Original Article

A model to predict communications in dynamic social networks

Ahmad Ghadamkheir [1], Seyed Alireza Derakhshan [1*], Ashraf Shahmansoury [1]

[1] Department of Information Technology Management, South Tehran Branch, Islamic Azad University, Tehran, Iran.

***Corresponding author and reprints: Seyed Alireza Derakhshan,** Visiting Professor**,** Department of Information Technology Management, South Tehran Branch, Islamic Azad University, Tehran, Iran.

Email: ard1331@gmail.com

Received: 14 Oct 2022 **Accepted:** 23 Feb 2023 **Published:** 04 Mar 2023

Abstract

Background: social networks are dynamic due to continuous increases in their members, communications, and links, while these links may be lost. This study was conducted with the aim of investigating the link and communication between social network users using the centrality criterion and decision tree.

Methods: After checking the nodes in the network for each pair of unrelated nodes, some common nodes in the proximity list of these two groups were extracted as common neighbors. Analysis was performed based on common neighbors, association prediction process, and weighted common neighbors. Prediction accuracy improved. Centrality criteria were used to determine the weight of each group. New Big Data techniques were used to calculate centrality measures and store them as features of common neighbors. Personal characteristics of users were added to build complete data for training a data mining model. After modeling, the decision tree model was used to predict communication.

Results: There was an increase in sensitivity, which indicated model power in identifying positive categories (i.e., communications) when users' characteristics were used. It means that the model could identify potential latent communications. It can be stated that users are more willing to make a relationship with users similar to them through common neighbors. Personal characteristics of users and centrality were effective in method efficiency, while removal of these properties in the learning process of the decision tree model caused a reduction in efficiency criteria.

Conclusion: Prediction of latent communications through social networks was promising. Better results can be obtained from further studies.

Keywords: Big Data; Communication; Decision Trees; Forecasting; Social Networking.

Cite this article as: Ghadamkheir A, Derakhshan SA, Shahmansoury A. A model to predict communications in dynamic social networks. *Soc Determinants Health*. 2023;9(1):1-13. DOI: http://dx.doi.org/10.22037/sdh.v9i1.39715

Introduction

A social network is a web-based service through which users connect with their friends, families, colleagues, and others on their pages. Users can join new communities and experience new social activities through these networks. Social networks include abundant knowledge about communications between people and a set of nodes that are connected through edges. These networks may be presented as different web pages, newspapers,

neighbors, and organizations (1). Users exchange a large volume of information through social, political, sports, and economic applications. Because social networks are dynamic networks, any change in the number of members and their communications may reduce the accuracy of prediction-related algorithms (2). Graph theory is one of the methods used for forecasting and consists of nodes and edges. Graph theory comprised some features, including centrality, the most important criterion that considers the significance of each node in the network structure. Centrality identifies important nodes in the social network, so predicting these nodes leads to better but not sufficient results in the whole network. In communication networks, the node serves as a connection, branch point, or endpoint of the connection. A node is a physical network that is an active electronic device connected to the network and can send, receive, or resend the data on a communicational channel (3). In the form of clustering, the effect of network structure on communication prediction efficiency indicates that clustering is effective in increasing prediction accuracy and efficiency. However, this technique does not work in solitude networks with poor connections. The network structure has been integrated with communities' information to determine the behavior and interests of users and to predict communication in the social network of Tweeter. Users' characteristics (e.g., education level, book title, keywords, and age) are used to predict communications in the two-part network (4). It is essential to use data to model users' behavior in social networks and the relationship between their behaviors and social phenomena (5). The challenge of the connection between increasing users of social networks is necessary (6). The extant study was conducted on the data of social networks by using centrality criterion and decision tree to examine the link and communication between users of these substantial networks.

Methods

The proposed method of this study comprises the following steps: In the first phase, the available raw data were processed initially, and the required data, common neighbors between pairwise unconnected nodes, were extracted from the data. The input data of this step includes the adjacency list that indicates the graph structure of a social network. The output of this step is used as input for the next step. The block diagram of the proposed method has been illustrated in Figure 1.

In the second phase, big data analysis and MapReduce processing were used to measure centrality criteria for input nodes of this phase in social networks. The centrality criteria indicate the importance of nodes, so they were stored as nodes' features. The stored data served as input data for the next step. Then, users' calculated features and personal characteristics were used to build a training and evaluation dataset. In the last phase, training data were finally used to design the decision tree model to predict communications in the network.

