Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 11, No. 4, 2023, 535-545.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Identification of Influential Nodes in Social Networks based on Profile Analysis

Zeinab Poshtiban, Elham Ghanbari* and Mohammad Reza Jahangir

Department of Computer Engineering, Yadegar-e-Imam Khomeini (RAH) Shahre Rey Branch, Islamic Azad University, Tehran, Iran.

Article Info

Article History:
Received 30 June 2023
Revised 08 August 2023
Accepted 24 September 2023

DOI:10.22044/jadm.2023.13276.2457

Keywords:
Social Networks, Profile Analysis, Emotion Retrieval, User Identification.

**Corresponding author: el.ghanbari@iau.ir (E. Ghanbari).*

Abstract

Analyzing the influence of people and nodes in social networks has attracted a lot of attention. Social networks gain meaning, despite the groups, associations, and people interested in a specific issue or topic, and people demonstrate their theoretical and practical tendencies in such places. Influential nodes are often identified based on the information related to the social network structure, and less attention is paid to the information spread by the social network user. The present study aims to assess the structural information in the network to identify influential users in addition to using their information in the social network. To this aim, the user’s feelings were extracted. Then an emotional or affective score was assigned to each user based on an emotional dictionary, and his/her weight in the network was determined utilizing centrality criteria. Here, the Twitter network was applied. Thus the structure of the social network was defined, and its graph was drawn after collecting and processing the data. Then the analysis capability of the network and existing data was extracted and identified based on the algorithm proposed by the users and influential nodes. Based on the results, the nodes identified by the proposed algorithm are considered high-quality, and the speed of information simulated is higher than other existing algorithms.

1. Introduction

Online social networks have become integral to communication, skill-sharing, and content sharing. The users actively post on these platforms, providing information such as their location, interests, and educational background. This profile information serves as the basis for grouping users and facilitating content sharing and direct interaction among individuals. The manual input of profile attributes by the users highlights their crucial role in online social networks, particularly when individuals are members of multiple platforms. The analysis of people and nodes within social networks has garnered significant attention. Organizations, campaigns, and institutions are particularly interested in understanding people's behavior and interests within these networks. As a result, several studies have focused on exploring this issue [1] [3]. These studies aim to shed light on the influence and impact of individuals and nodes within social networks, as well as how their behavior and interests can be analyzed and understood.

The recent studies have shown the effectiveness of applying social network information modeling in various domains including fraud detection, quote analysis, and marketing [2]. The presence of communication links between data within each domain presents a unique opportunity to enhance the performance of models. This observation supports the hypothesis that individuals' friends within social networks often share similar interests, traits, abilities, and views. For instance, when it comes to fraud and malicious activities, the concept of hemophilia is particularly relevant. If a person is involved in fraudulent behavior, it is likely that their colleagues and connections within the social network may also be implicated. This interconnectedness can lead to the formation of campaigns centered around fraudulent activities [5] [6]. Campaigns encompass various types such as advertising, political, social, and electoral campaigns. The key to a successful campaign lies in determining the strategies and policies necessary to attract the main elements that possess the required commitment, expertise, strategy, and motivation to achieve the campaign's objectives. This critical foundation of a campaign is known as the "campaign core," which defines the strategic vision essential for the campaign's success.

The objective of this paper is to identify and detect influential nodes within social networks and analyze their campaigns, which often exert a guiding role and influence the dynamics of social networks. These influential nodes serve various purposes including cultural, educational, and security aspects. The paper aims to identify influential individuals in social networks and analyze their behavior to achieve organizational, governmental, and other objectives. Additionally, the paper aims to assess individuals' profiles on social networks and analyze their behavior based on emotions and weight. To achieve these goals, we focus on previous events and content created by individuals and determine their weight within the network. This weight plays a crucial role in identifying influential individuals within a group or social network. Furthermore, we analyze individuals' behavior within campaigns by examining their previous events and activities. By considering the weight and sentiment coefficient, the analysis of social networks can be improved.

Traditionally, influential nodes are identified based on the structure of the social network, with less emphasis on the content propagated by users themselves. Although some studies have analyzed individuals in online social networks, the user-generated information has not been fully utilized. This paper aims to overcome this limitation by incorporating the user information in addition to the network structure to more accurately identify influential users. In this paper, we collect and store the physical structure or topology of the social network and categorize and analyze the information spread by nodes within the network. The topology helps evaluate the node's status and centrality within the network, while the analysis of information spread allows for understanding public opinion about specific nodes. All of this information is utilized to detect influential nodes.

The critical hypothesis of this paper is that the user's content and the structure of the network in which they are located play significant roles and should be utilized to identify and predict their behavior. To achieve this, the proposed framework extracts users' feelings by collecting data and assigns emotional or affective scores to the users, determining their weight in the network using centrality criteria. The proposed framework demonstrates superiority over existing algorithms, as evidenced by the results obtained from Twitter.

The discussion presents the main objectives, methodology, and hypothesis of the study, highlighting the importance of incorporating user-generated content and network structure to identify influential nodes within social networks. Additionally, it emphasizes the potential advantages of the proposed method compared to the existing algorithms. First, the previous works are discussed. Then the proposed algorithm is introduced and evaluated. Finally, conclusions and future works are expressed.

