Cluster Computing (2019) 22:S4035–S4045
https://doi.org/10.1007/s10586-018-2616-y

Identification and classification of best spreader in the domain of interest over the social networks

A. N. Arularasan[1] · A. Suresh[2] · Koteeswaran Seerangan[1]

Received: 6 March 2018 / Revised: 15 March 2018 / Accepted: 18 March 2018 / Published online: 27 March 2018

Abstract
The emerging social networks promptly create greater opportunities for fast-developing viral marketing. The online social networks (OSNs) play an essential role in the information diffusion among the social users within the community. The social network being large-scale, it leads to the inconvenience in identifying the influential spreaders in a specific domain, as every social user receives the information from different sources through multiple connections over the network. Although, analyzing the complex social network is indispensable to determine the influence spreaders with the knowledge of understanding the dynamics of information evolution. The existing solutions of the influential measurement techniques lack in neglecting the redundant links and quantifying the temporal information among the social users while estimating the diffusion importance of a social user. Moreover, these techniques fail in analyzing the structural relationships in the domain. To overcome these obstacles, this paper presents a de-duplicated k-shell influence estimation (DKIE) model in the social network by classifying the influential spreaders based on the domain of interest using k-shell decomposition and N-gram similarity. The DKIE model incorporates two major phases such as generic influential spreader identification and domain-specific influential spreader identification. The first phase measures the diffusion importance of each active social user based on the structural relationships of the social network using k-shell decomposition method. It separates the core-like groups and true core and identifies the best spreaders regardless of the redundant links. The second phase exploits the topic of the discussion of the best spreaders and consequently, measures the topic-wise influence to categorize the domain-specific best spreaders using N-gram similarity measurement. The experimental results illustrate the effectiveness of DKIE approach.

Keywords Viral marketing · Social network · Core-like group · k-shell · Redundant link · Influence · Classification · N-gram

1 Introduction

In recent years, the social networks and microblogging sites have become the most prominent and popular communication medium among the Internet users. In day-to-day life, the social users spread more than billions of messages within the different social networks such as Twitter, Facebook, Tumblr, and Google Plus in which the users have the ability to post and share their opinions with a specific community freely. The advent of online social networks (OSNs) [1] substantially increases the amount of real-time information stream, including daily chats politics, sports, entertainment and celebrity gossips, which covers almost the entire global information. Viral marketing requires the information hub influencing a large number of

✉ A. Suresh
prisu6esh@yahoo.com

A. N. Arularasan
arularasan@live.com

Koteeswaran Seerangan
s.koteeswaran@gmail.com

[1] Department of Computer Science and Engineering, School of Computing, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu 600062, India

[2] Department of Computer Science and Engineering, Nehru Institute of Engineering and Technology, T.M.Palayam, Coimbatore, Tamil Nadu 641105, India

spreaders in the social network. This spreading phenomenon increases the exposure range of a particular product with low cost. Thus, there is an emerging need for analyzing the massive information exchange while considering the significance of the information diffused in the OSN [2]. With this intent, tracing the information propagation level along with the cause of each spread is a fundamental challenge in the social network. To identify the cause of the information diffusion, the analysis methods need to find the origin of the contagion and the path of the propagation in OSN [3].

Instead of exploiting the structural relationships of the social network for information hub identification, considering the real information exchange among the social users is beneficial to viral marketing. Even though, identifying the influential users in a particular domain, regardless of the textual information is a complicated task while ensuring the spreading efficiency in the global social network. Hence, the categorization of the globally exchanged information in the social network is necessary to identify the domain-specific information hub [4] since every social user discusses the information over multiple domains. In recent years, most of the recent works related to the information hub identification exploit the k-shell method to discover the propagation level of each user based on the explicit social relationship of the network [5, 6]. However, these k-shell decomposition methods lack in eliminating the duplicate links during the influence measurement in the social network. Moreover, the complete view of the social network requires the analysis of information exchange among the users to determine the influence of each user in a specific domain [7]. In contrast with the existing k-shell based influential user identification approaches, this paper targets to find the nature of the relationship among the social users based on the topic of the shared information.

The main contributions of the de-duplicated k-shell influence estimation (DKIE) approach are as follows.

- The main objective of the DKIE methodology is to classify the domain-specific information hubs or the best spreaders with the knowledge of utilizing both the structural relationships and the topic of the discussion within the social network.
- The DKIE approach exploits the structural social relationships to determine the core-like groups having the redundant links and true core, and consequently, measure the diffusion importance of each social user using the k-shell decomposition method along with the entropy function.
- With the aim of globally identifying the best spreaders, the DKIE approach takes into account of the best spreader in the core-like group when the spreading efficiency is higher than the true core rather than ignoring the entire core-like group.
- Instead of identifying the structural relationship based best spreaders, the DKIE approach analyzes the tweet information of each best spreader using N-gram similarity method to understand the tendency of each user in a specific domain.
- Finally, the DKIE approach classifies the domain-specific best spreaders by exploiting the N-gram similarity measurement with the knowledge of the corpus and WordNet ontology.
- The experimental results demonstrate that the DKIE approach precisely classifies the domain-specific best spreaders in the social network under the multiple domains of interests.

