Received 29 September 2019
Revised 29 December 2019
Accepted 1 February 2020

A social network based approach to identify and rank influential nodes for smart city

Bharat Arun Tidke
VIT-AP University, Amaravati, India, and
Rupa Mehta, Dipti Rana, Divyani Mittal and Pooja Suthar
Department of Computer Engineering,
Sardar Vallabhbhai National Institute of Technology, Surat, India

Abstract

Purpose – In online social network analysis, the problem of identification and ranking of influential nodes based on their prominence has attracted immense attention from researchers and practitioners. Identification and ranking of influential nodes is a challenging problem using Twitter, as data contains heterogeneous features such as tweets, likes, mentions and retweets. The purpose of this paper is to perform correlation between various features, evaluation metrics, approaches and results to validate selection of features as well as results. In addition, the paper uses well-known techniques to find topical authority and sentiments of influential nodes that help smart city governance and to make importance decisions while understanding the various perceptions of relevant influential nodes.

Design/methodology/approach – The tweets fetched using Twitter API are stored in Neo4j to generate graph-based relationships between various features of Twitter data such as followers, mentions and retweets. In this paper, consensus approach based on Twitter data using heterogeneous features has been proposed based on various features such as like, mentions and retweets to generate individual list of top-k influential nodes based on each features.

Findings – The heterogeneous features are meant for integrating to accomplish identification and ranking tasks with low computational complexity, i.e. O(n), which is suitable for large-scale online social network with better accuracy than baselines.

Originality/value – Identified influential nodes can act as source in making public decisions and their opinion give insights to urban governance bodies such as municipal corporation as well as similar organization responsible for smart urban governance and smart city development.

Keywords Smart city, Centrality measures, Heterogeneous features, Topic authoritative

Paper type Research paper

1. Introduction

Smart city (Gallion and Eisner, 1980; Yeh, 2017) portrays a list of institutional, physical, social and economic infrastructure that provide smart application-based services that define a level of aspiration of its citizens (node). Smart cities are first and foremost about people (Ahmed *et al.*, 2016); the traits of smart city life are greatly digitized and data-driven. Issues accompanying individuals and communities (Gallion and Eisner, 1980; Alawadhi *et al.*, 2012; Yeh, 2017; Alp and Öğüdücü, 2018; Charalabidis *et al.*, 2019) in the paradigm of urban and smart cities development have been ignored in the outlay of a cavernous understanding of the smart technical and scientific aspect.

Researchers emphasize on harnessing online social network (OSN) platforms (Alawadhi *et al.*, 2012; Yeh, 2017), which are crucial in decision-making for specific topics that have

Kybernetes
Vol. 50 No. 2, 2021
pp. 568-587

0368-492X
DOI 10.1108/K-09-2019-0637

significance for their citizens, and in what manner these citizen's opinions can contribute to topic as well as overall development of smart cities. OSN methods and tools (Cha *et al.*, 2010; Ahmed *et al.*, 2016) are important to understand the preferences of the citizens. OSN data are also linked with location background (Ahmed *et al.*, 2016; Brandt *et al.*, 2017) such as location detection, or environmental and cultural indication. Smart cities need urban governance to consider the interests and opinions of citizens. However, taking opinions from all citizens can create chaos while making decisions.

Researchers (Alp and Öğüdücü, 2018) claimed that 1 per cent of nodes on Twitter, those that are influential, have an impact on the 25 per cent of the information transmission in society. Research in computing the influence of a user (node) from OSN data (Pal and Counts, 2011; Song *et al.*, 2017) is a notional problem. There is no concrete characterization for influential node (Riquelme and González-Cantergiani, 2016). Consequently, novel influence metrics and methods (Wei *et al.*, 2016; Riquelme and González-Cantergiani, 2016) are continually evolving; most of them agree on various evaluation criteria. These varied influence metrics comprise the description of different categories of nodes such as influential nodes (Al-Garadi *et al.*, 2018), opinion leaders (Yang *et al.*, 2018), topical experts (Wei *et al.*, 2016), authoritative actors (Pal and Counts, 2011; Oro *et al.*, 2018) and influence spreader or disseminators (Song *et al.*, 2017). The research aims to identify the influential nodes that comprise abovementioned categories of nodes as compared to influence maximization, which mainly considers nodes that spread or propagate information or influence of the influential nodes.

Twitter becomes popular platform (Riquelme and González-Cantergiani, 2016) to collect large volume of data in terms of relationships and user generated contents such as tweets. These data include varied topics, such as daily conversation, news headlines, philosophy, research and technology. Compatibly, users on Twitter display proficiency and authority (Oro *et al.*, 2018) on varied topics. A topical influential node (Wei *et al.*, 2016; Pal and Counts, 2011; Oro *et al.*, 2018) in a social networking site such as Twitter recurrently posts valuable and relevant information for particular topic. These influential nodes with many followers and large number of mention, retweets and likes can be experts or may have authority on the topic at hand (Alp and Öğüdücü, 2018; Riquelme and González-Cantergiani, 2016; Al-Garadi *et al.*, 2018).

1.1 Motivation

Smart cities need urban governance (Charalabidis *et al.*, 2019) to consider the interests and opinions of citizens. Researchers emphasize on the role of social media platforms in critical decision-making on specific domains that have significance for their citizens and in the manner in which these citizens can contribute to overall development of smart cities. Issues accompanying individuals and communities, in the paradigm of urban and smart cities development, have been ignored in the outlay of a cavernous understanding of the smart technical and scientific aspect. However, taking opinions from all citizens can create chaos while making decisions. The influential citizens can act as a source to provide an insight into what citizens are talking or feeling about any topic. The perception of existing works on identification and ranking of influential nodes (IRIN) have high discrepancy, which shows an influential node varies with variation in network topology and nature of information flows in OSN. Formation of an efficient process for IRIN is important in collecting, analyzing and organizing unstructured revealing information into intelligible concepts for various decision support applications in the multidimensional or multilayer network.

In addition, real-world social networks are not often static. The structure, user-generated contents and the influence strength related to nodes present in the social networks evolve

continually. As a result, set of influential nodes and their respective influence score should be continuously rationalized conferring the evolution of social network data.

1.2 Contributions

- To identify the need for citizen-centric methodologies in smart cities to enhance the decision-making capabilities based on time stamp.
- Analysis of correlation between various features, evaluation metrics, approaches and results
- Ensure proposed aggregation approach can be stretched to test different heterogeneous features on diverse data sets and applications.
- Use well-known existing techniques to find topical authority and sentiments of influential nodes.
- Influential nodes can act as source in making public decisions and their opinion give insights to urban governance bodies such as municipal corporation as well as similar organization responsible for smart urban governance and smart city development.

2. Related work

Aral and Walker (2012) demonstrated that influential nodes get less influence from common users. These people build more relationships with other influential people with time, which results in the evolution of their network. This network of the known influential node may be used to extract the influence of other individuals or groups whose dynamics prevail over time.

Badar *et al.* (2014, 2016) proposed approaches to discover the effect of the connection between author demographics, research performance of scholar and their degree centrality. They show that the research demographics of authors will not affect impact of author in terms of high degree centrality; however, research performance can lead to high degree centrality. Wang *et al.* (2017) introduced a multi-attribute ranking method using location and neighborhood of node in the network by applying K-shell decomposition method. They show that their approach ranks influential nodes, with low computational complexity i.e. O (n).