2. Previous Works

Data mining, which refers to the process of discovering knowledge, has a relatively recent history, and has found applications in various fields [3] [4]. One significant aspect of study within this field pertains to privacy concerns in social networks. To tackle this issue, the researchers have employed machine learning algorithms including Bayesian networks to model the causal relationships among individuals in these networks. The findings indicate that personal characteristics can be accurately inferred, particularly when individuals share strong connections. Remarkably, even in a society where individuals tend to conceal their personal attributes, privacy information can still be deduced [8].

Wu *at al*. [9] investigated link prediction as key step in many complex network analysis areas such as social friend recommendation, which aims to infer the missing or future connections between two nodes. To this aim, they presented a general framework of similarity-based link prediction from the view point of influential common neighbors' identification. It incorporates both global-view contributions of common neighbors as well as locality influence into a top-k prediction framework. The experimental results demonstrated that this proposed metrics achieves better performance than the existing state-of-the-art local and global similarity methods [9].

Influence maximization is one of the topics that have attracted much attention in social networks analysis. An important challenge in influence maximization is to find the most influential nodes based on their structural location in the network. In this paper [10], the authors proposed a new method to estimate influence of nodes in information networks. They proposed an improved cluster rank approach that takes into account common hierarchy of nodes and their neighborhood set. A number of experiments are conducted on synthetic and real networks to reveal effectiveness of the proposed ranking approach. We also consider ground-truth influence ranking based on susceptible–infected–recovered model, on which performance of the proposed ranking algorithm is verified [10].

Further, Raamakirtinan *et al*. [11] identified influential users on Facebook based on an emotion-based perspective, and presented a hybrid view in which the influence of a user is calculated using the Sentiment Weighted Page Ranking (SWPR) algorithm. Based on the SWPR, each interaction between two nodes is regarded, and the related tendency is calculated. Then the degree of centrality is considered to calculate the overall rank, and the sentiment related to a user is fed as his/her weight. After testing the proposed method with 532 nodes and tracking data, the top 10 users ranked by SWPR were spread higher and faster than the influential users with page ranking. The results indicated that the spread rate increases as more friends of users receive the message. Network structure and user positions do not suffice to identify influential users accurately. Considering and processing the users' content including wall posts and their feelings, along with the network structure gives the system more understanding about the users, resulting in limiting influential users more accurately compared to other methods. In other words, regarding the feelings of the user as a weight leads to more accuracy than traditional methods.

Furthermore, Deng *et al*. [12] proposed a frequent pattern mining-based cross-social network user identification algorithm that analyzes user-generated data in a personalized manner. The user data weight is optimized by the posterior probability-based information entropy weight allocation algorithm. The user data weight is not only dependent on the information entropy, but is also based on the prior probability. Therefore, the weight of each user data is more reliable. The experimental results show that the user behavior data is analyzed, which improves the precision rate, recall rate, as well as the F1 of user's matching results [12].

In another study, Xiao *et al*. identified groups of fake accounts in online social networks [13] based on ranking learning to classify a whole class of accounts as malicious or legitimate. Statistically, the key features utilized in the model included context texts created by the user such as name, email address, company or university, which contained the frequency of patterns in the category such as all of the emails shared with a common letter/number pattern and a comparison of text frequency across the entire user base such as rareness or non-rareness of all of the names. The applied system categorized clusters of fake accounts to determine whether they were created by the same actor or not, instead of predicting each individual account. The evaluation led to AUC 0.98 and AUC 0.95 on the performed test and out-of-sample test data, respectively, indicating strong performance. The above-mentioned model identified more than 250,000 fake accounts since its establishment [13].

Finally, Liu *et al*. studied an evaluation method for important nodes that affect security in complex networks. They considered the global structure of a network and proposed a network security evaluation index of important nodes that is free of prior knowledge of network organization based on the degree of nodes and nearest neighborhood information. A node information control ability index is proposed according to the structural whole characteristics of nodes. An algorithm ranks the importance of nodes based on the above two indices and the nodes' local propagation ability. The influence of nodes on network security and their own propagation ability are analyzed by experiments through the evaluation indices of network efficiency, network maximum connectivity coefficient, and Kendall coefficient. The experimental results show that the proposed algorithm can improve the accuracy of important node identification; this analysis has applications in monitoring network security [14].

3. Proposed Approach

Most of the methods applied to identify influential people in social networks rely solely on the structure of the networks, and a few seek to narrow and reduce the size of the network and use the content produced by the users for the aforementioned purpose. The present study aims to assess the content produced by people in addition to the structure of the social network to achieve more appropriate results in the direction of information classification. To this aim, the proposed structure is fully introduced. As shown in Figure 1, the utilized data is pre-processed, as well as determining the weight of each person in the network and his/her emotional weight applying the evaluation criteria of node centrality in the network, resulting in knowing the structure of the social network, drawing and displaying its network or graph, and acquiring the ability to analyze the network and available data.