1.1 Problem statement

Most of the earlier research works in viral marketing have been focusing on classifying the domain-specific information hubs in the social network to facilitate the recommendation. To classify the social users, the system needs the analysis of the textual information shared among the users within the network. In a large social network, analyzing the massive information collection of the numerous social users is an arduous task, which develops the time and cost complexity. Hence, the information hub classification system requires the completion of information hub identification as a prerequisite process in large social networks. Several existing information hub identification methods employ the structural relationship of the social network rather than analyzing the tweet information to reduce the burden of the identification process. Even though, these methods lack in accurately identifying the influential users due to the global network structure. To tackle this constraint, the recent information hub identification techniques, exploits the k-shell method to improve the accuracy of the influential users. However, there is a possibility of the redundant links occurrence in the true core and ignoring of the efficient core-like group due to the lack of analyzing the spreading efficiency in the true core and the elimination of the entire core-like group respectively. Moreover, neglecting the non-active users is essential while measuring the diffusion importance of each social user in the network. Thus, to overcome these issues, the DKIE approach focuses on the identification of the best spreaders based on the structural relationships and consequently, on the determination of domain-specific best spreaders based on the topic of the discussion.

1.2 Paper organization

The structure of the paper is organized as follows: Sect. 2 deliberates the existing research works related to the problem adopted. Section 3 describes in detail about the proposed methodology which correspondingly explains the component involved in this proposed methodology to attain the optimal solution for the problem. Section 4 highlights the experimental evaluation and proves that the proposed methodology obtains a better solution by comparing with the existing method. Section 5 presents the conclusion which summarizes the work done in this paper.

2 Related works

In the social network, the earlier information hub identification works employ the following methods such as degree distribution, node clustering, degree centrality, communities, degree correlation, and k-shell. Most of the recent researches focus on the k-shell or k-shell decomposition method of the social network to identify the core of the network in terms of influential spreaders in the fast developing field of viral marketing [8]. Even though influential spreader has a large number of neighbors, it negatively impacts the spreading efficiency. Mixed degree decomposition (MDD) method [9] takes into the account of the number of removed and existed links during the decomposition process to distinguish the spreading influence of each user within the same group using k-shell value. An approach [10] measures the influence of social user after differentiating the core-like group and true core based on link diversity of shells, which includes the redundant links of a shell while identifying the super-spreaders. Later, an advanced k-shell method [11] further improves the accuracy of the influential spreader identification by exploiting the spreading dynamics and removing the redundant links. Also, there are several influential spreader identification types of research [12, 13] focused on the k-shell strategy in the social network. An improved and efficient ranking method [14] assigned the influence rank to the social users based on the shortest path distance between the center node and it is a neighbor node which has the highest k-shell value in the social network. A large k-shell decomposition framework [15] presents a linear-time randomized algorithm to decompose a large graph into the most extreme k-shell based on the random edge contraction technique.

Several conventional research works [16, 17] target on the temporal quantification in the social network since, the real social networks comprise the enormous data about user activities, which varies according to the dynamic nature. The personalization of the user's profile is mandatory since the users have interest on various topics. There are only a few researchers have presented that the identification of the influential spreaders is not only depended on the number of followers but also with the similar intention of the users. An approach [18] determines the influential spreaders based on the wide popularity of the social users and the influence on multiple topics over the social network. It does not automatically mean that these influential spreaders have the ability to influence their activities and their interests. An approach [19] considers the user connections in the Facebook social network and analyzes the relationships among the social users. Whereas, the work [20] exploits the messages in the communication network of Yahoo, and the work [21] takes into the account of the digital bibliography and library project (DBLP) to evaluate the collaborations among the scientific authors. The supervised learning technique [22] classifies the short messages shared on a social network by exploiting labeled training set and identifying the information paths over the data to determine the domain of the information. Domain-based user influence estimation [23] employs the classification method to categorize the information hubs under various domains based on the interest of each hub. An approach [24] presents a novel model to measure the influence of the social user by analyzing the multiple paths of the information. It analyzes the dynamics of the information contagion according to the change of topic of the discussion. TopicRank [25] identifies the topic-sensitive influential spreaders by employing the latent Dirichlet allocation (LDA) model which learns the topic of each propagation message over the social network. It lacks in considering the redundant links in the social network, which degrades the performance of influence measurement. However, the existing works have analyzed only the user connection and their relationship they lack in precisely identifying the information hub for a specific domain. Moreover, accurately classifying the massive information based influential spreaders regardless of the training set is still in its infancy stage. In [26, 27] has discussed the four human-in-the-loop simulations conducted at NASA Ames Research Center illustrating the framework and key aspects. In [28, 29] the introspective reports on cooperative behaviour seems to be efficient for concept map participants. In [30, 31] it shows that the actions are supported by the infusion systems highlighting the different sources of user availability.

3 An overview of the proposed methodology

With the extreme popularity and rapid growth of the Online Social Networks, the demand for viral marketing has significantly increased to disseminate the product advertisements among the social users effectively. To recommend the appropriate products to the social users through influential users, the personalization or classification of information hub is essential in OSNs. Accordingly, to determine the domain-specific best spreaders, effective best spreader identification is necessary. Many current information hub identification and classification methods rely on the structural and tweets information to identify the domain-specific best spreaders on the social network. The structural information based information hub identification method selects the spreaders with minimum spreading efficiency due to the existence of the core-like group. In real-world social network systems, the core-like group contains a very few outbound links instead of having the high diffusion value. Hence, determining the redundant links is crucial while measuring the diffusion importance across the nodes to form the true core group excluding the redundant links. The DKIE approach incorporates two phases such as identifying the best spreaders based on the diffusion measurement in k-shell and classifying the best spreaders using N-gram method to classify the information hubs in the OSNs precisely.

Identifying the best spreaders based on the diffusion measurement in k-shell The DKIE approach identifies the best spreaders by finding the active users and forming the true core with abundant outbound links using k-shell decomposition method. Initially, it discovers the potential active users using timestamp information and performs de-duplication in two steps. The two steps of de-duplication involve in separating the core-like group and measuring the diffusion importance regardless of redundant links, which leads to the formation of the true core group. The DKIE approach applies the k-shell decomposition method with entropy function on all the active users pertaining in the true core to measure the propagation level of each spreader in the social network in which entropy function estimates the uncertainty of the diffusion to determine the spreading capability of each spreader.