The influential citizens (node) (Ahmed *et al.*, 2016) can act as a source to provide an insight into what citizens are talking or feeling about a particular topic. The identified topical influential node aids in reducing the complexity for many applications such as sentiment analysis (Liu, 2012; Brandt *et al.*, 2017) and fake news or rumor detection (Wang *et al.*, 2017). Posts, opinions or replies from topical influential nodes on same or different topics can be useful for finding out polarity in terms of positive, neutral or negative for detecting trends (Moreno-Munoz *et al.*, 2016), recommendation of product and services (O'Mahony and Smyth, 2018) and while taking decision on various aspects that can be used by businesses and governments organizations to act accordingly. For illustration, if a citizen wishes to follow topical influential node on Twitter for getting contents as well as user information that are very much appropriate for a smart city-based topics such as "Healthcare- or sports-related activity." It also enables citizens to grab attention on particular topic with relevant information quickly and by removing irrelevant nodes while looking at influential score improves the trustworthiness of the information.

Numerous centrality based techniques (Borgatti, 2005; Newman, 2005; Lee *et al.*, 2010; Riquelme and González-Cantergiani, 2016; Al-Garadi *et al.*, 2018), which include degree

centrality, closeness centrality, betweenness centrality and eigenvector centrality existed for IRIN. These techniques detect the solitary position of node which ultimately estimates their importance for a specified network. To compute influence spread, influence maximization models such as independent cascade model (Goyal *et al.*, 2011) and linear threshold (Wang *et al.*, 2012) are mainly applied by researchers and scientist for various application domains. These techniques for IRIN in Twitter faces few challenging issue such as influence of a node is change and replicate due to network topology (Tidke *et al.*, 2018), similarly, correct set of influential nodes are different for diverse application. Another challenge lies to select correct metric from numerous combinations or disagreeing ranking pattern.

Gao *et al.* (2015) discussed the importance of integrating multiple features for accomplishment of more accurate IRIN. In the recent past, techniques such as hierarchical (Zareie and Sheikhahmadi, 2018), multilayer network (Alp and Öğüdücü, 2018), multi-criteria decision-making (MCDM) (Mardani *et al.*, 2015) and multi-attribute integrated measurement (MAIM) (Wang and Zhao, 2015) came into existence for IRIN comprehensively. However, these approaches as well as traditional centrality measures are not competent in large and dynamic networks due to their high computational complexity (Zareie, and Sheikhahmadi, 2018). For example, closeness and betweenness centrality takes $O(n^3)$ to find most central node in given network (Riquelme and González-Cantergiani, 2016). Similarly, an iterative approach, i.e. PageRank (Page *et al.*, 1998), a concrete method to identify influential nodes in broad range of applications takes O (n*m) computational complexity where n denotes number of nodes in network and m represents the number of edges. Mariani *et al.* (2015) pointed out that PageRank fails for IRIN on large-scale networks. Therefore, in large directed network scenarios, it is crucial to recognize the low computationally cost diverse features, which are key for IRIN in consequence.

Tidke *et al.* (2019) proposed a consensus based approach to identify and rank top-k influential nodes using various keywords based on particular topic. They used aggregation approach based on heterogeneous features such as tweets, likes, mentions and retweets. Wang *et al.* (2019) proposed deep learning approach latent feature representation of users for predicting influence spread.

2.1 Topical authority

Weng *et al.* (2010) put forward one of earlier work in topic based influence ranking approach "Twitter-Rank" which concentrates on both the subjective similarity and the topological structure. The approach is divided into phases: first is to detect topic from Twitter data, i.e. tweets using latent Dirichlet allocation (LDA) model for each individual node. In next phase, it generates weighted graph for separate topic by considering topical similarity and in degree as weights before applying PageRank. Xiao *et al.* (2014) proposed two methods, i.e. RetweetRank and MentionRank to identify influential nodes using hashtag based community. They consider topic from news to identify influential nodes. They created two set of influential nodes based on retweet and mention rank which suggest that RetweetRank finds influential nodes whose tweets are important on specific topic and draw attention of other nodes. MentionRank finds influential nodes with authority on topic and they compare their approach with follower rank and PageRank.

Similarly, Alp and Öğüdücü (2018) proposed Personalized PageRank that assimilates network topology and the contents from Twitter. Their goal is to find top k influencers who are powerful or have authority on a particular topic. Yang *et al.* (2018) introduces novel betweenness centrality algorithm, to find top opinion leaders and compared their results with independent cascade model based approaches. They showed the improvement of their approach by performing various experiments on various social network data. Sun *et al.* (2016)

implemented a machine learning approach based on features such as followers, following, tweet count, retweet count etc. to identify influential nodes and compare their proposed approach with follower rank and weighted rank. However, they have not mentioned how various features are integrating or they apply any learning model to ensure their evaluation criteria. In addition, they have not used any standard evaluation measures to check correctness of their results.

2.2 Sentiment analysis

Jansen *et al.* (2009) explained and compared different classification and manual coding techniques approach on user's brand sentiments using Twitter data and they found out that social networking site like Twitter is a platform where user can communicate with various commercial businesses. Trattner and Kappe (2013) suggested opinions and sentiments can be useful for purchasing online products as they performed several experiments on Facebook stream Bigdata. Ma *et al.* (2008) performed sentiment analysis using three different models based on heat diffusion process and comes up with opinion that same can be applied to social network Bigdata analytics also in scalable manner. Social media data (Aggarwal, 2011; Bello-Orgaz *et al.*, 2016) can improve the quality of new products based on customer opinions or reviews.

3. Methodology

The research in influence analysis governs broad areas of application and needs various information which encompasses structure of social network that can be useful to detect influential users (Wei *et al.*, 2016; Riquelme and González-Cantergiani, 2016) and indication in terms of other entities which symbolizes their activity. The influence analysis on social network can be performing on either entire network, i.e. global influence or some part of entire network, i.e. local influence.

3.1 Problem statement

Let the number of Twitter node be $\delta = (x_i)_k$ (k is the total users), which are aspirant authorities on particular topic. Node x post tweets on any topic contains Username, Mention, Hashtags, Likes, Retweets, Text Links, Image, etc. These various features are extracted from the tweet based on several topic (T), related keywords $T_k = (t_1, t_2 \ldots t_k)$, where, k is the number of keywords.

The topical influential or authority node problem is defined as identification and ranking of k number of aspirant authorities using their expertise and influence on the topic T based on time. The sentiments of influential nodes can be act as a preprocessing step while finding sentiments of masses or people for given query related to particular topic.

3.2 Proposed architecture

Figure 1 shows the overall strategy of proposed work for identification and ranking of topical influential node. First, we fetch the data from Twitter based on keyword query given by the user. Any number of keywords can be used to fetch the data for a topic. For example, if user wants to retrieve information about Cricket, the keywords can be Test match, World_Cup _2019_England, IPL, etc. or can be any hashtags related to Cricket on Twitter, i.e. #bcci or #icc.