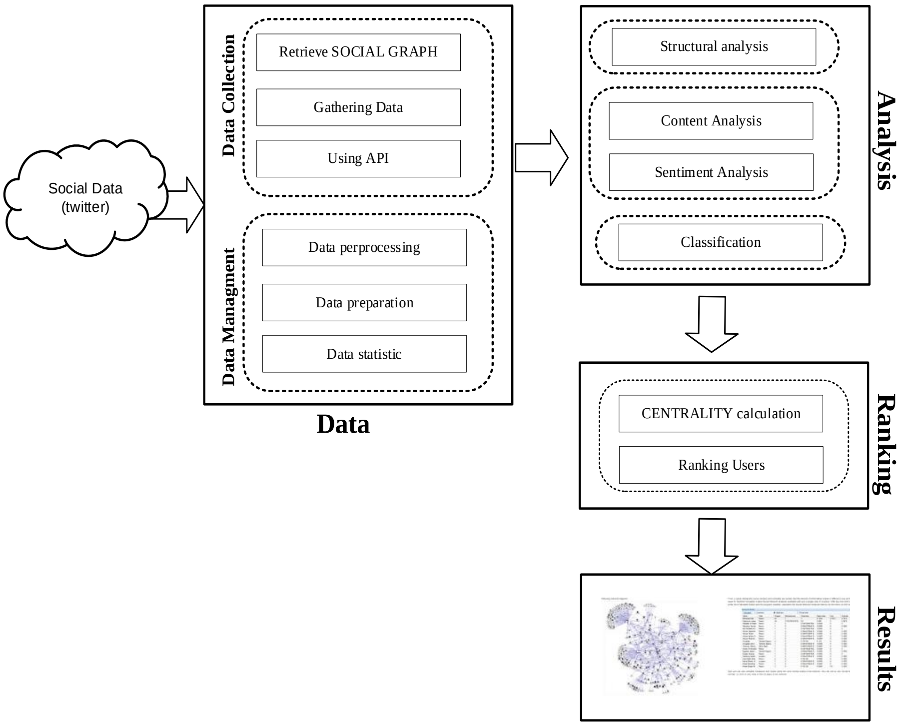

Figure 1. Proposed approach.

By considering emotional scores alongside other factors such as network structure, content dissemination patterns, and user behavior, we can identify nodes that have a higher likelihood of being influential. This comprehensive approach provides a more nuanced understanding of influence dynamics within the network.

The proposed approach involves several steps. At the first, we collect relevant data from the network. Then we clean and pre-process the collected data to remove noise, irrelevant information, and standardize the text. This step often includes tasks such as tokenization, removing stopwords, and handling special characters or emojis. After that, we apply sentiment analysis techniques to determine the emotional polarity of each piece of text associated with a node. This analysis is performed using lexicon-based approaches. The sentiment analysis process assigns sentiment scores or labels to each text and aggregates them to calculate an emotional score for each node. Then we normalize the emotional scores to ensure they are on a consistent scale or range. This step helps in comparing and interpreting emotional scores across different nodes in the network. By analyzing the emotional scores of nodes in the network, we can identify individuals who evoke strong emotional responses from others. Nodes with high emotional scores, particularly positive ones, are more likely to have a significant influence on the network. Their content, opinions, and actions are more likely to be noticed, shared, and followed by other network members.

Moreover, emotional scores can help uncover nodes that have strong emotional connections with other influential nodes. These nodes may act as bridges or connectors within the network, spreading information, ideas or sentiments more effectively. Finally, we use the emotional scores to identify nodes with higher emotional engagement or influence. Nodes with consistently high positive emotional scores or those that evoke strong emotional responses from others are likely to be influential within the network. Each section is fully explained as follows.

Collecting and managing data: Here, data is collected using the APIs of the network (Tweepy API) and entered the data management stage after selecting the studied social network. The collected data is pre-processed before performing analysis operation. Depending on the type of collected data, pre-processing usually involves different steps such as eliminating stop words from data, removing special characters or converting data to root. Finally, the statistical characteristics of the data are extracted, and the data is ready for the analysis stage.

Analyzing and classifying network data and allocating emotional score: In this section, the topology and structure of the network are analyzed and content published by users is evaluated. Then the emotions contained in the content are classified and categorized to rank the users in the next step and identify the important people in different campaigns. Content and sentiment analysis, which is regarded as the most critical part of this section, seeks to classify the sentiments in the data into positive, negative, and neural categories. To this aim, the emotional load of the words is categorized utilizing the dictionary in order to calculate the emotional score in the network nodes since the feelings and tendency of network nodes to other ones play a significant role in addition to the network structure. Thus the emotional score of one person towards another one is calculated by applying a dictionary. Then the collected information is matched with the dictionary, the general feeling of the users is scored, and the general feeling towards the subject under study is determined. The scores in the dictionary may include a range of positive, negative or zero numbers, which represent positive, neural, and negative emotions. Finally, a score is assigned to the node indicating its degree of centrality and used in the final calculations, as well. Thus we used lexicon-based approach for emotion analysis. In this approach, a pre-defined sentiment lexicon or dictionary is used, where words are assigned emotional scores. Each sentiment-bearing word in a text is looked up in the lexicon, and its corresponding emotional score is assigned to that word. The emotional scores can be positive, negative or neutral. The overall emotional score for a sentiment can be calculated by aggregating the scores of individual words in the text. To calculate the emotional score for each person, an approach is used that takes into account their rate of following other people. This method combines information from the emotional scores, network topology, and content published by the users to identify specific campaigns. In this calculation, a dictionary is utilized to assess the emotional score of each person. For instance, let's consider the following example:

<Synset Label=”-1”
Example="The inexperienced nurse did not understand the patient's condition"
Gloss="one who lacks experience and skills"
Sense="beginner, inexperienced, clumsy"
POS="Adjective"
SynID="1"
ID="N"/>

In this example, the Sense tag indicates a negative feeling associated with the tweet. The negative sentiment is determined by words like "inexperienced" and "clumsy," which have a negative charge according to the dictionary. Consequently, the experimental emotional score assigned to this tweet is -1. The POS (Part of Speech) indicates the position of the mentioned words in this tweet, where they are identified as adjectives.