Classifying the best spreaders using N-gram method After identifying the best spreaders, the DKIE approach delves the tweets of each best spreader regarding the domain of discussion to classify the domain-specific best spreaders. It classifies the best spreaders by exploiting the user-generated content of tweets, corpus comprising a collection of information about multiple domains, and WordNet ontology. The DKIE approach applies the N-gram similarity method between the tweets and the corpus information to categorize the domain-specific best spreaders accurately. It facilitates the system to understand the dynamics of domain-wise information contagion over the social network (Fig. 1).

3.1 Identifying the best spreaders based on the diffusion measurement in k-shell

The DKIE approach aims at classifying the best spreaders based on the topic of the discussion in the OSNs. Before personalizing the best spreaders, the system needs to perform the best spreader identification process as the prerequisite process. Hence, the DKIE approach estimates the influence among the social users by applying the k-shell decomposition method for the explicit structural relationships to find the best spreaders. Initially, it discovers the active users in the social network based on the social activity logs and timestamp information of each user with other users to ease the best spreader identification. Then, it determines and removes the redundant links in the social network structure, and creates the true core from the core-like group using k-shell and entropy function based influence. Eventually, it identifies the best spreaders excluding the uncertainty of redundant spreading capability.

3.1.1 Determining the active users and manipulating the de-duplication using diffusion importance

The DKIE approach identifies the active social users by exploring the corresponding user's social activity logs and timestamp data. The social users involve in two different kinds of activity logs such as (1) creating new content in terms of uploading music and images, writing messages and blogs, and (2) consuming others' content in terms of downloading music and images, and reading messages and blogs. Also, timestamp information is used to identify whether the user is either active or inactive at a particular time. In a large explicit social network, the DKIE approach considers the graph structure of the social network based on the social relationships between the social users to determine the adoption of information propagation across the network. The social network comprises a core-like group containing many users with low spreading efficiency. The core-like group is likely to lead the inaccurate identification of the best spreader due to the repeated densely connected nodes. Hence, identifying the core-like group is necessary before identifying the best spreaders. From Eq. (1), the probability value ($P(C_G)$) determines that nodes and its neighboring nodes either from the core-like group or the true core group. In the core-like group, the tendency of each user depends on the influence disseminated within the community of membership.

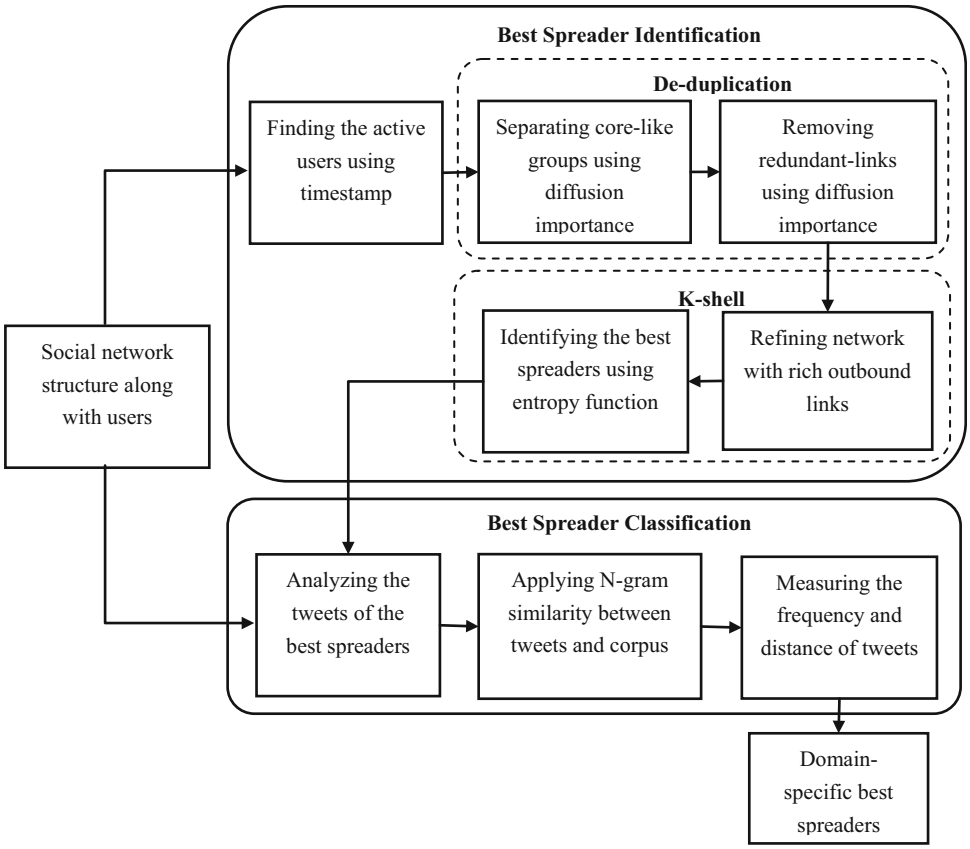

Fig. 1 Classification of information hubs using k-shell method

$$P(C_G) = \frac{\left|\left\{\{N_i\} \cap \left\{\bigcup_{j=0}^{k} N_j\right\}\right\}\right|}{|N_i|}, \quad (1)$$

where, 'i' and 'j' represents the nodes in the social network, and 'k' refers the number of neighborhood nodes to the ith node. 'N_i' and 'N_j' denotes the number of neighboring nodes of node 'i' and 'j' respectively. Equation (1) facilitates the process of core-like group identification based on the structural feature of the explicit social relationship. Accordingly, the DKIE approach separates the core-like group using the threshold value of the probability score. Even though the core-like group has minimum spreading efficiency, neglecting the core-like group structure from the best spreader identification process is ineffective, since the core-like groups may comprise large number neighborhood nodes than the true core group on the social network. Hence, considering the maximum spreading influence of one of the nodes from the core-like group is crucial during the best spreader identification. It leverages the system to increase the spreading efficiency and avert the redundant spreading process.