3.2.1 Network-based analysis. The tweets fetch using Twitter API are stored in Neo4j to generate graph based relationships between various features of Twitter data such as followers, mentions and retweets .In this paper, consensus approach based on Twitter data

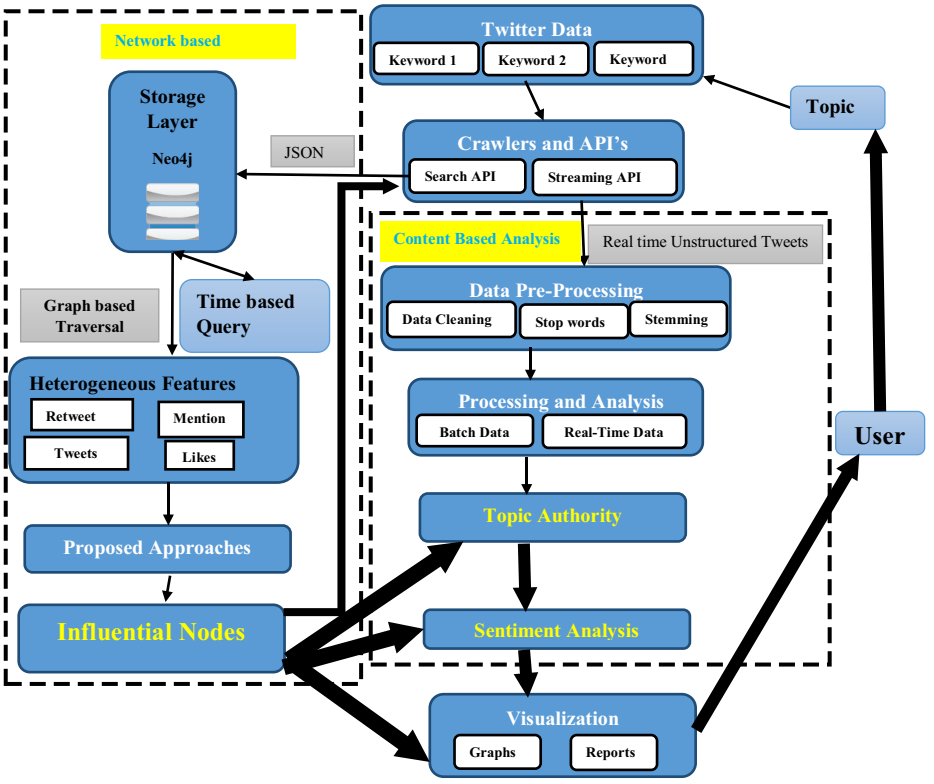

Figure 1.
Overview of proposed architecture

using Heterogeneous Features has been proposed that considers based features, i.e. likes, mentions and retweets, to generate individual list of top-k influential nodes based on each features.

The heterogeneous features is assimilation of above mention features which is implemented to accomplish identification and ranking task with low computational complexity, i.e. O(n), which is suitable to large scale OSN. Now, the list generated from each feature will be consensus aggregated for getting final ranked top-k nodes.

3.2.2 Preliminaries. For $\delta = \{x_1, x_2, \ldots, x_i\}$, where x_i indicates the cardinality of δ. Node influence (NI) ranking on δ is orderly list: $\beta = (x_1 > x_2 > \cdots > x_i)$, where $x_i \in \delta$. Ranking influential node problem in proposed approach extract top-k ranked node based on consensus aggregation that fulfils the criteria of heterogeneous surface learning features (HSLF) will be present in the top k list denoted by μ.

Definition 1. Node $x \in \delta$ exist in θ, $\theta(i)$ signifies the rank of i in θ. Assume, every individual heterogeneous feature provides list of influential nodes, now, the multiple list generated based on each Heterogeneous Features contributes n rankings $\{\theta_{i1}\ \theta_{i2} \cdots \theta_{in}\}$. A consensus rank-position aggregation (RPA) of μ_i is a ranking:

- on the set of list of top-k nodes generated by proposed HSLF as compares to complete rankings aggregation of all nodes N, that minimizes the size of list by Δ (€, N) where Δ is the subtraction of € from N and we denote the set RPA of N by k(N).

Consider a scenario where proposed HSLF features generates m ranking of n nodes to compute NI as shown in equation (1):

$$NI = \begin{array}{c} \\ HSLF_1 \\ HSLF_2 \\ \vdots \\ HSLF_m \end{array} \begin{array}{c} \begin{array}{cccc} a_1 & a_2 & \cdots & a_n \end{array} \\ \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1n} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{m1} & \theta_{m2} & \cdots & \theta_{mn} \end{bmatrix} \end{array} \tag{1}$$

The ith row of NI denotes the rank list generated by $SLHF_1$ i.e. $\theta_i = [\theta i_1\ \theta i_2 \cdots \theta_{in}]$ and $\theta_{ij} = (1 \leq \theta_{ij} \leq n)$ represent the place of nodes in the ordered list. Now we have the data set of $NIi.e.D_{NI} = (HSLF^1, HSLF_2, \cdots, HSLF_m)$ with various permutation of node ranking. The goal is to generate consensus based rank aggregation of n nodes that best signifies the data set. The following features are used as HSLF for proposed approach:

- *Likes*: The HSLF likes gives the popularity of node's tweets among other nodes (Riquelme and González-Cantergiani, 2016).
- *Mentions*: The HSLF Mention effectiveness used to propagate information to remote node rather than the neighborhood only and give importance to the proper set of nodes that have authority on the topic or relate to it (Xiao *et al.*, 2014).
- *Retweets*: The HSLF retweet effectiveness gives the information on importance of nodes contents on topic which will be propagate by other nodes. In other words, the retweets count gives the ability of a node to produce information that is propagated to other node. A user can send information to her followers by retweeting posts of other users (Riquelme and González-Cantergiani, 2016).

3.2.3 Time stamp features extraction. In real dynamic Twitter data, for network of relationships such as "friends-followers", it is highly unlikely to take rapid and severe variations in topology in short duration, or can be termed as slowly evolving network. However, other features such tweets, mentions, likes and retweets, which belong to the category of content-based data, will come as stream of data. Querying such streaming network data can give us insights in the influence strengths of different nodes in different time stamp. In proposed approach, list generated by various features are time stamp-dependent.

3.2.4 Proposed algorithms. Two different approaches are proposed for ranking of influential nodes, first is ACRA (Algorithm 1) which rank influential nodes with higher score in upper position, however, the nodes with same influential score is randomly placed based on some criteria such as first features nodes can be place in upper position second feature node is placed in second position and so on or can be placed based on their name. To overcome this challenge, WACRA is proposed (Algorithm 2).

Algorithm 1. ACRA

Input: Set of V nodes from Twitter as a graph $G = (V, E)$ and edge (m, n), value of k, n: no. of nodes

Output: Top-k nodes using **ACRA**

1. ***For*** 1 to n

- Compute and Sort $HSLF_1$ Likes score
- Compute and Sort $HSLF_2$ Mention score
- Compute and Sort $HSLF_3$ Retweet score

2. *For* *input k*

- Extract the top k nodes from each feature generated list

3. Create *NI* matrix using $SLHF_1$ to $SLHF_m$ as m * k matrix using equation (5)

For i = 1 to m

For j = 1 to k

- Assigned **value** to each node based on its **Position Rank** in the list

$$Vj = k - j + 1 \tag{2}$$

Where, Vj = value of jth node

- Generate a key-value pair $< V, Vj >$
- Where key represents node and value represents its computed score

For End

For End

4. Compute an ACRA to top-k IRIN

- Let P *denotes* the list of the aggregated weights as score for pairs having same key
- Calculate ACRA using following equation

$$ACRA = \underset{\forall z \in P}{Generate} < Z_{key}, \frac{P_{score}}{m} > \tag{3}$$

- Now, Sort the ACRA in descending order based on score and fetch top-k IRIN

5. *End*

Algorithm 2. WACRA

Input: *Step 3 of Algorithm 1*

Output: Top-k nodes using WACRA

1. Assign weights to each node based on its threshold criteria

For i = 1 to m

For j = 1 to k

$$HSLF_1 = 0.9 * Vj$$

$$HSLF_2 = 0.95 * Vj$$

$$HSLF_3 = 1 * Vj$$

- Generate a key-value pair $< V, Vj >$
- Where key represents node and value represents its computed score

For End

For End

2. Compute an WACRA to top-k IRIN

- Let P *denotes* the list of the aggregated weights as score for pairs having same key
- Calculate ACRA using following equation

$$ACRA = \underset{\forall z \in P}{Generate} < Z_{key}, \frac{P_{score}}{m} > \quad (4)$$

• Now, Sort the ACRA in descending order based on score and fetch top-k IRIN

3. *End*

3.3 Topical authority and sentiment analysis

Topical authority can put forward new ideas or involve in decision-making, in other words he can be expert having proper and deep knowledge about specific topic and post quality content around a specific topic. These topical authorities have opinions in terms of positive or negative while interacting or posting opinions on particular topics. For example, during a voting or making decision, noticing the influential node opinion could help the political party or organization what segment of citizens probably are in favor of party or which employee are upset about particular strategy of the organization.