Data Preprocessing

According to Figure 1, the applicable data in this step depicts a social network graph illustrated as an adjacency list. In this section, the adjacency list of all nodes of the graph was explored, and all common neighbors (i.e., nodes existing in adjacency list per pairwise unconnected nodes) were extracted per pairwise nodes that are not adjacent (neighbor) to each other. The common neighbors were taken as the output of this phase and transferred to the next step.

Calculation of Centrality Criteria

One of the big data techniques called Hadoop with MapReduce programming was used to measure centrality criteria. Hadoop is an open-code software framework that allows distributed big data processing on some clusters from servers.

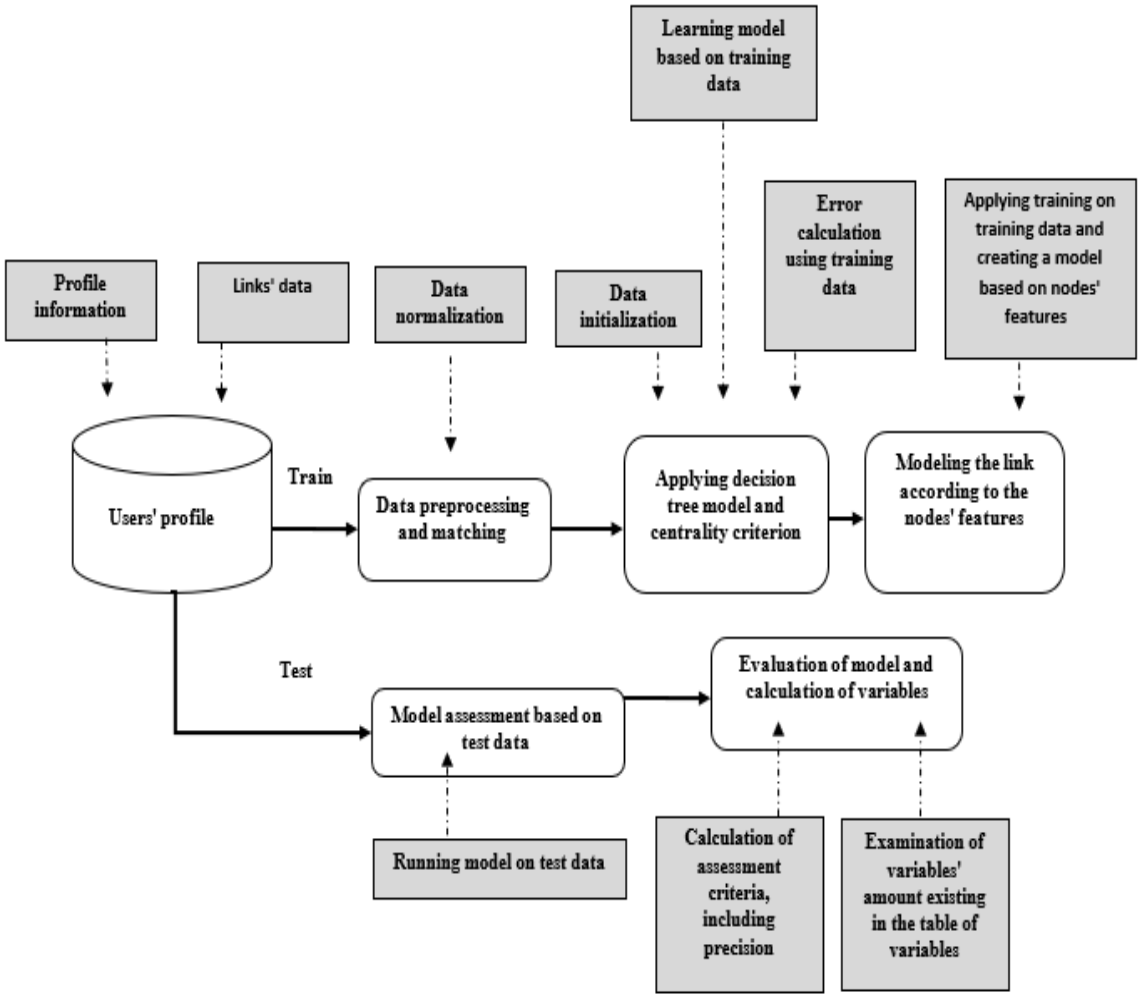

Figure 1. Block scheme of the proposed model to predict communications in social network