Instead of separating the core-like group from the best spreader identification process, exploiting the core-like groups based on the spreading efficiency creates a greater impact in identifying the best spreader, since the numerous interconnected social networks are likely to comprise a few redundant links while comparing the number of links in the core-like group. To determine the highly influenced social user of the core-like group, the DKIE approach considers both the structural feature of explicit social relationship and correlation between the number of outbound links and spreading efficiency of each social user bounded in the core-like group. It eases the system to accurately determine the best spreader rather than ignoring the complete structure of the core-like group. Hence, according to Eq. (2), the DKIE approach measures the diffusion importance of each node in the core-like group regardless of the redundant links.

$$\begin{aligned} &\text{Spread\_Eff}(C_G(U_i)) \\ &= \frac{\alpha * \left[\log\left(N_i^L\right)\right] + \beta * \left[\log\left(\left(\frac{N_i^M}{N_i^L}\right) * D(C_G)\right)\right]}{\log_2 U_N}, \end{aligned} \quad (2)$$

where, N_i^L denotes the number of outbound links, i.e., neighbors of an ith node of the core-like group and N_i^M refers the number of outbound links which are not a neighborhood of its one of the neighbors. $D(C_G)$ represents the total number of nodes in the core-like group or density of the core-like group. α and β are the random variables in which $\beta > \alpha$ and $\alpha + \beta = 1$. U_N represents the number of users in the social network. By measuring the diffusion

importance of each node based on the outbound links, the DKIE approach filters out the redundant links across the network. Also, from the perspective of spreading efficiency, the DKIE approach accurately determines the truecoreness of each node and identifies the true core group in the real-world social networks.

3.1.2 Refining network structure and finding the best spreaders using the k-shell method

The DKIE approach employs the k-shell method to refine the network structure into the true core group without non-active users and consequently, identifies the best spreaders. The best spreader is either from the core-like group or true core group based on the diffusion importance or spreading efficiency. To identify the best spreaders, the DKIE approach exploits both the local influence and global diversity of each node in the social network. Similar to the Eq. (2), the DKIE approach formulates the Eq. (3) along with the centrality score, which is used to measure the local influence of each user ($I_L(U_i)$) based on its neighbors. In Eq. (3), 'n' and C(n) denotes the number of direct neighbors of ith node and centrality of nth neighbors in the network structure respectively. N_i represents the direct neighbors of the ith user. α, β, and γ are the random variables in which $\alpha + \beta + \gamma = 1$ and $\beta > \alpha > \gamma$. The DKIE approach exploits the degree centrality measurement in terms of the total number of direct and indirect neighbors [32].

$$I_L(U_i) = \frac{\alpha * \left[\log\left(N_i^L\right)\right] + \beta * \left[\log\left(\left(\frac{N_i^M}{N_i^L}\right) * (U_N)\right)\right] + \gamma * \log\left[\sum_{n \in N_i} C(n)\right]}{\log_2 U_N}. \tag{3}$$

Moreover, the global diversity relies on the overall social network structure to analyze the global characteristics of the social users. The DKIE approach applies the k-shell decomposition with Shannon's entropy to obtain the global node information in a complex social network. From the Eq. (4), entropy function indicates two cases such as (i) a node has the capability to establish the connection with all layers when the entropy (H) is high, (ii) all the nodes have the connections within the same layer when the entropy value is '0'.

$$H_i\left(N_i^K\right) = -\sum_{j=1}^{K_S^M}\left(P_i\left(N_i^j\right) * \log_2 P_i\left(N_i^j\right)\right), \tag{4}$$

where $N_i^K = \{1, 2, \ldots, K_S^M\}$ represents k-shell values of the neighbors of ith user or node. Equation (5) calculates the probability value of jth layer of neighbors ($P_i(N_j)$) in which '$|N_i^j|$' indicates the number of neighbors of the ith user appearing in the jth layer.

$$P_i\left(N_i^j\right) = \frac{\left|N_i^j\right|}{\sum_{j=1}^{K_S=M} N_i^j}. \tag{5}$$

The DKIE approach exploits the normalized score of the global diversity value to determine the overall influence of the social user. Equation (6) refers the global diversity of each user ($\hat{H}_i(N_i^K)$) in the k-shell structure based on the entropy value. The maximum value of the entropy ensures that the neighbors of each node significantly are more diverse in the k-shell structure.

$$I_G(U_i) = \hat{H}_i\left(N_i^K\right) = \frac{H_i\left(N_i^K\right)}{\log_2 K_S^M}. \tag{6}$$

Finally, the DKIE approach utilizes both the local ($I_L(U_i)$)and global value ($I_G(U_i)$) of each node to determine the exact tendency of the corresponding user in the complex social network.

$$\text{Overall influence}(U_i) = \begin{cases} I_L(U_i) * I_G(U_i), & if\ U_i \neq C_G(U_i) \\ \text{Spread\_Eff}(C_G(U_i)), & f\ U_i = C_G(U_i) \end{cases}. \tag{7}$$

Using Eq. (7), the DKIE approach sorts the best spreaders based on the maximum value of overall influence in descending order. It retains the number of best spreaders in a specific OSN regardless of the domain. To support a viral marketing, discovering the domain-wise best spreaders in the social network is necessary.