- *JSON to CSV*: Tweets are normally textual data in JSON format. We convert the JSON tweets in text and store in csv.
- *Removal of symbol*: Tweets data are often noisy to perform any machine learning task. They consist of number of unique symbols such as #, @ may not give any useful details while mining text. Therefore, before processing all symbols are removed.
- *Removal of irrelevant words*: Eliminating irrelevant words from tweets such as common Pronouns, Conjunction, and Prepositions. We consider the remaining highly relevant words for further process. All irrelevant words are stored in one dictionary (stop-word list). We used NLTK stop-word dictionary to remove all irrelevant words.
- *Stemming*: After removing irrelevant words next we perform stemming process that discards prefixes or suffixes of the words and construct a lexicon for speedy stemming process.
- *Lexical diversity:* It is the ratio of number of unique to the number of total words in data set. The main advantage of using lexical diversity is that its works better while comparing equal length text. The value near or equal to 1 indicates all or most of the words in a data set are unique, vice versa lexical diversity near or equal to 0 implies most or all words are duplicate.
- *Top words:* The next step is to compute the frequency of individual word called as top words in the tweets of node. The top 10 words with highest frequency is store. The top words are matched with hashtags of topic. The more top words match the hashtags more authority node has on particular topic.

Algorithm 3: Word Count (To count the existence of individual term (t) in each Tweet)

```
Let: Term- (t), TweetId- Tid, C-Count, Sum- S
  class C
  procedure Count (Tweet)
  for each (t) ∈ Tweet
  tick (((t), Tid), 1)
  class S
       procedure Sum ((t), Tid), K [1,1,..., n])
  A = 0
  for each n ∈ K
```

```
        S = S +1
        return (((t),Tid),K)
          Now, to compute authority of node, the following equation is
used
```

$$\text{Authority of node on topic } T = \frac{\text{Total number of top keywords matches}}{\text{Lexical diversity score}} \quad (5)$$

The next step is to find top–k topical expert based on *equation* (8);

Finally, calculate sentiments of each influential nodes using sentiment technique of TextBlob library to determine the polarity of a node on particular topic. The topic is categorized as, or negative, neutral, or positive based on polarity score. If less than 0, then the node has negative opinion on topic T equal to 0 it is neutral and greater than 0 then nodes has positive opinion on topic T.

4. Experiments

In this chapter, a series of experiments performed using real Twitter data to evaluate our theoretical analysis and proposed approach.

4.1 Experimental setting

All relationships in terms of graphs are stored in Neo4j graph database. Algorithms are implemented in python with py2neo library support. Experiments are run on a PC with Intel (r) core(TM) i7-7700 cpu @ 3.60ghz dual CPU and 16 GB RAM with OS Ubuntu 16.04.5 LTS (XenialXerus).

4.2 Data set

The analysis of proposed algorithms is carried out using Twitter online data which is one of the most used microblog site having 330 million monthly active users and increasing rapidly every day. The data is fetched using various keywords (Keyword for Fetching Twitter Data) from September 17, 2018, to November 13, 2018. Total number of nodes and relationships fetch with respect to keyword are mention with labels in Table I:

- *Politics*: Indian_Politics, loksabha, Indian_Parliament, Indian_Elections:

4.3 Neo4j

Neo4j reduces the requisite of several graph modelling techniques and platform with proficient, convenient algorithms by giving in-built procedures and functions. Moreover, provides stable time traversals in large graphs for both breadth and depth, since it represents nodes and relationships in efficient manner. The scalability of Neo4j enables large with billions of nodes to traverse efficiently on commodity machines.

In addition, it provides flexibility in property graph schema, which enables it to adapt in dynamic networks scenario. Figure 2 shows the graph structure of various relationships

Table I. Statistics of nodes and relationships of topics

	Nodes	Relationships
Politics	Tweets: 62,422 Users: 41,326	Followers: 581,761 Mentions:170,332 Retweets: 32,814

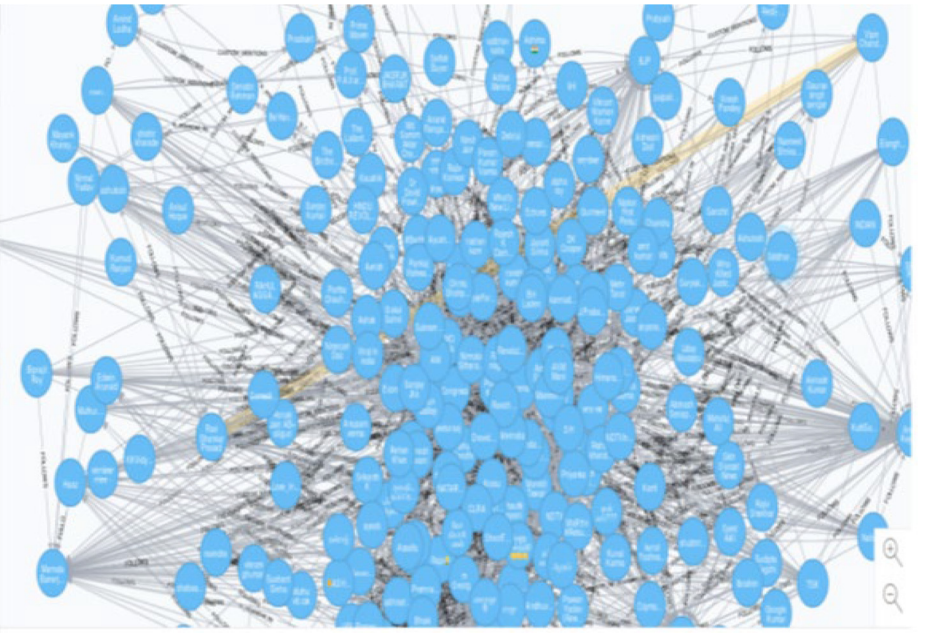

Figure 2.
Example snapshot of Twitter data in Neo4j Graph database

stored in neo4j. Twitter data relationships such as friends, mentions and retweets can be formed efficiently using Neo4j.