By applying this approach to evaluate the emotional scores of individuals based on their interactions, following patterns, and the sentiment analysis of their content, we can gain insights into their emotional engagement and identify specific campaign within the network.

The sentiment analysis process involves matching the collected information with the provided dictionary to score the general feeling or emotion related to the subject of study. The sentiment tags assigned to the tweets are categorized as positive, negative or neutral. In the lexicon used, the scores assigned to words are +1, 0 or -1, representing positive, neutral, and negative emotions, respectively. The specific range of scores assigned to words may vary depending on the intended application and preferences. For example, consider the following example that is shown in Table 1.

Table 1. A sample of data with sentiment.

ID	Tweet text	Sen.
1221	The ios system is not useful in this situation -	-
1222	I haven’t had an iPhone until now. But Android is really better. iPhone has nothing. I am much more comfortable with Android.	-
1267	You can't say give it all. You have to work with it to know how it is.	+
1310	Who thought Android is better??? Is Android better than Apple???	+
1315	When you say iOS is better than Android	0
1388	Let's be realistic, iOS is better than Android, but because it is closed, we don't use it, but Windows Phone was better than all of them because it was open and had good speed.	+

In this table, the first column represents the user ID, the second column contains the text posted on Twitter, and the third column categorizes the sentiment of each tweet based on the dictionary. The sentiments are classified into three categories: positive, negative or neutral. By analyzing the sentiment categories assigned to tweets and calculating the overall emotional scores for the users based on their interactions and content, we can determine the general feeling towards the subject of study. This information contributes to identifying specific campaigns and gaining insights into users' emotional engagement within the network. The example of users' feelings is used to calculate a score matrix, which is then utilized to determine the total points for each user. To calculate the sum of points for a hypothetical user (let's say user number one), you would need to add up the emotional scores assigned to their sentiments. The sum of the points of the hypothetical user number one can be calculated as $sum = \sum_{u_j} p = -1-1+1+1+0+1 = +1$.

Thus the calculation of emotional scores for nodes in a network involves several steps, after data collection and pre-processing, we apply sentiment analysis techniques to determine the emotional polarity (positive, negative or neutral) of each piece of text associated with a node. This analysis is performed using lexicon-based approaches. After that, we calculate emotional score. Thus once sentiment scores are assigned to the texts, aggregate them to calculate an emotional score for each node. Finally we analyze the emotional scores to identify nodes with higher emotional engagement or influence. Nodes with consistently high positive emotional scores or those that evoke strong emotional responses from others are likely to be influential within the network. As illustrated in Figure 2, the sum of the scores for the hypothetical user number one is calculated as 1, where p equals the acquired emotional score and the calculated emotion based on the sum of the scores equals +1. In addition, the sum of the scores for user number one equals to +1 considering the sum of the scores calculated for the users in advance.

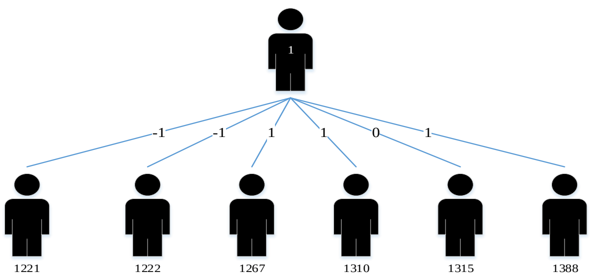

Figure 2. Mapping users' scores into a hypothetical user.

Emotions should be calculated based on the available data for all of the nodes when there are N nodes i.e. the feeling of the first and second node in relation to the next N-1 node, and the like. Therefore, the complexity of the operation equals to $\sum C = N(N-1) \approx N^2 = O(N^2)$, where C indicates the complexity of each calculation. However, the network graph is not considered complete and the complexity of the calculation is normally less than what was indicated. All of the nodes are not connected to each other and do not react to each other's opinions. Finally, the weighing should be assigned to the network graph. In addition, the obtained figures are divided by the total scores achieved and normalized considering that the calculated weight should be standard. Further, the users can be ranked with the highest scores by the information calculated for the entire network. After calculating emotional scores for each user in the network, the result is a numerical score representing the emotional state of each user. In the next step, we incorporate topological information from the network. This refers to the structural characteristics of the network such as the connections and relationships. It provides insights into how users are connected, their positions within the network, and their influence on others. Thus we used this information to spread the emotional score to the other nodes in the networks. It means that the emotional scores and topological information are obtained for each user, and are combined to create a feature set for user classification in the next step. This can be done by concatenating the emotional scores and topological features into a single feature vector for each user. And finally, with the combined feature set, a classification algorithm is applied to predict or classify the users based on their characteristics. This algorithm could be a supervised learning method such as logistic regression, support vector machines or decision trees, which learns patterns from the combined features to differentiate users into specific classes or categories.

Calculating users' rank using the ranking algorithm: In this stage, the centrality of the campaigns should be re-evaluated based on the scores of the previous stage, as well as ranking the users. The influential nodes that play a significant role in campaigns should be found. Such nodes can be called central nodes in campaigns as well because they produce their targeted content by their influence among the users and spread it in line with their objectives. The method of influencing each other by the users is among the issues that should be considered while calculating central nodes. Figure 3 shows the explanation of the aforementioned issue.