3.2 Classifying the best spreaders using N-gram method

Instead of estimating the influence of users based on the explicit structural relationship in the social network, the DKIE approach assesses the influence of the social users in a specific domain with the help of N-gram similarity-based classification. The classification of influential users facilitates in analyzing the best spreaders under the specific domain. To perform the N-gram classification method, the system requires the tweets of the social users. The N-gram technique also requires the domains and the training set. The DKIE approach calculates all the feasible n-grams of each keyword (w) in a tweet (T) instead of assigning a classic weighted vector to each tweet, as the tweets mostly comprise abbreviations, small textual errors, and minimal term variations. Hence, the DKIE approach utilizes the N-gram based classification system in natural language processing application to deal with the textual information with abbreviations and errors effectively. To ease the viral marketing, topic categorization is an essential task in text

document analysis based on the pre-determined set of categories. The DKIE approach applies the Dice coefficient based N-gram similarity method [33] between the tweets and the information stored in the corpus to identify the intention of the social users in a particular domain. It analyzes the frequency of tweets of each best spreader in the corpus and measures the distance of the tweets with the topic using the WordNet ontology. Thus, N-gram similarity method precisely categorizes the domain-wise tweets of all the social users (U_S) by utilizing the knowledge of corpus and WordNet ontology.

The DKIE approach focuses on estimating the influence of social users based on the interaction of each user among others by exploiting the topic based information model on tweets (T) and retweets (RT). Each social user (U_i) sends numerous tweets (T_s) involving multiple domains to their neighbors (N_i) in the social network. The DKIE approach not only exploits the tweet information of each social user, but also focuses on the retweets to determine the domain-specific influence of the social user. Initially, it measures the retweeting capability (RT_C) of each user (U_i) in a particular domain (D) using the Eq. (8).

$$RT_C(U_i, D) = \sum_{N_i^{RT} \in RT(U_i,D)} \left[\frac{1}{\omega * \left(\frac{1}{A\left(N_i^{RT}\right)} \right) + (1-\omega) * \left(\frac{1}{RT\left(N_i^{RT}\right)} \right)} \right], \tag{8}$$

where, $|RT(N_i^{RT})|$ represents the number of retweets of retweeting neighbors of ith user in a particular domain, $A(N_i^{RT})$ refers the authority of retweeting neighbor ith user i.e. $A(N_i^{RT}) = |T_{S,D}(N_i^{RT})|$ denotes the number of tweets of the neighbor of ith user in a specific domain (D).In Eq. (9), $\omega < (1-\omega)$ in which 'ω' is a random parameter, $0 < \omega < 1$. $RT(U_i, D)$ indicates the retweeter of the ith user in Dth domain. Finally, the DKIE approach identifies the domain-specific best spreaders by estimating the tweets based influence using Eq. (9).

$$\text{Influence}(U_i, D) = \frac{\log\left|T_{i,D}\right| * \sum_{T_S \in T(i,D)} \left[\sum_{U_X \in RT(T_S)} \left(\frac{|RT(U_X)|}{|N_X|} \right) * RT_C(U_X) \right]}{|T_i|}, \tag{9}$$

where, $|T_{i,\ D}|$ represents the total number of tweets of the ith user in Dth domain. T_s denotes each tweet of the ith user and U_x refers the retweeters of tweet 'T_s'. $|N_x|$ is the total number of direct neighbors of user U_x and $|RT(U_x)|$ is the number of retweeters of user U_x. '$|T_i|$' represents the total number of tweets of an ith user. At the end of the process, the DKIE approach sorts the best spreaders in the descending order based on the estimation of a degree of influence of each user on each topic, which expresses the domain of interest and propagation level through posting activity.

4 Experimental evaluation

This section evaluates the performance of the DKIE approach while comparing with the existing TopicRank approach [25]. It conducts an experiment on the real-world dataset to demonstrate the performance improvement in terms of the domain-specific influence spread maximization over the existing algorithm.

4.1 Experimental setup

The experimental evaluation implements the DKIE approach using Java platform. The DKIE approach employs the social network of Twitter dataset incorporating structural relationships and the tweet information to identify the domain-specific influential spreaders. It performs the experiments on Linux Ubuntu 12.04 LTS 64-bit machine, 2.9 GHz Intel CPU, and 32 GB memory. It exploits text mining analysis tools to process the tweet information gathered from the dataset. Moreover, it exploits the text corpus and WordNet ontology to categorize the topic-wise tweets.

4.1.1 Dataset

The evaluation of the DKIE approach exploits a popular microblogging site of the Twitter dataset. The dataset contains 2.5 million of users who exchange 15 million of tweets across multiple domains. The proposed evaluation takes into the account of 2333 Twitter users, including football players, Olympic athletes, UK parliament members, Irish politicians, and the Rugby Union players. Also, it comprises the tweets with a specific tweet ID regarding keywords along with the retweeting neighbors and the number of retweets. The proposed evaluation employs the corpus with the five unique domains comprising the domain related keywords. The five domains are Football, Olympics, Politics-UK, Politics-IE, and Rugby.

4.1.2 Evaluation metrics

Precision is the ratio between the number of recovered influential users those that are relevant, influential users and the total number of recovered influential users in a specific domain.

Recall is the ratio between the number of recovered influential users those that are relevant, influential users

and the total number of relevant influential users in a specific domain.

F-measure or F-score is the discrepancy and balance between the precision and recall.

$$\text{F - measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Domain-specific retweet influence (DSRI) It is the ratio between the number of domain-specific retweets of the recovered influential spreaders and the total number of the retweets of the relevant spreaders on a specific topic.

4.2 Experimental results

The experimental results illustrate the significant performance improvement of the DKIE approach than over the existing approach while varying the real-time validation scenarios of the number of users, the number of tweets, and the number of domain categories.