4.4 Evaluation measures

The experimental authentication on the level to which position of node signify influence is carried out using three different measures i.e. ACC, F-score and MCC (Tidke *et al.*, 2019):

- *Accuracy (ACC):* Accuracy is mainly perceptive evaluation measure and it is basically a ratio of accurately predicted results to the total results. It evaluates the complete usefulness of the model by calculate the likelihood of the correctness of predicted output (equation 9). In addition, error rate in accuracy measure is the probability of incorrect output given by prediction model:

$$Accuracy(ACC) = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

- *F-measure:* F-measure is a well-known evaluation measure for binary classification problem. Therefore, it fits the criteria for evaluating our proposed influence analysis methodology which is based on binary detection of influential nodes in a given network of well-known influential node. F-measure (equation 10) is a fusion of precision and recall, which are efficient metrics for influential node identification where the imbalance problem exists. The F-measure is the harmonic mean of precision and recall, i.e:

$$F - measure = \frac{2{*}P{*}R}{(P + R)} \tag{7}$$

Where, $precision(P) = \frac{TP}{(TP+FP)}$ and $recall(R) = \frac{TP}{(TP+FN)}$

- *Matthews correlation coefficient (MCC):* This measure is frequently used in binary classification problem for imbalanced data set as it takes mutual accuracies and error rates on both values (True and False), and considers all classes of confusion matrix. MCC gives value ranges between 1 for best case output to −1 for the worst case output. If prediction models perform arbitrarily the MCC gives value near to 0.

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{8}$$

- *Kendall's Tau Rank Correlation:* Kendall's Tau (τ) (Lapata, 2006) can be inferred as function of the likelihood of detecting concordant and discordant pairs. It calculates the difference probability in detected data two features same ordered list as opposed to the probability of having different orders. The probability score ranges from −1 to 1 as in equation (8).

For $\beta = (x_1 > x_2 > \ldots > x_i)$, Let μ_1 and μ_2 indicate two separate orderings of β, and $\gamma(\mu_1, \mu_2)$, then, the minimum count of contiguous swaps required to match μ_1 to μ_2. The following equation calculates Kendall's τ as:

$$\tau = 1 - \frac{2\gamma(\mu_1, \mu_2\ )}{N(N-1)/2} \tag{9}$$

Where, N denotes number of nodes being ranked.

4.5 Result analysis

The challenge of positioning node with same score is overcome in proposed WACRA (Algorithm 2) which assign score based on its weighted feature. The followers count shows the effectiveness of WACRA as compare to ACRA, for all standard list, except for few @k, for all remaining @k WACRA performs better as compare to ACRA. Hence, in experimental analysis WACRA is used for comparison with standard list.

4.5.1 Time stamp influential node ranking. Network topology, contents of social network as well as influence spread capabilities among nodes evolve frequently that needs identification and rank of influential nodes under a dynamic situation. To tackle this problem of time stamp influential node identification and ranking, proposed approach explore the time stamp IRIN problem as an extension to WACRA and evaluation measures under time stamp Twitter data.

The time complexity of various measures proposed in related work are compared. This analyse the time complexity of used HSLF, as compared to other available features. Figure 3 shows that the used features such as Likes, Mention and Reweets take less time to generate list as compare to all other features, except, PageRank which also generates list in less time complexity. However, PageRank is not suitable as number of relationships grows with time, as page rank is an iterative approach and iterations may increase with increase in number of relationships. Figure 4 gives the details of data distribution which is extracted from Twitter to analyze with respect to time stamp.

4.5.2 Correlational analysis among selected features. A crucial challenge in IRIN from Twitter data is recognizing a typical set of features to build a model for an IRIN task. The

theory suggested that worthy feature sets comprise features that are highly correlated with particular task, however uncorrelated with each other (Lapata, 2006).

Highly correlated features are likely to be linearly dependent and therefore gave same results or have similar influence, in case of IRIN task on the dependent feature. Consequently, two features with high correlation, one can afford to remove one feature while computing influential nodes. Figure 5 shows that all three selected features have low

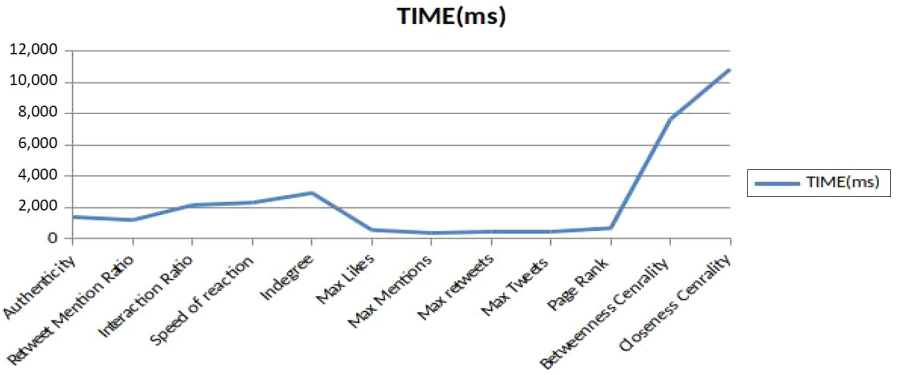

Figure 3.
Time comparison of various features of Twitter data

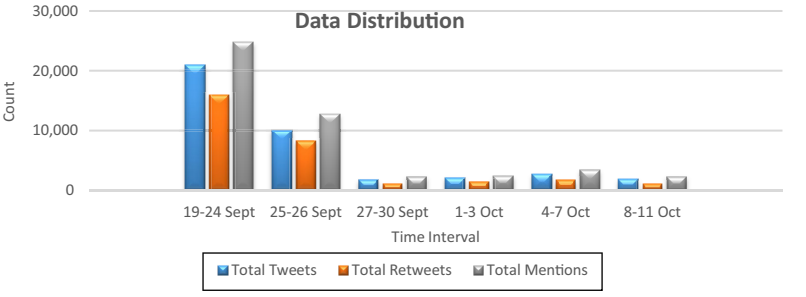

Figure 4.
Data distribution of each feature with respect to time

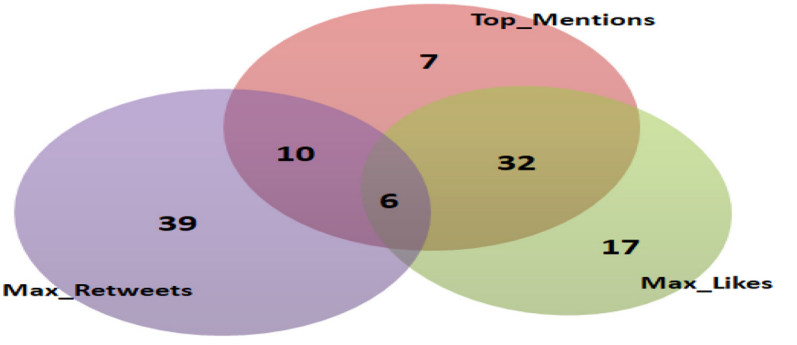

Figure 5.
Correlational analysis of different Twitter features

correlation among themselves with males them perfect candidates for features while generating list of influential nodes.

4.5.3 Benchmark list and evaluation. In OSN system, large number of nodes and nodal activities occur in real time, which creates problems to obtain a benchmark for time stamp-based influential analysis. There are various baseline methods to find interval-based influencers that depend on user interaction.

Pal and Counts (2011) provides many features for evaluating influential nodes in microblogs such as mention impact (MI) and information diffusion (DI). Al-Garadi *et al.* (2018) and Alp and Öğüdücü (2018) used retweet rate of influential node known as spread score to rank them. Many authors (Wei *et al.*, 2016; Alp and Öğüdücü, 2018) used manual annotation to verify their results. Similarly, used real spread dynamics to identify influential nodes. These evaluation measures are either judges influence maximization or use retweets as measure to determine spread. Still, there is lack of standard benchmark for evaluation. A benchmark could be artificially generated based on the characteristics of the Twitter data to evaluate performance of different algorithms applied for finding interval-based influencers. A benchmark for time stamp based analysis would highly depend on user's activity in terms of numbers of likes obtained, number of retweets, and number of mentions a user receives as these are the core features that reflect the influence of the user. The following two lists are used to compare proposed approach:

- *Retweet–mention ratio (RTMN)*: RTMN ratio is a Twitter metric proposed to evaluate ranking of influential nodes (Oro *et al.*, 2018).
- *Ensemble list*: A list generated by combination these features could be set as standard list for evaluation of top-k interval based influencer users (Wei *et al.*, 2016).