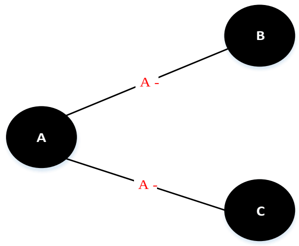

Figure 3. The effect of users on each other.

It is hypothesized that feeling A is negative towards nodes B and C based on the calculated emotional scores (consider -1). Thus node A naturally does not republish or retweet the messages published by nodes B and C due to its negative feelings. However, node A is influenced by nodes B and C or the user and probably republishes their published content when the emotional score of A is neutral to node B and positive to node C. To this aim, the people's influence coefficient should be calculated by considering and taking advantage of the emotional scores of other people due to the possibility of changing people's feelings towards each other or changing their attitudes. Therefore, the influence coefficient is affected by the feeling factor of the clusters around the user. Thus an improved version of the PageRank algorithm is provided to conduct ranking in the best method. The rank of each group in the network is calculated as follows:

$$R(u_i) = \frac{1-d}{N} + d \sum \frac{R(u_j)}{L(u_i)} \times S_{u_j} \times C_u \tag{1}$$

$$S_{u_j} = \frac{sent.u_j \rightarrow u_i}{\sum sent}$$

$$C_u = centrality$$

where R is considered as rank, d is regarded as the damping coefficient, which equals 0.85 by default, L(u) indicates the number of followers of u, and S_{u_j} represents the emotional score of the node, which is divided by the total value in the network for normalization to become a standard value in the range of 0-1, and C_{u_i} represents the centrality score of the node.

4. Experimental Settings

Here, the utilized data is presented, and the evaluation criteria are introduced. Finally, the basic settings for testing and checking the proposed method are indicated.

4.1. Dataset

To use Twitter as a data source in this paper, we have done the following steps: at the first, we obtain access to the Twitter API (Application Programming Interface), which allows programmatic access to the Twitter's data. To receive the flow of users' information, Tweepy library based on the python programming language was applied to connect to twitter. This usually requires creating a developer account, creating an application, and obtaining API keys and access tokens. Then we utilize the Twitter API to collect the relevant data for our work. This included retrieving user profiles, posts, follower/following relationships, and other relevant metadata. In the next step, we clean and pre-process the collected data to ensure data quality and prepare it for analysis. This may involve removing duplicates, handling missing values, normalizing text, and removing irrelevant or sensitive information. After that, we construct the social network by representing users as nodes and their relationships as edges between nodes based on the collected data. This step establishes the structure of the social network. As demonstrated in Figure 4, the data include the information of a network containing 500 nodes with 12,500, edges, and an average degree of 50.

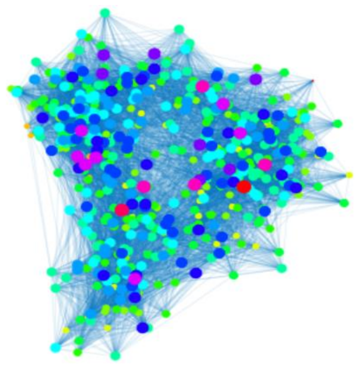

Figure 4. Twitter communication network used.

Table 2 indicates the measures of the collected network. The rows show the number of nodes, number of edges, average degree, average degree centrality, average closeness centrality, average eigenvector centrality, and average PageRank centrality, respectively.

Table 2. Characteristics of the Twitter social network used in this research work.

Measures	Number
number of nodes	500
number of edges	12500
average degree	50
average degree centrality	50
average closeness centrality	0.0011
average eigenvector centrality	0.0020
average PageRank centrality	0.0020

4.2. Evaluation measures

Based on the evaluation criteria, all of the nodes are included in the calculations using the proposed algorithm according to their position in the network topology and obtain a standardized score. Any node or user who acquires more scores obtains a higher rank [18].

The method of disseminating the information by the nodes should be investigated after they are classified by the order of significance utilizing the proposed algorithm. Then the hypothetical information is disseminated in the network and its process is studied. To this aim, well-known methods such as susceptible-inferred (SI) and susceptible-inferred-susceptible (SIS) are applied. The SIS model is mostly used to simulate the spread of disease in the network. Based on the SIS model, a person recovers with the probability of mu and is considered as susceptible to the disease again after infecting with the probability of beta. All of the people are in S state first. One person is randomly infected when the disease starts and the disease spreads in the network with such a model. Not all of the people in the network are infected since there is a rate at which people recover.

Here, the SI model was utilized to analyze the method of disseminating information in the network. Every person can receive information in the SI model potentially. All of the people are in S state at first. Then one person is selected among the top people in terms of scores, by whom information is disseminated in the network. In the next step, the speed of information dissemination in the network is assessed by tracking the data and its flow in the network. High speed of information dissemination means that the selected node is regarded as a more appropriate one that can disseminate the information faster.

5. Experimental Results

In this section, the evaluations performed on the proposed algorithm are compared with PageRank [15] and the Betweenness algorithm [16]. There are a lot of state of the art algorithms for centrality, but the using of them depend on several factors including the specific characteristics of the network, the research question or application, and the computational resources available. Different algorithms may excel in different scenarios. Here are some commonly used algorithms for centrality measures such as PageRank, Betweenness, Eigenvector, Exclusivity, and Hyperlink-Induced Topic Search [19] [20]. While new centrality measures continue to emerge, PageRank and Betweenness centrality retain their relevance and usefulness due to their effectiveness, broad applicability, interpretability, robustness, computational efficiency, and their role as established benchmarks in the field of network analysis. Moreover, there are a lot of research papers that use these algorithms, especially PageRank or Betweenness, for their evaluation such as [21-23].