4.2.1 Number of users versus precision

Figure 2 illustrates the precision of both the DKIE and the TopicRank approaches while varying the number of social users from 0.5 million to 2.5 million and the Number of Core-Like Groups (NCLG) from 5 to 15. The precision value decreases while increasing the number of users on the social network. The DKIE approach gradually decreases the precision value by 0.94% when the number of users is increasing from 0.5 million to 2.5 million and NCLG is 15. However, the Topic Rank approach suddenly decreases the precision value by 2.59% at the same scenario. Since, the DKIE approach identifies the best spreaders with the consideration of the CLG in the social network, which ensures that the best spreaders have higher influence than others over the social network. When increasing the NCLG from 5 to 15, the performance of both the DKIE and TopicRank decrease due to the burden of processing of the CLG. At the point of 2.5 million of the users, the DKIE approach attains 3.63% of precision higher than the TopicRank approach when NCLG is 10, as the CLG separation and the diffusion measurement of the DKIE approach maintain the precision even when increasing the number of users.

4.2.2 Number of users versus recall

Figure 3 shows the recall performance variation of both the DKIE and the TopicRank approaches for the variation of the number of users and the NCLG. The effectiveness of the domain-specific information hub classification system not only depends on the number of users, but it also depends on the increasing number of CLG on the social network. The DKIE approach maintains the recall value when increasing the number of users, but the TopicRank approach only maintains the recall until reaching a certain number of users after that, it suddenly decreases due to the lack of the CLG consideration during influence measurement. Moreover, the DKIE approach employs the k-shell decomposition method along with the entropy function, which globally measures the influence without the uncertainty of the spreading efficiency. The Topic Rank approach measures the influence having the redundant links, which degrades the diffusion importance value of each social user. At the point of 1 million users when NCLG is 10, the performance of the Topic Rank decreases by 2.31% more than the DKIE approach, but, at the point of 2 million users, the recall of Topic Rank approach decreases much more by 3.1% than the DKIE approach. Since measuring the influence regardless of the CLG and k-shell is likely to provide the benefits to only a few number of users, which is inaccurate when the number of social users is high.

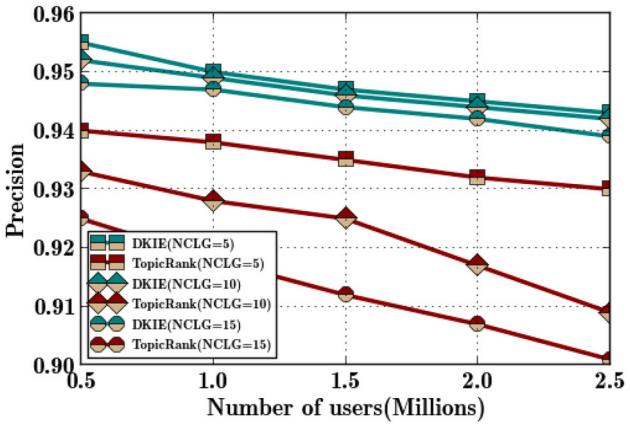

Fig. 2 Number of users versus precision

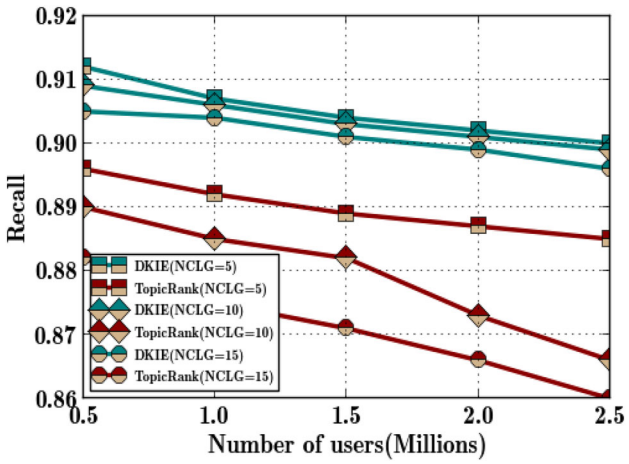

Fig. 3 Number of users versus recall

4.2.3 Number of tweets versus F-measure

Figure 4 depicts the F-measure performance variation while varying the number of tweets from 3 to 15 million and the number of domains (ND) is in the range of 5, 10, and 15 over the Twitter social network. If tweet information comprises additional ND, the domain-specific best spreader classification is a much more complex process, which is likely to degrade the performance of the classification system. The F-measure value decreases when increasing the number of tweets and the ND in the tweets. When ND is 10, the DKIE approach gradually decreases the F-measure value by 1.07%. However, the TopicRank approach decreases by 1.18% while varying the number of tweets from the 3 million to 15 million. Even though, ND = 10 of the TopicRank approach provides better performance than the DKIE approach of ND = 15 until reaching the number of tweets at 6 million, the DKIE approach provides appreciable result than TopicRank approach after crossing 6 million tweets. The TopicRank approach misclassifies the best spreaders when increasing the ND and tweets due to the absence of the retweets influence based domain-specific best spreaders classification.

