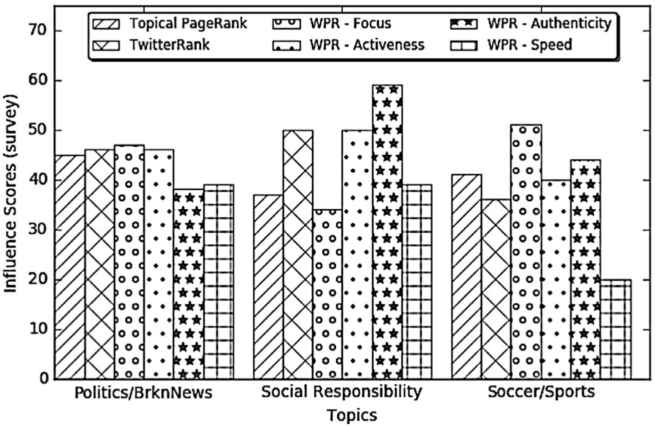


Fig. 7. Influence scores obtained from survey.

tant to state that, showing limited number of tweets of limited number of influencer users might have biased the results of the survey.

The results of the experiments conducted in this paper can be summarized as follows:

- Splitting the global network to topical networks is an effective approach for determining influential users. As we demonstrated in a recent work [1], running PageRank on topical networks rather than global network improves spread score significantly.
- It is useful to incorporate user specific features to network features. However, the performance of various user specific features depends on the network topology. Authenticity feature yields satisfactory results in most of the topical sub-networks in terms of spread scores.
- Using "spread score" for evaluating social influence is effective since it is simple to calculate and gives similar results compared to the state-of-the art measures.

6. Conclusion and future work

In this paper, we propose a novel method, *PPR* algorithm, that incorporates both the user specific features and topological network features to identify topical social influencers in a network. The proposed method is based on the idea that users may have different expertise and/or interests in various topics which leads them to be topic-specific influencers. The results obtained from the proposed algorithm and the baseline methods are evaluated by human participants using a survey. The survey results also verify the influencer users identified by *PPR* algorithm.

Several nodal features; *focus rate, activeness, authenticity*, and *speed of getting reaction*, are incorporated into network features like friend/following information. The proposed method is also implemented in a distributed manner to increase the speed of calculations. Experimental results on a large data-set collected from Twitter show that using both user specific features and network features is more effective in identifying topical influencers.

In this paper, we also conducted different experiments in order to compare commonly used evaluation measures for evaluating social influencers. The empirical results illustrate that a simple evaluation measure, spread score, is also effective to verify influencer users.

As a future work, all user features can be combined to a derived feature in order to capture the concept of topical influence. Another future work can be using the *PPR* algorithm for other purposes than detecting the topical influencers. For instance, we are going to investigate if *PPR* algorithm can be used for identification of "troll" accounts. We believe that some demographic information and authenticity feature can be used with *PPR* to detect trolling in social media.