This framework based on Java Language is designed to distribute processes on thousands of machines with high fault tolerance. This framework can also be used on the local machine. MapReduce is a programming model that expresses a large distributed computation as a sequence of distributed operations on a key/value pairs dataset. Figure 2 depicts the MapReduce process. MapReduce computation has two phases: a map phase and a reduce phase. The input to the computation is a dataset of key-value pairs. In Figure 2, Map and Reduce tasks have been shown as circles, and key-value pairs have been depicted in colored rectangles. In the map phase, the framework splits the input data into many fragments and assigns each fragment to a map task. Each map task consumes key/value pairs from its assigned fragment and produces a set of intermediate key/value pairs (K':V'). Following the map phase, the framework sorts the intermediate dataset by key and produces a set of (K':V'*) tuples so that all the values associated with a particular key appear together. This step, called Shuffle, is

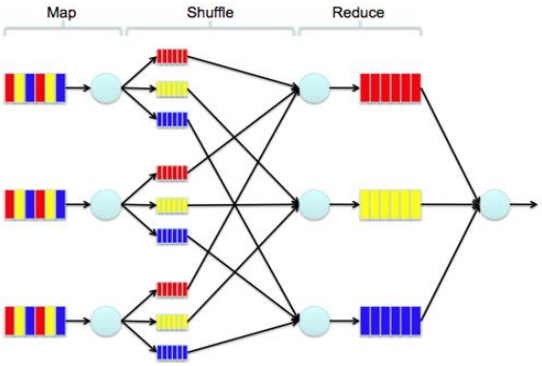

Figure 2. MapReduce Process

shown in Figure 2. As seen in the figure, all values appear together with equal keys. In other words, all values with the same color have been sorted together per a particular key. The framework also partitions the set of tuples into several fragments equal to reduced tasks.

In the reduce phase, each reduces task consumes the fragment of (K': V'*) tuples assigned to it and transmutes the tuple into an output key/value pair (K, V). Finally, the final output will be a set of key/value pairs (K,V).

According to MapReduce Process, calculation of centrality criteria has been described based on this programming model.

Calculation of Degree Centrality Criterion based on MapReduce Processing

The pseudocode shown in Figure 3 was used to calculate degree centrality according to MapReduce processing. The graph adjacency list is the data required to calculate degree centrality. Therefore, a node with its adjacency list is taken as an input key-value pair. In the map phase, input key/value nodes with their adjacency list were read, and intermediate key/values were produced. As we know, all values are sent to the reducer with the equal key in this processing paradigm. Therefore, an input key that is a node in the network can be considered an intermediate key. Then, it is possible to take a numerical value "1" as an intermediate value per node existing in its adjacency list and send the produced intermediate key-value pair to the output. In the Shuffle phase, the sorted intermediate key-values and all values related to a unique key go to a reducer. In reduce phase, a list of numbers 1 exists as a values list per entered key. Therefore, the values "1" can be counted to compute the final value of degree centrality for a node. Then, the output key in reduce phase, i.e., the key entered into it, is one node in the network, while the output value in this phase is in the sum of values "1" as the final value of degree centrality for the considered node.

Calculation of Closeness Centrality Criterion based on MapReduce Processing

The closeness centrality criterion indicates the average distance between a source node and other nodes existing in the graph. In other words, this criterion expresses how a node is close to other nodes in the graph. The more precise definition of this criterion indicates information diffusion speed from one node to another one. Therefore, this criterion implies the importance of a node.

```
1. Class Mapper
2.      Method Map (nodeID, AdjacencyList)
3.      For each nodeID in AdjacencyList
4.              Output (nodeID, 1)
-------------------------------------------------
1. Class Reducer
2.      Method Reduce (nodeID, ListOfValues [1, 1, 1, ..... ])
3.              sum = 0
4.              For each value in ListOfValue
5.                      sum = sum + value
6.              Output (nodeID, sum)
```

Figure 3. The pseudocode used for calculation of degree centrality criterion

```
- Input: Common Neighbors with their Adjacency List, Graph of Social Network
- Output: Closeness Centrality of each Common Neighbor
---
1.    For each node in CommonNeighbors
2.        allShortestDistances = BFS (node, SocialNetworkGraph)
3.        closenessCentrality = 1/Mean(allShortestDistances)
```

Figure 4. The pseudocode used for calculation of closeness centrality criterion

The higher the closeness centrality criterion, the closer the node in a graph to other nodes. In this case, information will spread throughout the network more rapidly. All of the shortest paths from a source node to all other graph nodes must be calculated for this criterion. Because there are numerous users in social networks, the closeness centrality criterion cannot be computed simply using classic methods. Therefore, the MapReduce model was used to compute this criterion based on the pseudocode presented in Figure 4.