To this aim, the SI model is applied to the three algorithms indicated in the first, second, and third nodes, and the results are discussed in comparison with time. Table 3 and Figure 5 show the result of applying the SI model to the indicated algorithms in the first selected node.

Table 3. The results of the SI model - 1st node.

Method/Time	0	40	80	120
Betweenness	25	85	106	150
PageRank	25	190	290	350
Proposed	25	395	450	497

As illustrated, the proposed algorithm shows a better performance than the basic ones. For example, 150 users or nodes receive the disseminated information at time 120 in the Betweenness algorithm. The aforementioned number equals to 350 and 497 for PageRank and the proposed algorithm, respectively. The higher number means that the information reaches more nodes in the network, leading to a better performance of the algorithm. Also the higher values suggest that the proposed algorithm achieves a broader dissemination of information among more nodes in the network, indicating a better performance in terms of spreading information. In addition, the higher number indicates that the graph of the algorithm is placed at a higher level.

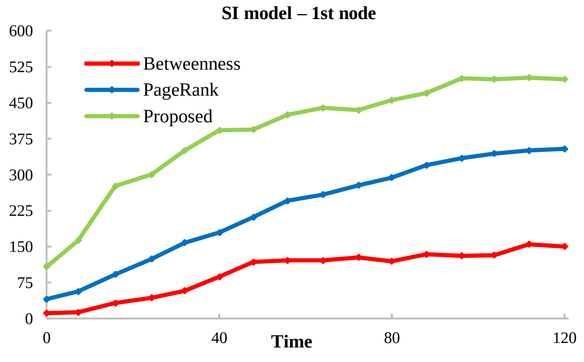

Figure 5. Simulation results of the SI model - 1st node.

As displayed, the X and Y indicate the time consumption of the methods and the number of nodes, respectively. Figure 5 is related to the first node from the ranking results according to the proposed algorithm. As shown, the information disseminated in the network applying the proposed algorithm presents better results than the other two algorithms, along with more selected nodes. The selection and dissemination speed of the proposed algorithm is faster than the other two ones as the number of its nodes is more than that of other

methods. The statement highlights several aspects in which the proposed algorithm outperforms the basic algorithms: Number of selected nodes: The proposed algorithm shows a higher number of selected nodes compared to the basic algorithms. This indicates that the proposed algorithm is able to identify and target more nodes in the network for information dissemination. Selection speed: The proposed algorithm is faster in selecting nodes for dissemination compared to the basic algorithms. This suggests that the proposed algorithm is more efficient in identifying and prioritizing nodes to spread information, and finally, the dissemination over time: The proposed algorithm demonstrates a better dissemination of information over time compared to the basic algorithms. This means that the information spreads more effectively and reaches a larger portion of the network as time progresses when using the proposed algorithm.
Figure 6 and Table 4 indicate the selection of the second node as the source and start of information dissemination.

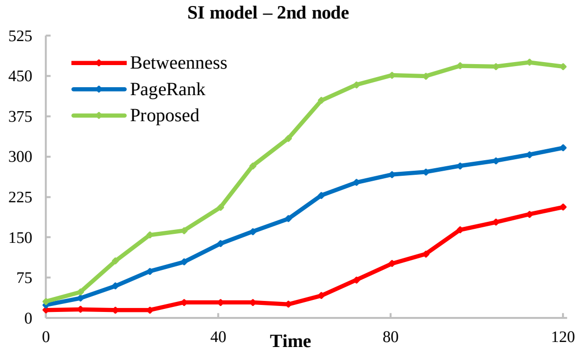

Figure 6. Simulation results of the SI model - 2nd node.

Figure 6 is related to the second node obtained from the ranking results according to the proposed algorithm. As observed, the efficiency of the three algorithms is close to each other. However, the proposed algorithm shows better performance after some time. Such superiority can be observed in the number of selected nodes, selection speed, and its dissemination in time compared to the two basic algorithms.

Table 4. The results of the SI model–2nd node.

Method/Time	0	40	80	120
Betweenness	25	30	112	206
PageRank	27	135	260	325
Proposed	30	210	450	462

As displayed in Figure 7, the third node is demonstrated as the source for information dissemination, from which results similar to the above-mentioned graph are achieved.

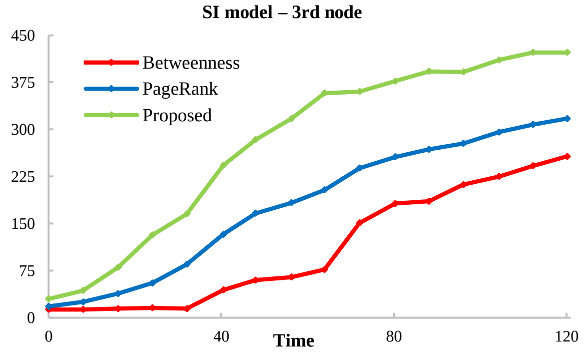

Figure 7. Simulation results of the SI model – 3rd node.

As represented in Table 5, the results are presented for the third node as a source for information dissemination and the proposed algorithm demonstrates a better performance than the basic ones. For instance, 257 users or nodes receive the disseminated information at time 120 in the Betweenness algorithm. The above-mentioned number equals to 317 and 425 for PageRank and the proposed algorithm, respectively, indicating a big difference between the proposed algorithm and the basic ones.