References

[1] Z.Z. Alp, Ş.G. Öğüdücü, Influential user detection on twitter: analyzing effect of focus rate, in: Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on, IEEE, 2016, pp. 1321–1328.
[2] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web., Technical Report, 1999-66, 1999.
[3] J. Weng, E.-P. Lim, J. Jiang, Q. He, Twitterrank: Finding topic-sensitive influential twitterers, in: Proceedings of the Third ACM International Conference on Web Search and Data Mining, in: WSDM '10, ACM, New York, NY, USA, 2010, pp. 261–270, doi:10.1145/1718487.1718520.
[4] A. Goyal, F. Bonchi, L.V.S. Lakshmanan, A data-based approach to social influence maximization, Proc. VLDB Endowment 5 (1) (2011) 73–84, doi:10.14778/2047485.2047492.
[5] S. Jendoubi, A. Martin, L. Liétard, H.B. Hadji, B.B. Yaghlane, Two evidential data based models for influence maximization in twitter, Knowl. Based Syst. 121 (2017) 58–70.
[6] F. Bravo-Marquez, E. Frank, B. Pfahringer, Building a twitter opinion lexicon from automatically-annotated tweets, Know.-Based Syst. 108 (C) (2016) 65–78, doi:10.1016/j.knosys.2016.05.018.
[7] S. Poria, E. Cambria, A. Gelbukh, Aspect extraction for opinion mining with a deep convolutional neural network, Knowl. Based Syst 108 (2016) 42–49. New Avenues in Knowledge Bases for Natural Language Processing. doi: 10.1016/j.knosys.2016.06.009.
[8] B. Altnel, M.C. Ganiz, A new hybrid semi-supervised algorithm for text classification with class-based semantics, Knowl. Based Syst. 108 (2016) 50–64. New Avenues in Knowledge Bases for Natural Language Processing. doi: 10.1016/j.knosys.2016.06.021.
[9] A. Muhammad, N. Wiratunga, R. Lothian, Contextual sentiment analysis for social media genres, Knowl. Based Syst. 108 (2016) 92–101. New Avenues in Knowledge Bases for Natural Language Processing. doi: 10.1016/j.knosys.2016.05.032.
[10] M. Deutsch, H.B. Gerard, A study of normative and informational social influences upon individual judgment., J. Abnormal Social Psychol. 51 (3) (1955) 629–636.
[11] R. Lippitt, N. Polansky, S. Rosen, The dynamics of power; a field study of social influence in groups of children., Human Relat. 5 (1952) 37–64.
[12] A.R. Cohen, Some implications of self-esteem for social influence., 1959.
[13] E. Katz, P.F. Lazarsfeld, Personal influence, the part played by people in the flow of mass communications, Can. J. Econ. Polit. Sci. 6 (21) (1957).
[14] A. Goyal, Social Influence And Its Applications, The school of the thesis, The University of British Columbia, Vancouver, 2013 Ph.D. thesis.
[15] S. Aral, D. Walker, Tie strength, embeddedness, and social influence: a large-scale networked experiment, Manage. Sci. 60 (6) (2014) 1352–1370, doi:10.1287/mnsc.2014.1936.
[16] S. Aral, D. Walker, Identifying influential and susceptible members of social networks, Science 337 (6092) (2012) 337–341, doi:10.1126/science.1215842.
[17] L. Gallos, S. Havlin, M. Kitsak, F. Liljeros, H. Makse, L. Muchnik, H. Stanley, Identification of influential spreaders in complex networks, Nat. Phys. 6 (11) (2010) 888–893.
[18] D.M. Romero, W. Galuba, S. Asur, B.A. Huberman, Influence and passivity in social media, in: Proceedings of the 20th International Conference Companion on World Wide Web, in: WWW '11, ACM, New York, NY, USA, 2011, pp. 113–114.
[19] M. Cha, H. Haddadi, F. Benevenuto, P.K. Gummadi, Measuring user influence in twitter: the million follower fallacy., ICWSM 10 (10–17) (2010) 30.
[20] T. Lou, J. Tang, Mining structural hole spanners through information diffusion in social networks, in: Proceedings of the 22Nd International Conference on World Wide Web, in: WWW '13, ACM, New York, NY, USA, 2013, pp. 825–836, doi:10.1145/2488388.2488461.
[21] L. Liu, J. Tang, J. Han, S. Yang, Learning influence from heterogeneous social networks, Data Min. Knowl. Discov. 25 (3) (2012) 511–544, doi:10.1007/s10618-012-0252-3.
[22] A. Pal, S. Counts, Identifying topical authorities in microblogs, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, in: WSDM '11, ACM, New York, NY, USA, 2011, pp. 45–54, doi:10.1145/1935826.1935843.
[23] C.P. Sankar, S. Asharaf, K.S. Kumar, Learning from bees: an approach for influence maximization on viral campaigns, PloS one 11 (12) (2016) e0168125.
[24] W.-X. Lu, C. Zhou, J. Wu, Big social network influence maximization via recursively estimating influence spread, Knowl. Based Syst. 113 (2016) 143–154.
[25] W. Liu, K. Yue, H. Wu, J. Li, D. Liu, D. Tang, Containment of competitive influence spread in social networks, Knowl. Based Syst. 109 (2016) 266–275.
[26] Y. Yang, J. Tang, C.W.-k. Leung, Y. Sun, Q. Chen, J. Li, Q. Yang, Rain: social role-aware information diffusion., in: AAAI, 2015, pp. 367–373.
[27] M. Granovetter, Threshold models of collective behavior, Am. J. Sociol. 83 (6) (1978) 1420–1443, doi:10.1086/226707.
[28] J. Goldenberg, B. Libai, E. Muller, Talk of the network: a complex systems look at the underlying process of word-of-mouth, Mark. Lett. 12 (3) (2001) 211–223, doi:10.1023/A:1011122126881.