Based on this intuition, we generate a standard list as follows:

- node present in top-k list of all three core features;
- node present in top-k list of any two core features; and
- node present in top-k list of only one core feature.

4.5.4 Comparative analysis. Politics data from Twitter in the context of Indian user are extracted using various keywords (Table II). Figure 6 depicts that for politics data set, WACRA performs better in term of accuracy as compare with RTMN rank list. The WACRA is the aggregation of three measures, i.e. likes, mention and retweets based features which gives the topic based influential nodes. F1-measures and MCC gives identical results when compared for all time stamp.

The feature such as mention and retweet are common in proposed approach as well as standard list generated by RTMN ratio. The accuracy measure is biased toward TP and TP value of confusion matrix which results in high accuracy. However, F1-measures and MCC gives somewhat correct comparison while evaluating results generated by WACRA. Similarly, Figure 7 again depicts that for politics data set WACRA performs better in term of accuracy.

	Predicted as positive	Predicted as negative
Actually positive	True positives (TP)	False negatives (FN)
Actually negative	False positive	True negative

Table II.
Confusion matrix

To tackle the problem-evaluating ranking of influential nodes, a correlation based approach i.e. Kendall tau "τ" proved to be efficient. Proposed approach with RTMN and Ensemble list is compared using Kendal tau.

Figure 8 depicts the correlational analysis between WACRA and RTMN with different time stamp. It is observed that there is minor change in the correlation pattern of both the list with respect to time. Figure 9 again shown no sign of major pattern change or in other word there is no drift in the correlation between two lists.

However, with large data set and more successive time stamp, it may possible that there may be change in drift of correlation patterns among various lists.

4.5.5 Experimental application. To check the experimental application of proposed approaches for IRIN, a small case study is performed on tweets of sample random nodes, location-based random sample nodes and WACRA-based influential nodes for two topics, i.e. Rafale (an issue which becomes political as well as economical debate related to the procurement of 36 combat aircraft for cost of Rs 58,000 crore by India from Dassault Aviation France's) and Statue of Unity (debate related to the gigantic statue of the late Indian politician, statesman and independence activist, Shri Sardar Vallabhbhai Patel in Gujarat, which is worth Rs 2,989 crore).

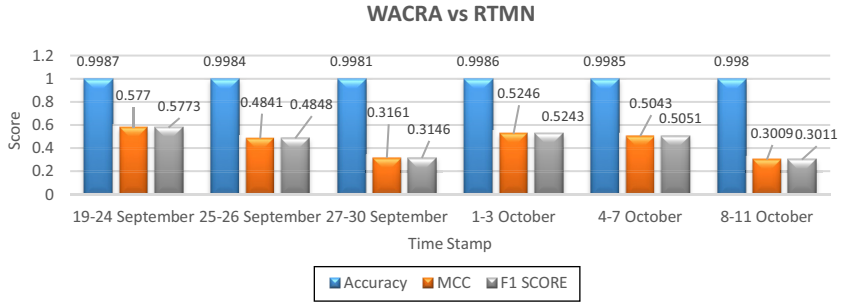

Figure 6.
Comparison of WACRA with RTMN for different time stamps

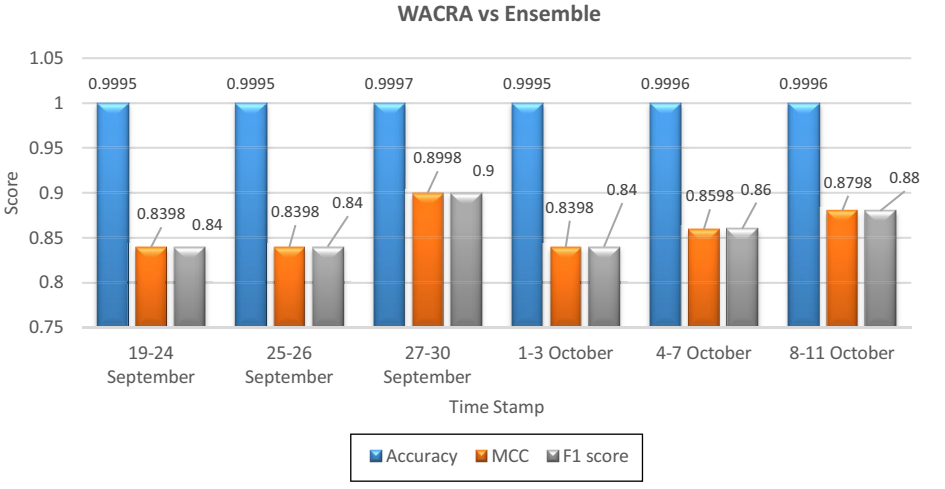

Figure 7.
Comparison of WACRA with ensemble list for different time stamps

Figure 10 shows that sentiments of random nodes are slightly positive on Rafale, with most of the tweets fall in neutral category. The location-based nodes have more positive sentiments for Rafale. For influential nodes, Mr Narendra Modi did not send a single tweet, so score is 0, however, Mr Arun Jaitley and Mr Subramanian Swamy gave their positive opinion on Rafale, in contrast, Mr Rahul Gandhi and Mr P. Chidambaram have negative opinion on it than positive.

Figure 11 shows that sentiments of random nodes are positive on the Statue of Unity as compared to those on Rafale. The location-based nodes again have positive sentiments for Rafale. Among the influential nodes, Mr Narendra Modi and Mr Arun Jaitley are slightly positive and mostly neutral on the Statue of Unity, while Mr Subramanian Swamy has

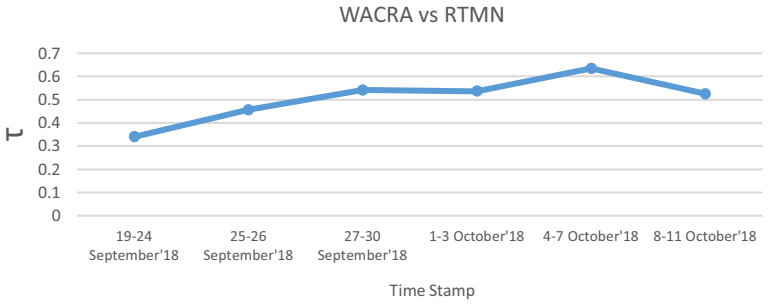

Figure 8.
Correlational analysis of WACRA with RTMN for different time stamps

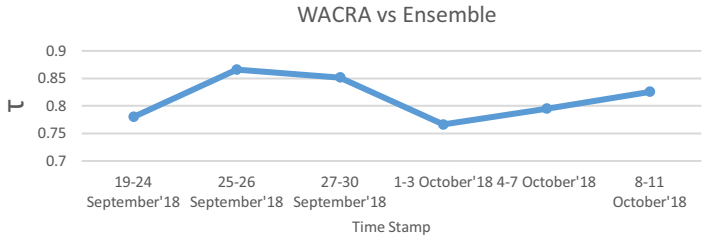

Figure 9.
Correlational analysis of WACRA with ensemble list for different time stamps

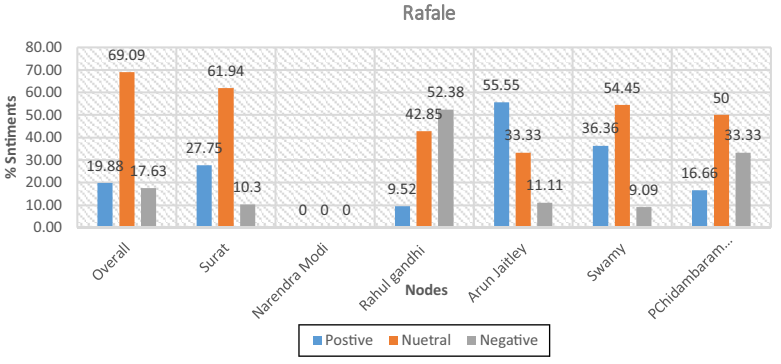

Figure 10.
Sentiment comparison among various nodes for topic "Rafale"

positive and neutral opinions, going by his tweets. In contrast, Mr Rahul Gandhi and Mr P. Chidambaram have negative opinions about the Statue of Unity.