Table 5. The results of the SI model - 3rd node.

Method/Time	0	40	80	120
Betweenness	20	44	188	257
PageRank	23	122	258	317
Proposed	30	248	375	425

As observed, the proposed algorithm in all three cases shows high efficiency. However, according to the prediction of the algorithm, in the third node, the nodes affected by the selected node are less than the similar cases in the first two cases due to the lower score obtained by the third node compared to the first two nodes. In other words, the low-score algorithm affects the network less, and fails to form the core of a campaign. The opposite is true, as well.
Generally, the proposed algorithm has performed better than other methods due to the use of topological information from the network and the emotional information of users in the proposed algorithm. Only the physical and topological information of the network is utilized in the two algorithms of PageRank and Betweenness. However, the emotional score was applied in the proposed structure by changing the PageRank algorithm. Finally, using information dissemination methods in the network such as SI proved that the effectiveness of the proposed algorithm in the presented structure is more than the basic ones.

6. Conclusion

Social networks cannot be analyzed easily due to their increasing expansion and different cultural, human, and emotional conditions, which should be addressed. Accordingly, the behavior of nodes in social networks is affected by the type and topology of the network. In addition, the quality and feeling of the nodes of the network towards each other play a significant role in their analysis.

The present study aims to evaluate the social network nodes utilizing qualitative information about the content of such networks. To this aim, the nodes in the network are classified, as well as demonstrating that the influential nodes in the network can be identified better and faster. The parameters applied here to examine the content of networks include the emotional information of users to each other and their centrality in the network. This study investigated the Twitter network. To analyze the effect, a network with 500 nodes was collected from Twitter, and its structure was reviewed after data pre-processing. The nodes with a high degree of rank centrality were considered as the most central ones. To this aim, the required information was collected using the Twitter API and the emotional scores of the nodes towards each other were calculated, resulting in assigning a score to each user and achieving emotional centrality in the network. Some information was disseminated in the network and its dissemination behavior was monitored by classifying nodes or network users. Based on the results, the proposed algorithm achieves better results compared to the PageRank and Betweenness algorithms, and the information broadcast by such node is disseminated in the network with greater speed and volume since the above-mentioned algorithms only utilize the network topology as a criterion for centrality and hypothesize that the node is regarded as central only in the structure and topology of the network, while in the proposed algorithm, the central nodes are determined better and more accurately by integrating the structural–topological features of the network and the people’s behavior in the network. The aforementioned results were confirmed for all three cases of high-ranking nodes and the proposed algorithm demonstrated better results. Thus considering the structure of the network leads to an incomplete analysis of the social network. Adding qualitative information in network analysis can help identify appropriate and effective nodes better. Testing the results indicates the speed of practical influence of nodes in social networks in addition to confirming the issue.

Other structural metrics of the network can be analyzed, as well. There are a large number of metrics that can be effective in network analysis. Such metrics can include microcosm phenomena in social networks. In addition, the dynamic nature of social networks can be considered more, and the topological changes of social networks can be studied. Future studies can be conducted on the changes in people’s opinions or network.

References

[1] N. Johnson, B. Turnbull, and M. Reisslein, "Social media influence, trust, and conflict: An interview-based study of leadership perceptions", *Technology in Society*, vol. 68, no. 101836, 2022.

[2] U. N. Dulhare, K. Ahmad, and K.A.B. Ahmad, *Machine Learning and Big Data: Concepts, Algorithms, Tools and Applications*, John Wiley & sons, 2020.

[3] M. Bahutair, A. Zaher, and K. Ibrahim, "NodeRank: Finding influential nodes in social networks based on interests", *The Journal of Supercomputing*, vol. 78, no. 2, pp. 2098-2124, 2022.

[4] A. K. Jain, S. R. Sahoo, and J. Kaubiyal, "Online social networks security and privacy: comprehensive review and analysis", *Complex & Intelligent Systems*, vol. 7, no. 5, pp. 2157-2177, 2021.

[5] S. Hajjam and H. Badir, *Community detection in social networks In Principles of Social Networking*, Springer, 2022.

[6] U. Can and B. Alatas, "A new direction in social network analysis: Online social network analysis problems and applications", *Physica A: Statistical Mechanics and its Applications*, vol. 535, pp. 122372, 2019.

[7] J. C. Bansal, M. Paprzycki, M. Bianchini, and S. Das, *Computationally Intelligent Systems and Their Applications*, Springer, 2021.

[8] S. Rathore, P. K. Sharma, V, Loia, Y. Jeong, and J. H. Park, "Social network security: Issues, challenges, threats, and solutions", *Information sciences*, vol. 421, pp. 43-69, 2017.

[9] J. Wu, J. Shen, B. Zhou, X. Zhang, and B. Huang, "General link prediction with influential node identification", *Physica A: Statistical Mechanics and its Applications*, vol. 523, pp. 996-1007, 2019.

[10] A. Zareie, A. Sheikhahmadi, M. Jalili, and M. S. K. Fasaei, "Finding influential nodes in social networks based on neighborhood correlation coefficient", *Knowledge-based systems*, vol. 194, pp. 105580, 2020.

[11] S. Raamakirtinan and L. M. J Livingston, "Identifying influential users in Facebook-a sentiment-based approach", *Indian Journal of Science and technology*, vol. 9, no. 10, pp. 1-9, 2016.