BFS (Breadth-First Search), the algorithm was used to compute the shortest path from a source node to all nodes existing in the network. Hence, this algorithm was the main core of closeness centrality criterion and was implemented based on the MapReduce paradigm.

BFS Algorithm based on the MapReduce Paradigm

In addition to the adjacency list, two-color and distance features were saved from implementing a distributed DFS algorithm for each node existing in the network graph. The distance indicated the distance between one node and a source node, while color indicates its current situation. When one node is white, it has not been explored while the BFS algorithm is running. If the node is observed, it will turn gray. After the algorithm observed all children of that node, the parent node turned black. It is worth noting that only the source node is in gray with zero distance in the first step.

```
1. Class Mapper
2.      Method Map (nodeID, DataStructure)
3.          d = DataStructure.Distance
4.          For each nID in DataStructure.AdjacencyList
5.              Output (nID, d + 1)
6.          Output (nodeID, DataStructure)
---
1. Class Reducer
2.      Method Reducer (nodeID, ListOfValues [d, d, d, d, … ])
3.          d_min = ∞
4.          M = 0
5.          For each d in ListOfValue
6.              If isStructure(d)
7.                  M = d
8.              Else if d < d_min
9.                  d_min = d
10.         M.Distance = d
11.         Output (nodeID , M)
```

Figure 5. The pseudocode of the BFS algorithm

Input data of map phase included adjacency list related to the network graph along with the mentioned extra characteristics, each node of graph (as the key) and a data structure, including node adjacency list, color, and distance (as value) indicate the input key/values of the mapper. As seen in Figure 5, in the phase map, an intermediate key-value is built per node located in the adjacency list of the input node. Each node in the adjacency list of input nodes was created as an intermediate key. For the intermediate value of data structure comprising as adjacency list, the current distance plus "1" was chosen as the distance, and gray color was considered the color of the node in the intermediate key. This key/value pair was written in output. In addition to sending nodes existing in the adjacency list, the main node plus its value is written as a key in output. In this case, after the MapReduce process was run, the input graph structure appeared in the output completely. Because implementation of MapReduce-based BFS algorithm is iterative processing and the full graph structure is required per iteration, the reason for the iterative pattern of this algorithm has been explained herein. In the next step, the shuffle step is implemented to sort and send all values per an equal key to an equal reducer. In reduce phase, input keys are the same graph nodes, and the list of values is the list of data structures explained before. In this phase, the input key that is a node of the graph was taken as the output key to creating output key-value. Then, all distances in the data structure were examined, and the shortest distance to the source node was selected and written as the distance in the data structure. Then this node turned black. The considerable aspect of the calculation of this algorithm is that all graph nodes are not explored within one implementation of MapReduce processing, so that some nodes may remain gray. Therefore, this processing was done iteratively until there was no gray node in the graph.

BFS algorithm was implemented per neighbor to calculate closeness centrality criterion for those common neighbors extracted in the first step then entered into this step. The output of the BFS algorithm included all shortest distances from considered nodes to all nodes of the network. Finally, the inverse mean of calculated distance was the final value of closeness centrality criterion for that common neighbor.

Preparing Data to Build Model

The extant study assumed that personal characteristics of users in their social network profiles, including their education, job status, and living place, could be used to improve the precession of communication prediction. In this case, users with more subscriptions and similarities make more relationships. Therefore, the personal characteristics of common neighbors were collected using the computed centrality criteria in the previous step, and then a larger dataset was formed.

Training and building a data mining model, such as a decision tree learning method with the observer, require training data that perfectly defines inputs and outputs. In other words, training data belong to a certain category before the training process begins. The case of communication prediction is a kind of categorization. The extant study aimed at predicting the label or category to which a new observed sample belongs. In this case, the data were divided into two separate categories, and the data of common neighbors were used to determine data labels. Before common neighbors were extracted, the data were divided into two historical categories. For instance, 2003-2010 were assigned to the first category, while 2001-2013 were placed n the second category. The first and second phases of the proposed method were implemented using data on first-time intervals in the next step.

All communications that did not exist in the first time interval in the network were

examined in the second interval. Label 1 was assigned to the data sample if the communication was created, 0, otherwise.

Prediction of Communication

After building training data, the last and main phase of the algorithm was proposed. In this strep of algorithm, the prediction process was done using a data mining model focusing on the decision tree model. Therefore, the decision tree model was created by dividing the data prepared in the previous step into training and test data sets.