4.2.4 Number of followers versus domain-specific retweet influence

The comparative results of the DKIE and the TopicRank approaches are shown in the Fig. 5 while varying the number of followers from 0.4 million to 2 million in the social network. It also relies on the k-shell distance that is in the range of 10, 20, and 30. The impact of the number of followers and the k-shell distance on the domain-specific retweet influence (DSRI) potentially reveals the performance of the domain-wise best spreader identification

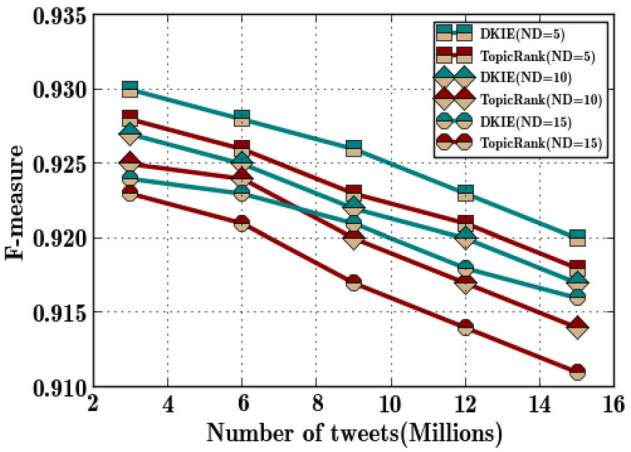

Fig. 4 Number of tweets versus F-measure

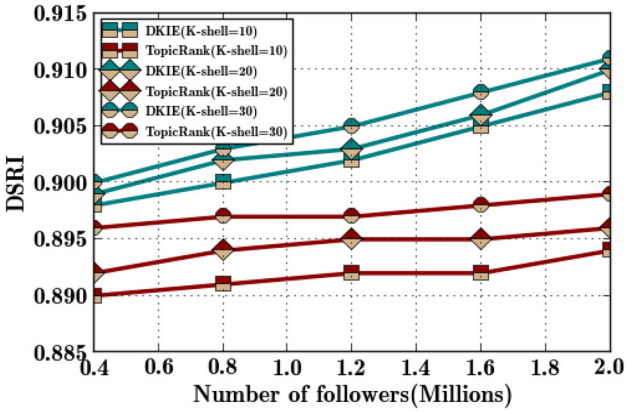

Fig. 5 Number of followers versus DSRI

system. If a social user has few followers in the maximum of k-shell distance, that social user has the high social influence than the number of followers in the minimum k-shell distance. The DKIE approach rapidly escalates the DSRI value by 1.22% when increasing the followers from 0.4 to 2 million and k-shell distance is 30. It focuses on both the tweet and the retweet importance while measuring the influence score throughout the k-shell structure. Whereas, at the same scenario, the TopicRank approach increases the DSRI value only by 0.33%, which measures the influence of every social user regardless of the follower's influence. Moreover, the DKIE approach exploits the corpus and the WordNet ontology to accurately categorize the domain-wise tweets of the best spreaders rather than measuring the learning based topic-wise diffusion probability of the tweets.

5 Conclusion

We have presented an information hub classification model DKIE, which overcomes the obstacles of the domain-specific best spreaders classification in the large-scale social network. The proposed DKIE model employs both the structural relationship and the tweets to precisely identify and classify the best spreaders according to the intention of the users. The significant performance improvement is attained by exploiting the k-shell decomposition and the N-gram classification method with the assistance of the corpus and the WordNet ontology. The k-shell decomposition method globally analyzes the diffusion importance of each social user over the social network to identify the generic best spreaders. Consequently, the N-gram classification method explores the tweets of the best spreaders to determine the domain-specific best spreaders. The experimental results reveal that the DKIE system yields the domain-specific best spreaders with higher influence than the TopicRank system while ensuring

91% of resale value when testing it on the real-world Twitter dataset.