4.5.6 Correlation of node influence. To check the correlation of influential nodes position with various time stamp. Again, Kendall's tau correlation is used. Figure 12 depicts correlation for Mr Narendra Modi (NM) and Mr Rahul Gandhi (RG) to check variation in position for WACRA list for various time stamp. NM influence correlation changes drastically for the period September 25-26 and then remains correlated with remaining time stamp. However, for RG correlation for his position for different time stamp remains highly correlated except little drift in the period October 4-7.

Location-based influential nodes presence can be for politics data. For illustration, we consider the whole world map and place influential nodes on that map based on profile location details.

This information leads user to find out the important location where influential people are present and spreads their influence. For example, an influential node Shehla_Rashid (Screen_name) is visualized with her basic information such as Image link, Name, Location and number of Followers.

5. Conclusion and future work

The notion of information extraction in OSN originates from applicability of different factor prompt to the invention of novel measures for IRIN. These measures are usually based on

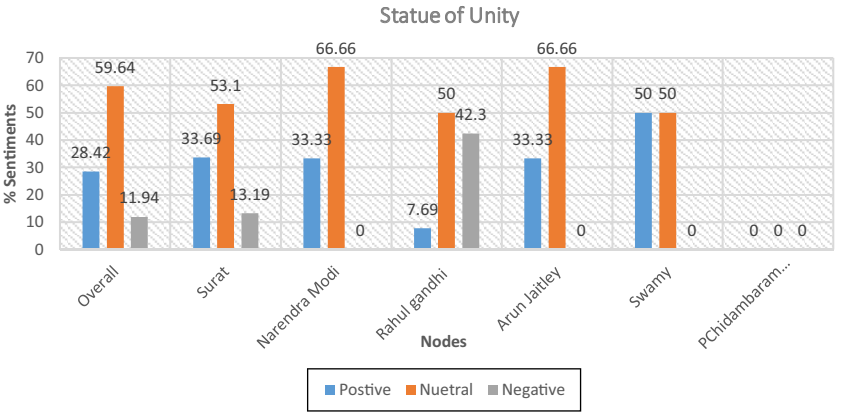

Figure 11.
Sentiment comparison among various nodes for topic "Statue of Unity"

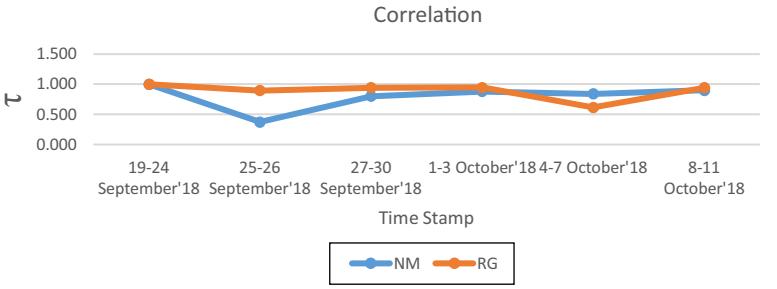

Figure 12.
Correlational analysis of two influential nodes position for different time stamps

their capability to acclimatize current ranking evaluation approaches with focus on user actions and their relationships in OSN, which evolves over time.

This paper presented novel approaches to identify and rank influential nodes for particular topic based on time stamp. Interestingly, proposed approach for topic based influential nodes entirely different from the most of state-of-the-art approaches for IRIN without using the follower/following network. Analysis of correlation between various features, evaluation metrics, approaches and results are performed to validate selection of features as well as results. In addition, proposed aggregation approach can be stretched to test different heterogeneous features on diverse data sets and applications. Identified influential nodes can act as source in making public decisions and their opinion give insights to urban governance bodies such as municipal corporation as well as similar organization responsible for smart urban governance and smart city development

In future work, efficient storage structure can be explored to make graph storage, traversing as well as processing more efficiently. In addition, different data sets of different domains can be identified to apply the proposed work with improved algorithms and models based on applications needed that could be an interesting extension to the current work.

References

Aggarwal, C.C. (2011), "An introduction to social network data analytics", *Social Network Data Analytics*, pp. 1-15.

Ahmed, K.B., Bouhorma, M. and Ahmed, M.B. (2016), "Smart citizen sensing: a proposed computational system with visual sentiment analysis and big data architecture", *International Journal of Computer Applications*, Vol. 152 No. 6, pp. 20-27.

Alawadhi, S., Aldama-Nalda, A., Chourabi, H., Gil-Garcia, J.R., Leung, S., Mellouli, S. and Walker, S. (2012), "Building understanding of smart city initiatives", In *International Conference on Electronic Government*, Springer, Berlin, Heidelberg, pp. 40-53.

Al-Garadi, M.A., Varathan, K.D., Ravana, S.D., Ahmed, E., Mujtaba, G., Khan, M.U.S. and Khan, S.U. (2018), "Analysis of online social network connections for identification of influential users: survey and open research issues", *ACM Computing Surveys*, Vol. 51 No. 1, p. 16.

Alp, Z.Z. and Öğüdücü, ŞG. (2018), "Identifying topical influencers on Twitter based on user behavior and network topology", *Knowledge-Based Systems*, Vol. 141, pp. 211-221.

Aral, S. and Walker, D. (2012), "Identifying influential and susceptible members of social networks", *Science*, Vol. 337 No. 6092, pp. 337-341.

Badar, K., Frantz, T.L. and Jabeen, M. (2016), "Research performance and degree centrality in co-authorship networks: the moderating role of homophily", *Aslib Journal of Information Management*, Vol. 68 No. 6, pp. 756-771.

Badar, K., Hite, J.M. and Badir, Y.F. (2014), "The moderating roles of academic age and institutional sector on the relationship between co-authorship network centrality and academic research performance", *Aslib Journal of Information Management*, Vol. 66 No. 1, pp. 38-53.

Bello-Orgaz, G., Jung, J.J. and Camacho, D. (2016), "Social big data: recent achievements and new challenges", *Information Fusion*, Vol. 28, pp. 45-59.

Borgatti, S.P. (2005), "Centrality and network flow", *Social Networks*, Vol. 27 No. 1, pp. 55-71.

Brandt, T., Bendler, J. and Neumann, D. (2017), "Social media analytics and value creation in urban smart tourism ecosystems", *Information and Management*, Vol. 54 No. 6, pp. 703-713.

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, P.K. (2010), "Measuring user influence in Twitter: the million follower fallacy", *Icwsm*, Vol. 10 Nos 10/17, p. 30.

Charalabidis, Y., Alexopoulos, C., Vogiatzis, N. and Kolokotronis, D.E. (2019), "A 360-degree model for prioritizing smart cities initiatives, with the participation of municipality officials, citizens and

experts", in *E-Participation in Smart Cities: Technologies and Models of Governance for Citizen Engagement*, Springer, Cham, pp. 123-153.