Results

A dataset extracted from DBLP social network was used to evaluate the proposed technique of the present study (7). However, an equal graph of these data must be created before dividing data. Therefore, every user was taken as a node in the graph, and then the created graph was used as the tested dataset. Reports of the characteristics of the dataset showed that the number of users was11590, the number of communications was 71150, Number of common neighbors was 12256.

Results of Empirical Experiments

The Effect of Personal Characteristics of Users on the Algorithm Efficiency

Personal characteristics of users can be used effectively in predicting latent communications in social networks. Personal characteristics comprise some information, such as education level, job status, and living place of users. Therefore, two tests were done to find the application results and non-application of these characteristics.

Test 1: Prediction of Latent Communications by Using Personal Characteristics of Users

In this test, the proposed method implemented both centrality criteria of nodes and users' characteristics used in the decision tree model. The results of communication prediction by using users' characteristics algorithm, Proposed method (Accuracy:0.95, Sensitivity:0.97, Precision:0.97, F-measures:0.97) was achieved.

Test 2: Prediction of Latent Communications without Using Personal Characteristics of Users

In this test, a decision tree was trained and evaluated after preparing the required datasets without using users' characteristics only by using the centrality features of nodes. The results have been reported in Table 1.

Table 1. Results of communication prediction without using users' characteristics

Algorithm	Accuracy	Sensitivity	Precision	F-measures
Proposed method	0.87	0.95	0.9	0.93
Proposed method	0.84	0.9	0.91	0.9
Proposed method	0.84	0.9	0.91	0.9
Proposed method	0.89	0.92	0.94	0.93
Proposed method	0.85	0.91	0.9	0.9

According to Table 1, the use of personal characteristics had a significant effect on the efficiency of the proposed method and its results. Accordingly, there was a reduction in all calculated evaluation criteria in these two tests when personal characteristics of users were not applied model training. Such reduction in the efficiency was at least about 2% in the Sensitivity measure and a maximum of 8% in the Accuracy metric. If we know the personal characteristics of common neighbors between two disconnected nodes and their situation in the network, prediction performance will be improved.

The Effect of Centrality Criteria as Weights Assigned to Nodes

Following tests were implemented to expand the precision and efficiency of the proposed method in predicting latent communications in the network and to know to what extent this assumption was correct. The obtained results have also been reported.

Test 1: Prediction of Latent Communications by using Centrality Criteria of Nodes

In this test, the proposed method was implemented considering all steps described above, i.e., both personal characteristics of users and two centrality criteria of nodes were used to create a decision tree model. Table 1 reports the relevant results.

Test 2: Prediction of Latent Communications without using Centrality Criteria of Nodes

In this test, a decision tree was trained and evaluated after only preparing the required datasets using users' characteristics. The results have been reported in Table 1.

According to results reported in Table 1, removing centrality features led to a considerable drop of a minimum of 4% and a maximum of 9% in Precision and Accuracy criteria, respectively, in all evaluation criteria. It is worth noting that the effect of the features extracted from the network structure was higher than the personal characteristics of users because the efficiency loss of the model was more considerable than two tests conducted in the first section. The reason may stem from some users that are not honest in providing their personal information. In this case, there will be more noise in the model's training data that, in turn, will reduce the precision and accuracy of the model. On the contrary, there will be no noise if the features extracted from the network structure (e.g., centrality criteria of each node) are used because these features are adopted from the network structure. Accordingly, the primary assumption of this study was confirmed, i.e., the knowing contribution of each common neighbor increases the performance power of the proposed method. Two other tests were also implemented for further assessments. In these tests, only one centrality criterion was considered for common neighbors.

Test 3: Prediction of Latent Communications by using Degree Centrality Criterion

In this test, the personal characteristics of users and the degree centrality criterion were considered for nodes within the building decision tree. The obtained results have been reported in Table 1.

Test 4: Prediction of Latent Communications by using Closeness Centrality Criterion

In this test, the personal characteristics of users and closeness centrality criterion were considered for nodes within the building decision tree. The obtained results have been reported in Table 1.

According to Table 1, degree and closeness centrality criteria had a significant effect on the performance of the decision tree in predicting latent communication in the network. As seen in test 3, the assignment of a contribution about degree centrality to each common neighbor led to improvement in all four criteria with a minimum 2% and maximum 5% rise in Precision and Accuracy criteria, respectively. However, the performance improvement of the model was less than the case of using both degree and closeness centrality in the building decision tree process.