References

1. Bonchi, F., Castillo, C., Gionis, A., Jaimes, A.: Social network analysis and mining for business applications. ACM Trans. Intell. Syst. Technol. **2**(3), 22 (2011)
2. Bakshy, E., Rosenn, I., Marlow, C., Adamic, L.: The role of social networks in information diffusion. In: ACM Proceedings of the 21st International Conference on World Wide Web, pp. 519–528 (2012)
3. Chen, D., Lü, L., Shang, M.S., Zhang, Y., Zhou, T.: Identifying influential nodes in complex networks. Physica A **391**(4), 1777–1787 (2012)
4. Barbieri, N., Bonchi, F., Manco, G.: Topic-aware social influence propagation models. Knowl. Inf. Syst. **37**(3), 555–584 (2013)
5. Rabade, R., Mishra, N., Sharma, S.: Survey of influential user identification techniques in online social networks. In: Recent Advances in Intelligent Informatics, pp. 359–370. Springer, New York (2014)
6. Hou, B., Yao, Y., Liao, D.: Identifying all-around nodes for spreading dynamics in complex networks. Physica A **391**(15), 4012–4017 (2012)
7. Mangal, N., Niyogi, R., Milani, A.: Analysis of users' interest based on tweets. In: 16th International Conference on Computational Science and Its Applications, pp. 12–23. Springer, New York (2016)
8. Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., Makse, H.A.: Identification of influential spreaders in complex networks. Nat. Phys. **6**(11), 888–893 (2010)
9. Zeng, A., Zhang, C.J.: Ranking spreaders by decomposing complex networks. Phys. Lett. A **377**(14), 1031–1035 (2013)
10. Liu, Y., Tang, M., Zhou, T., Do, Y.: Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition. Sci. Rep. **5**, 9602 (2015)
11. Liu, Y., Tang, M., Zhou, T., Do, Y.: Improving the accuracy of the k-shell method by removing redundant links: from a perspective of spreading dynamics. Sci. Rep. **5**, 13172 (2015)
12. Huang, X., Cheng, H., Qin, L., Tian, W., Yu, J.X.: Querying k-truss community in large and dynamic graphs. In: ACM Proceedings of the SIGMOD International Conference on Management of Data, pp. 1311–1322 (2014)
13. Kim, H., Yoneki, E.: Influential neighbours selection for information diffusion in online social networks. In: IEEE 21st International Conference on Computer Communications and Networks (ICCCN), pp. 1–7 (2012)
14. Lin, J.H., Guo, Q., Dong, W.Z., Tang, L.Y., Liu, J.G.: Identifying the node spreading influence with largest k-core values. Phys. Lett. A **378**(45), 3279–3284 (2014)
15. Akiba, T., Iwata, Y., Yoshida, Y.: Linear-time enumeration of maximal k-edge-connected subgraphs in large networks by random contraction. In: ACM Proceedings of the 22nd International Conference on Information and Knowledge Management, pp. 909–918 (2013)
16. Cheng, J., Ke, Y., Chu, S., Özsu, M.T.: Efficient core decomposition in massive networks. In: IEEE 27th International Conference on Data Engineering, pp. 51–62 (2011)
17. Michalski, R., Bródka, P., Kazienko, P., Juszczyszyn, K.; Quantifying social network dynamics. In: IEEE Fourth International Conference on Computational Aspects of Social Networks (CASoN), pp. 69–74 (2012)
18. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in Twitter: the million follower fallacy. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Vol. 10, No. 30, pp. 10–17 (2011)
19. Sun, E., Rosenn, I., Marlow, C., Lento, T.M.: Gesundheit! modeling contagion through Facebook news feed. In: ICWSM (2009)
20. Aral, S., Muchnik, L., Sundararajan, A.: Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proc. Natl. Acad. Sci. **106**(51), 21544–21549 (2009)
21. Di Caro, L., Cataldi, M., Schifanella, C.: The d-index: discovering dependences among scientific collaborators from their bibliographic data records. Scientometrics **93**(3), 583–607 (2012)
22. Zhang, S., Jin, X., Shen, D., Cao, B., Ding, X., Zhang, X.: Short text classification by detecting information path. In: ACM Proceedings of the 22nd International Conference on Information and Knowledge Management, pp. 727–732 (2013)
23. Cataldi, M., Mittal, N., Aufaure, M.A.: Estimating domain-based user influence in social networks. In: ACM Proceedings of the 28th Annual Symposium on Applied Computing, pp. 1957–1962 (2013)
24. Cataldi, M., Aufaure, M.A.: The 10 million follower fallacy: audience size does not prove domain-influence on Twitter. Knowl. Inf. Syst. **44**(3), 559–580 (2015)
25. Zhou, D., Han, W., Wang, Y.: Identifying topic-sensitive influential spreaders in social networks. Int. J. Hybrid Inf. Technol. **8**(2), 409–422 (2015)
26. Lachter, J., Brandt, S.L., Battiste, V., et al.: Enhanced ground support: lessons from work on reduced crew operations. Cogn. Technol. Work **19**, 279 (2017). https://doi.org/10.1007/s10111-017-0422-6
27. Kannan, N., Sivasubramanian, S., Kaliappan, M., Vimal, S., Suresh, A.: Predictive big data analytic on demonetization data using support vector machine. Cluster Comput. (2018). https://doi.org/10.1007/s10586-018-2384-8
28. Nonose, K., Okukubo, A., Yoda, Y., et al.: Support for creating introspective reports detailing behaviors with concept maps. Cogn. Technol. Work **18**, 71 (2016). https://doi.org/10.1007/s10111-015-0347-x
29. Suresh, A., Varatharajan, R.: Competent resource provisioning and distribution techniques for cloud computing environment. Cluster Comput. (2017). https://doi.org/10.1007/s10586-017-1293-6
30. Iacovides, I., Blandford, A., Cox, A., et al.: How external and internal resources influence user action: the case of infusion devices. Cogn. Technol. Work **18**, 793 (2016). https://doi.org/10.1007/s10111-016-0392-0
31. Chinnasamy, A., Sivakumar, B., Selvakumari, P., Suresh, A.: Minimum connected dominating set based RSU allocation for smart Cloud vehicles in VANET. Cluster Comput. (2018). https://doi.org/10.1007/s10586-018-1760-8
32. Gao, S., Ma, J., Chen, Z., Wang, G., Xing, C.: Ranking the spreading ability of nodes in complex networks based on local structure. Physica A **403**, 130–147 (2014)
33. Kondrak, G.: N-gram similarity and distance. In: Springer Proceedings of International Symposium on String Processing and Information Retrieval, pp. 115–126 (2005)

A. N. Arularasan is a Research Scholar in the Department of Computer Science and Engineering, School of Computing at Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India. He received his B.E. and M.Tech. in Computer Science and Engineering. He is a life member of ISTE and IAENG. His current research interests include Data mining & knowledge discovery and Social Network analysis.

A. Suresh currently working as the Professor & Head, Department of the Computer Science and Engineering in Nehru Institute of Engineering & Technology, Coimbatore, Tamil Nadu, India. He has been nearly two decades of experience in teaching and his areas of specializations are Data Mining, Artificial Intelligence, Image Processing, Multimedia and System Software. He has published 45 papers in International journals. He has published more than 40 papers in National and International Conferences. He has served as a reviewer for Springer, Elsevier, and Inderscience journals. He is a member of IEEE, ISTE, IACSIT, IAENG, MCSTA, MCSI, and Global Member of Internet Society (ISOC). He has organized several National Workshop, Conferences and Technical Events. He is regularly invited to deliver lectures in various programmes for imparting skills in research methodology to students and research scholars. He has published three books, in the name of Data structures & Algorithms, Computer Programming and Problem Solving and Python Programming in DD Publications, Excel Publications and Sri Maruthi Publisher, Chennai, respectively.

Koteeswaran Seerangan currently working as an Associate Professor in the Department of Computer Science and Engineering at Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai-62, Tamil Nadu, India. He has authored and co-authored several papers in various reputed journals and conference proceedings. He is a reviewer for more than a dozen of journals. His research interests include Theory of Computation, Software Engineering, Data Mining, Big Data and Cloud Computing. He is a Member of ACM, Senior Member of IEEE and Life Member of ISTE.