[12] K. Deng, L. Xing, L. Zheng, H. Wu, P. Xie, and F. Gao, "A user identification algorithm based on user behavior analysis in social networks", *IEEE Access*, vol. 7, pp. 47114-47123, 2019.

[13] C. Xiao, D. M. Freeman, and T. Hwa, "Detecting clusters of fake accounts in online social networks", in

Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security, pp. 91-101, 2015.

[14] Y. Liu, J. Wang, H. He, G. Huang, and W. Shi, "Identifying important nodes affecting network security in complex networks", *International Journal of Distributed Sensor Networks*, vol. 17, no. 2, pp. 155014772199928 ,2021.

[15] J. Heidemann, M. Klier, and F. Probst, "Identifying key users in online social networks: a pagerank based approach", in *Proceedings of the IEEE International Conference on Advanced Information Networking and Applications Workshops*, pp. 20-13, 2010.

[16] N. Kourtellis, T. Alahakoon, R. Simha, A. Iamnitchi, and R. Tripathi, "Identifying high betweenness centrality nodes in large social networks", *Social Network Analysis and Mining*, vol. 3, no. 4, pp. 899-914, 2013.

[17] D. Sivaganesan, "Novel influence maximization algorithm for social network behavior management", *Journal of ISMAC*, vol. 3, no. 01, pp. 60-68, 2021.

[18] S. Sovan, V. K. Dubey, and B. Sarkar, "Measure of influences in social networks", *Applied Soft Computing*, vol. 99, pp. 106858, 2021.

[19] Q. Liao and S. Sinthupinyo, "Exclusivity Centrality: A novel centrality measure for finding behavioral influencers on Twitter", in *Proceedings of 10th International Conference on Electronics, Computers and Artificial Intelligence*, pp. 1-6, 2018.

[20] P. D. Meo, M. Levene, F. Messina, and A. Provetti, "A general centrality framework based on node navigability", *IEEE Trans. Knowledge and Data Engineering*, vol. 32, no. 11, pp. 2088-2100, 2020.

[21] K. Hajarathaiah, M. K. Enduri, S. Anamalamudi, T. S. Rddy, and S. Tokala, "Computing influential nodes using the nearest neighborhood trust value and pagerank in complex networks", *Entropy*, vol. 24, no. 5, pp. 704, 2022.

[22] A. P. Ogan, P. M. LaCasse, and B. J. Lunday, "Socialnetwork analysis of Twitter interactions: a directed multilayer network approach", *Social Network Analysis and Mining*, vol. 13, no.1, pp. 65, 2023.

[23] W. Zhang, X. Guo, W. Wang, Q. Tian, L. Pan, and P. Jiao, "Role-based network embedding via structural features reconstruction with degree-regularized constraint", *Knowledge-Based Systems*, vol. 218, pp.106872, 2021.

شناسایی گره‌های تاثیرگذار در شبکه‌های اجتماعی مبتنی بر تحلیل پروفایل افراد

زینب پشتیبان، الهام قنبری* و محمدرضا جهانگیر

گروه کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه آزاد اسلامی، واحد یادگار امام خمینی (ره) شهر ری، تهران، ایران.

ارسال ۲۰۲۳/۰۶/۳۰؛ بازنگری ۲۰۲۳/۰۸/۰۸؛ پذیرش ۲۰۲۳/۰۹/۲۴

چکیده:

تحلیل نفوذ افراد و گره‌ها در شبکه‌های اجتماعی توجه بسیاری را به خود جلب کرده است .شبکه‌های اجتماعی با وجود گروه‌ها، انجمن‌ها و افراد علاقه‌مند به یک موضوع یا موضوع خاص معنا پیدا می‌کنند و افراد تمایلات نظری و عملی خود را در چنین مکان‌هایی نشان می‌دهند .در پژوهش‌های صورت گرفته، گره‌های تاثیرگذار اغلب بر اساس اطلاعات مربوط به ساختار شبکه اجتماعی شناسایی می‌شوند و کمتر به اطلاعات منتشر شده توسط کاربر شبکه اجتماعی توجه می شود. در این مقاله هدف بر این است که علاوه بر استفاده از اطلاعات کاربر در شبکه اجتماعی، از اطلاعات ساختاری شبکه نیز در جهت شناسایی کاربران تأثیرگذار استفاده شود .بدین منظور در ابتدا احساسات کاربر استخراج شده و بر اساس یک دیکشنری احساسی، به هر کاربر یک نمره احساسی یا عاطفی نسبت داده شود و وزن آن در شبکه با استفاده از معیارهای مرکزیت تعیین شود. شبکه اجتماعی مورد استفاده در این مقاله شبکه توئیتر است، لذا پس از جمع آوری و پردازش داده‌ها ساختار شبکه اجتماعی مشخص و گراف آن رسم می‌شود و قابلیت تحلیل شبکه و داده‌های موجود استخراج شده و بر اساس الگوریتم پیشنهادی کاربران و گرههای تاثیرگزار شناسیایی می‌شود. نتایج ارزیابی نشان می‌دهد که گره‌های شناسایی‌شده توسط الگوریتم پیشنهادی کیفیت بالایی داشته و سرعت انتشار اطلاعات شبیه‌سازی‌شده از آن‌ها بالاتر از سایر الگوریتم‌های موجود است.

کلمات کلیدی: شبکه های اجتماعی، تحلیل پروفایل، بازیابی احساسات، شناسایی افراد.