Moreover, the test 4 results reported in Table 1 indicated that closeness centrality could alone improve the precision and accuracy of the model. Compared to the case of using none of the centrality criteria n decision tree training, the case of using centrality resulted in a 1% rise in Accuracy and Sensitivity. The improvement percent was lower than the case in which both degree and closeness centralities were used in the decision tree process. Moreover, this improvement was less than test 3, in which only the degree centrality criterion was used to create a decision tree. Therefore, both degree and closeness centrality criteria had a significant effect on the efficiency of the proposed method to predict communication. Moreover, degree

Table 2. Evaluation of proposed method based on the method presented

Algorithm	Sensitivity	Precision	F-measure
Proposed method	0.98	0.94	0.96
Method with NB model	0.89	0.9	0.93
Method with RF model	0.92	0.9	094
Method with SVM model	0.91	0.91	0.93
Ordinary decision tree method	0.9	0.91	0.9
Neural network method	0.92	0.9	0.92

centrality was more important compared to closeness centrality.

According to Table 2, the proposed method in the present paper had better performance than all three methods presented, decision tree and neural network in terms of Sensitivity and F-measure. However, the proposed method was at last ranked in terms of precision. In both studies mentioned above, like the proposed method of extant study, the application of the machine learning approach led to better and more satisfying results when personal characteristics and network structure features were used rather than the absence of these features. However, the selection of personal characteristics was effective in the performance of the learning model. In the extant study, the use of personal characteristics, including education, living place, and job conditions, produced better results than other studies that used other information, such as users' interest, sending and receiving a message in the network, or other behaviors and social activities of users in the network.

Discussion

Social networks are dynamic due to continuous increases in their members, communications, and links, while these links may be lost. This study was conducted with the aim of investigating the link and communication between social network users using the centrality criterion and decision tree. According to the research result, there was an increase in sensitivity, which indicated model power in identifying positive categories (i.e., communications) when users' characteristics were used. It means that the model could identify potential latent communications. It can be stated that users are more willing to make a relationship with users similar to them through common neighbors. It can be stated that the assumption of this study was confirmed, i.e., knowing the contribution of each common neighbor increases the performance power of the proposed method.

The presence of both degree and closeness centrality criteria in this study were effective in the efficiency of the proposed method to predict communications. Moreover, the degree centrality criterion was more important than the closeness centrality.

Users' personal characteristics were indeed more significant than their behavioral traits; hence, these characteristics are more effective factors for the performance of the learning model because users first consider the profile and personal information of other users in the network before they decide to communicate. In the next step, users are more likely to make a relationship with other users if the personal characteristics in their profiles are desirable. The second priority is assessing the behavioral traits and activity background of the considered user. Therefore, the proposed method of this study had an appropriate and acceptable performance to produce satisfying results. In addition, the proposed method used personal characteristics mentioned in users' profiles, so it had better performance than those studies that used users' behavioral traits and activity background. Therefore, this method can predict latent communications in social networks and achieve promising results, although further studies are required.

In the present research, a weighted decision tree model was used by considering of importance of each characteristic for the final phase of communication prediction. It is possible to use other data mining models to further valuation and compare results with proposed methods.

This study used an outline dataset extracted from asocial network DBLP. The MapReduce programming model can be used when the big data volume is available. It is also possible to use flow processing techniques frequently used in big data techniques. These techniques allow a rapid flow of data that is producing or changing. The proposed method in this study has been compared to the current studies on predicting communications in social networks.

A machine learning-based method was proposed to predict communication in stoical networks of microblogs. In this research, different learning models, such as Naïve Bayes and Random Forest, were used to predict communication, and the results of each method were presented separately. In the method proposed to train these models, various features extracted from network structure (e.g., common neighbors between two disconnected nodes) or those related to nodes (e.g., number of followers in the network and number of received messages) were used (8). This study used some metrics, such as Precision, F-measure, and Sensitivity, to evaluate their proposed method (9). In another study, a machine learning approach-based method was used to predict latent communications in the social network of Twitter. This method used structural features of the network and considered behavioral traits and desires of users to train a learning model (10).

To evaluate the proposed methods of extant study based on the method presented in studies by Hosseini et al. (11), decision tree method, and neural network, the proposed algorithm was implemented with similar conditions of test 1.