Gallion, A.B. and Eisner, S. (1980), *The Urban Pattern. City Planning and Design*, D. Van Nostrand Company.

Gao, C., Zhong, L., Li, X., Zhang, Z. and Shi, N. (2015), "Combination methods for identifying influential nodes in networks", *International Journal of Modern Physics C*, Vol. 26 No. 6, pp. 1550067.

Goyal, A., Lu, W. and Lakshmanan, L.V. (2011), "Simpath: an efficient algorithm for influence maximization under the linear threshold model", *In Data Mining (ICDM), 2011 IEEE 11th International Conference on*, *IEEE*, pp. 211-220.

Jansen, B.J., Zhang, M., Sobel, K. and Chowdury, A. (2009), "Twitter power: tweets as electronic word of mouth", *Journal of the American Society for Information Science and Technology*, Vol. 60 No. 11, pp. 2169-2188.

Lapata, M. (2006), "Automatic evaluation of information ordering: Kendall's tau", *Computational Linguistics*, Vol. 32 No. 4, pp. 471-484.

Lee, S.H., Cotte, J. and Noseworthy, T.J. (2010), "The role of network centrality in the flow of consumer influence", *Journal of Consumer Psychology*, Vol. 20 No. 1, pp. 66-77.

Liu, B. (2012), "Sentiment analysis and opinion mining", *Synthesis Lectures on Human Language Technologies*, Vol. 5 No. 1, pp. 1-167.

Ma, H., Yang, H., Lyu, M.R. and King, I. (2008), "Mining social networks using heat diffusion processes for marketing candidates selection", in: *Proceedings of the 17th ACM conference on Information and knowledge management*, *ACM*, pp. 233-242.

Mardani, A., Jusoh, A. and Zavadskas, E.K. (2015), "Fuzzy multiple criteria decision-making techniques and applications–two decades review from 1994 to 2014", *Expert Systems with Applications*, Vol. 42 No. 8, pp. 4126-4148.

Mariani, M.S., Medo, M. and Zhang, Y.C. (2015), "Ranking nodes in growing networks: when PageRank fails", *Scientific Reports*, Vol. 5 No. 1, p. 16181.

Moreno-Munoz, A., Bellido-Outeirino, F.J., Siano, P. and Gomez-Nieto, M.A. (2016), "Mobile social media for smart grids customer engagement: emerging trends and challenges", *Renewable and Sustainable Energy Reviews*, Vol. 53, pp. 1611-1616.

Newman, M.E. (2005), "A measure of betweenness centrality based on random walks", *Social Networks*, Vol. 27 No. 1, pp. 39-54.

O'Mahony, M.P. and Smyth, B. (2018), "From opinions to recommendations", *Social Information Access*, Springer, Cham, pp. 480-509.

Oro, E., Pizzuti, C., Procopio, N. and Ruffolo, M. (2018), "Detecting topic authoritative social media users: a multilayer network approach", *IEEE Transactions on Multimedia*, Vol. 20 No. 5, pp. 1195-1208.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1998), *The PageRank Citation Ranking: Bringing Order to the Web*, Stanford InfoLab.

Pal, A. and Counts, S. (2011), "Identifying topical authorities in microblogs", *In Proceedings of the fourth ACM international conference on Web search and data mining*, *ACM*, pp. 45-54.

Riquelme, F. and González-Cantergiani, P. (2016), "Measuring user influence on Twitter: a survey", *Information Processing and Management*, Vol. 52 No. 5, pp. 949-975.

Song, G., Li, Y., Chen, X., He, X. and Tang, J. (2017), "Influential node tracking on dynamic social network: an interchange greedy approach", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 29 No. 2, pp. 359-372.

Sun, Q., Wang, N., Zhou, Y. and Luo, Z. (2016), "Identification of influential online social network users based on multi-features", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 30 No. 6, p. 1659015.

Tidke, B., Mehta, R. and Dhanani, J. (2018), "SIRIF: supervised influence ranking based on influential network", *Journal of Intelligent and Fuzzy Systems*, Vol. 35, pp. 1225-1237.

Tidke, B., Mehta, R. and Dhanani, J. (2019), "Consensus-based aggregation for identification and ranking of top-k influential nodes", *Neural Computing and Applications*, pp. 1-27.

Trattner, C. and Kappe, F. (2013), "Social stream marketing on Facebook: a case study", *International Journal of Social and Humanistic Computing*, Vol. 2 Nos 1/2, pp. 86-103.

Wang, B., Chen, G., Fu, L., Song, L. and Wang, X. (2017), "Drimux: dynamic rumor influence minimization with user experience in social networks", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 29 No. 10, pp. 2168-2181.

Wang, C., Chen, W. and Wang, Y. (2012), "Scalable influence maximization for independent cascade model in large-scale social networks", *Data Mining and Knowledge Discovery*, Vol. 25 No. 3, pp. 545-576.

Wang, Z., Du, C., Fan, J. and Xing, Y. (2017), "Ranking influential nodes in social networks based on node position and neighborhood", *Neurocomputing*, Vol. 260, pp. 466-477.

Wang, S. and Zhao, J. (2015), "Multi-attribute integrated measurement of node importance in complex networks", *Chaos: An Interdisciplinary Journal of Nonlinear Science*, Vol. 25 No. 11, p. 113105.

Wang, F., She, J., Ohyama, Y. and Wu, M. (2019), "Deep-learning-based identification of influential spreaders in online social networks", in *IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society*, Vol. 1, *IEEE*, pp. 6854-6858.

Wei, W., Cong, G., Miao, C., Zhu, F. and Li, G. (2016), "Learning to find topic experts in Twitter via different relations", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28 No. 7, pp. 1764-1778.

Weng, J., Lim, E.P., Jiang, J. and He, Q. (2010), "TwitterRank: finding topic-sensitive influential Twitterer", In *Proceedings of the third ACM international conference on Web search and data mining*, *ACM*, pp. 261-270.

Xiao, F., Noro, T. and Tokuda, T. (2014), "Finding news-topic oriented influential Twitter users based on topic related hashtag community detection", *J. Web Eng*, Vol. 13 Nos 5/6, pp. 405-429.

Yang, L., Qiao, Y., Liu, Z., Ma, J. and Li, X. (2018), "Identifying opinion leader nodes in online social networks with a new closeness evaluation algorithm", *Soft Computing*, Vol. 22 No. 2, pp. 453-464.

Yeh, H. (2017), "The effects of successful ICT-based smart city services: from citizens' perspectives", *Government Information Quarterly*, Vol. 34 No. 3, pp. 556-565.

Zareie, A. and Sheikhahmadi, A. (2018), "A hierarchical approach for influential node ranking in complex social networks", *Expert Systems with Applications*, Vol. 93, pp. 200-211.

Further reading

Cook, D.J. and Holder, L.B. (Eds). (2006), *Mining Graph Data*, John Wiley and Sons.

Tan, P.N. (2018), *Introduction to Data Mining*, Pearson Education India.

Tidke, B., (2018), and R. and Mehta, "A comprehensive review and open challenges of stream big data", in *Soft Computing: Theories and Applications*, Springer, Singapore, pp. 89-99.

Tidke, B., Mehta, R and Dhanani, J. (2017), "A comprehensive survey and open challenges of mining Bigdata", In *International Conference on Information and Communication Technology for Intelligent Systems*, *Springer*, *Cham*, pp. 441-448.

Corresponding author

Bharat Arun Tidke can be contacted at: batidke@gmail.com