[29] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the spread of influence through a social network, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '03, ACM, New York, NY, USA, 2003, pp. 137–146.

[30] A. Guille, H. Hacid, A predictive model for the temporal dynamics of information diffusion in online social networks, in: Proceedings of the 21st International Conference on World Wide Web, in: WWW '12 Companion, ACM, New York, NY, USA, 2012, pp. 1145–1152.

[31] K. Saito, K. Ohara, Y. Yamagishi, M. Kimura, H. Motoda, Foundations of Intelligent Systems: 19th International Symposium, ISMIS 2011, Warsaw, Poland, June 28–30, 2011. Proceedings, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 153–162.

[32] C. Long, R.C.-W. Wong, Viral marketing for dedicated customers, Inf. Syst. 46 (2014) 1–23, doi:10.1016/j.is.2014.05.003.

[33] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a social network or a news media? in: WWW '10: Proceedings of the 19th international conference on World wide web, ACM, New York, NY, USA, 2010, pp. 591–600, doi:10.1145/1772690.1772751.

[34] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM (JACM) 46 (5) (1999) 604–632.

[35] F. Riquelme, P. González-Cantergiani, Measuring user influence on twitter: a survey, Inf. Process. Manag. 52 (5) (2016) 949–975, doi:10.1016/j.ipm.2016.04.003.

[36] T.H. Haveliwala, Topic-sensitive pagerank, in: Proceedings of the 11th International Conference on World Wide Web, in: WWW '02, ACM, New York, NY, USA, 2002, pp. 517–526, doi:10.1145/511446.511513.

[37] G. Petz, M. Karpowicz, H. Fürschuß, A. Auinger, V. Stříteskỳ, A. Holzinger, Opinion mining on the web 2.0–characteristics of user generated content and their impacts, in: Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, Springer, 2013, pp. 35–46.

[38] E. Yildiz, C. Tirkaz, H.B. Sahin, M.T. Eren, A morphology-aware network for morphological disambiguation, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, in: AAAI'16, AAAI Press, 2016, pp. 2863–2869.

[39] A.K. McCallum, Mallet: a machine learning for language toolkit, 2002. http://mallet.cs.umass.edu.

[40] M. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, Mach. Learn. Res. 3 (2003) 993–1022.

[41] L. Hong, B.D. Davison, Empirical study of topic modeling in twitter, in: Proceedings of the First Workshop on Social Media Analytics, in: SOMA '10, ACM, New York, NY, USA, 2010, pp. 80–88, doi:10.1145/1964858.1964870.

[42] W.X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, X. Li, Comparing twitter and traditional media using topic models, in: Proceedings of the 33rd European Conference on Advances in Information Retrieval, in: ECIR'11, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 338–349.

[43] R. Mehrotra, S. Sanner, W. Buntine, L. Xie, Improving lda topic models for microblogs via tweet pooling and automatic labeling, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, in: SIGIR '13, ACM, New York, NY, USA, 2013, pp. 889–892, doi:10.1145/2484028.2484166.

[44] Z.Z. Alp, S.G. Ögüdücü, Extracting topical information of tweets using hashtags, in: 14th IEEE International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, December 9–11, 2015, 2015, pp. 644–648, doi:10.1109/ICMLA.2015.73.

[45] S. Rill, D. Reinel, J. Scheidt, R.V. Zicari, Politwi: early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, Knowl. Based Syst. 69 (2014) 24–33.

[46] E. Sulis, D.I.H. Farías, P. Rosso, V. Patti, G. Ruffo, Figurative messages and affect in twitter: differences between #irony, #sarcasm and #not, Knowl. Based Syst. 108 (2016) 132–143.

[47] I. Anger, C. Kittl, Measuring influence on twitter, in: Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, in: i-KNOW '11, ACM, New York, NY, USA, 2011, pp. 31:1–31:4, doi:10.1145/2024288.2024326.