The present study considered communications between two users, while there are other entities in social networks, including different groups and communities, software, blogs, pages, and games that can be predicted the user may communicate with which one of these bodies. Therefore, the proposed method in this study can predict different kinds of communications in a social network. However, global methods identify all path structures. The method introduced in the present paper defined a new node similarity index that uses all local and global features of a network. This method used the similarity between nodes in a graph that was created indirectly based on communicational data. In the method introduced by Ebadsichani et al. (12), the algorithm starts by placing each vertex in a separate association. There is no edge at the beginning, while one-by-one adding edges leads to integration of the associations at two ends of this edge if division modularity increases. Division modularity is computed based on the graph to which edges are added and indicate associations. If adding an edge does not create integration in associations, that edge will be an edge in the association. Hence, this case does not change the modularity rate. The number of divisions found in the process equaled the number of vertices (n). Every division has a modularity value, and after edges were added, the division with the largest modularity is taken as output. Hosseini Sedeh et al. (13) used users' interests overlapping to measure their similarities. Users' interests are determined based on their performances, such as their answers to the questions asked on a website like Stack Overflow or edition of a paper in Wikipedia. All users' actions are shown as a vector, and similarities between two users are indicated with the cosine between two vectors. In the method presented in the study (14), the similarity criterion is defined based on the common friends shared between two users. In this study, the subgraph of common friends and their

communications are extracted. The more the edges in this subgraph, the stronger the relationship between two users and the more alike they are.

The random walker algorithm was proposed in the paper. The random walker introduced by Yazdi et al., moves randomly on the graph and then goes from each vertex to adjacent vertex, considering the existing edges. The idea of Zhu's algorithm implies that random walker spends a longer time due to the high density of edges. Zhu used the random walker to define the distance between two vertices. The distance (d_{ij}) between two vertexes i and j indicates the average number of edges across which the random walker must pass to reach i from j. This method defines the global absorber of vertex i as the closest neighbor of this vertex (that has the lowest d_{ij} value). This method also defines the local absorber of vertex i as a vertex that i is its closest neighbor. Online social networks such as Facebook suggest new friends to users, and this is a process based on a transparent and clear social network in which users add each other as friends and create a network. A large part of the initial work on link prediction infers the new interactions between users by focusing on a unified network. However, users create several implicit social networks through their daily interactions, such as leaving comments on individuals' posts or ranking similar products shared between different users. The authors of the present paper introduced a method in which both implicit and explicit social networks were used to solve the group/item suggestion problem (15).

The extant study showed that complementary information of the user's item network could be successfully integrated with the friendship network to improve friendship suggestion procedures. In this method, the famous Katz algorithm was changed to use a multifaceted network and provide friendship suggestions. Finally, the real and fake datasets were used, and results showed that the proposed method suggested more accurate friendship relationships than two path-based algorithms with the same source (16).

Murata & Moriyasu, studied the effect of network structure on communication prediction efficiency in the form of clustering. The results indicated that clustering is effective in increasing prediction accuracy and efficiency. However, this technique did not work in solitude networks with poor connections (17). In Morata's research, link prediction is made on networks like Yahoo! Answers, where it is done by using graph theory and its features, as well as weighted edges between the nodes of this network. . The results obtained on the new databases have also been evaluated, and the results have been reported to be very effective (18). Today, users spend a lot of time surfing the Internet and social networks to make online purchases and social media. But one of the problems for business owners and managers of these media is a large number of these media, in other words, the existence of many competitors in this field. Therefore, users are surrounded by a large amount of information and have various options to use and spend time among the media. Having a proper understanding of user interests has become increasingly important for retailers who intend to create a personalized service for a target market and how these users relate to each other. Today, the size and number of online social networks are increasing day by day. For this reason, the analysis of social networks has become a popular issue in many branches of science. Relational prediction is one of the key issues in analyzing the evolution of social networks. With the increase in the size of social networks, the need to create and develop scalable communication prediction algorithms is felt more (19).

Recommendation

In this research, we considered only the communication that will be established between two users. But in a social network, there are other entities such as various

groups and communities, software, blogs, pages, and various games that can be predicted that in the future the intended user will Which one of these will communicate? Therefore, the method presented in this research can be used to predict different types of communication in a social network.

Conclusion

Some factors, such as personal characteristics and centrality criteria, affect the efficiency of the method. Removal of each factor in the learning process of the decision tree model indeed reduces the efficiency criteria. This study presented a method to predict communication in social networks. The proposed method benefited the advantages of big data techniques, users' characteristics and importance in the network, and data mining methods. Hence, this method had optimal performance in the prediction of latent communications. The method presented in the extant study not only used the information extracted from the network structure (e.g., centrality criteria of nodes) but also applied personal characteristics of users (e.g., education level, job status, and living place) to increase efficiency and precision of performance rather than other studies. Moreover, optimal results were obtained because the degree and closeness centrality features were considered a unique weight for each node in computations. Social networks' policymakers and watchdogs can use these results in their decisions.

Conflict of interest

The authors declare that they have no conflict of interests.

Authors' contribution

Ahmad Ghadamkheir and Seyed Alireza Derakhshan developed the study concept and design. Ashraf Shahmansoury acquired the data. Ahmad Ghadamkheir and Seyed Alireza Derakhshan analyzed and interpreted the data, and wrote the first draft of the manuscript. All authors contributed to the intellectual content, manuscript editing and read and approved the final manuscript.

Informed consent

Questionnaires were filled with the participants' satisfaction and written consent was obtained from the participants in this study.

Funding/financial support

There is no funding.

References

1. Marin A, Wellman B. Social Network Analysis: An Introduction. In: Scott, J. and Carrington, P.J., Eds., The Sage Handbook of Social Network Analysis, Sage Publications, Thousand Oaks;2011. https://www.scirp.org/(S(351jmbntvnsjt1aadkozje))/reference/referencespapers.aspx?referenceid=2414681
2. Zhu L, Guo D, Yin J, Steeg GV, Galstyan A. Scalable temporal latent space inference for link prediction in dynamic social networks. IEEE Transactions on Knowledge and Data Engineering. 2016;28(10):2765-2777. https://doi.org/10.1109/TKDE.2016.2591009
3. Carrington PJ, Scott J, Wasserman S. Models and methods in social network analysis. Cambridge University Press, Cambridge;2005. https://www.amazon.com/Methods-Network-Analysis-Structural-Sciences/dp/0521600979
4. Linyuan L, Ci-Hang J, Tao Z. Similarity index based on local paths for link prediction of complex networks. Physical Review. 2009;80(4):046122-046136. https://journals.aps.org/pre/abstract/10.1103/PhysRevE.80.046122
5. Leicht EA, Holme P, Newman ME. Vertex similarity in networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2006;73(2):026120-026132. https://doi.org/10.1103/PhysRevE.73.026120
6. Wattenhofer M, Wattenhofer R, Zhu Z. The YouTube Social Network. Proceedings of the International AAAI Conference on Web and Social Media. 2021;6(1):354-361. https://ojs.aaai.org/index.php/ICWSM/article/view/14243
7. Koren Y, Bell R, Volinsky C. Matrix Factorization Techniques for Recommender Systems. IEEE Computer. 2009;42(8):30-37. https://doi.org/10.1109/MC.2009.263
8. Liu W, Lu L. Link prediction based on local random walk. EPL (Europhysics Letters). 2010;89(5):58007-58016. https://doi.org/10.1209/0295-5075/89/58007

9. Ahmed C, ElKorany A, Bahgat R. A supervised learning approach to link prediction in Twitter. Social Network Analysis and Mining. 2016;6(1):1-11. https://doi.org/10.1007/s13278-016-0333-1
10. Han S, Xu Y. Link Prediction in Microblog Network Using Supervised Learning with Multiple Features. Journal of Computers. 2016;11(1):72-82. https://doi.org/10.17706/jcp.11.1.72-82
11. Hosseini M, Sultan Aghaei MR, Zamani Borujeni F. A Study of Link Prediction Methods in Social Networks. in the First National Conference on Distribution Computing and Big Data Processing. Shahid Madani University of Azerbaijan;2015. https://civilica.com/doc/590362/
12. Ebadsichani R, Khaiambashi MR, Khosravi Farsani H. Study of Demographic Factors on Link Prediction in Social Networks. in the 8th International Conference on Information Technology and Knowledge, Iranian Information and Communication Technology Association;2016. https://www.sid.ir/paper/893168/fa
13. Hosseini Sedeh M, Sultan Aghaei MR, Zamani Borujeni F. Presenting a New Method for Predicting the Type of Linkage in Heterogeneous Social Networks, in the 3nd Internatioanl Conference on Knowledge -Base Engineering and Innovation;2016. https://civilica.com/doc/623109/
14. Yazdi Sh, Mirzaei K. Musavi Zadeh Meybodi MT. The use of multiple categories to predict the link between the entities of a social network, in the First International Conference on Web Research;2016. https://civilica.com/doc/378220/
15. Yazdi Sh. Presentation of a hybrid model using genetic algorithm to predict links in social network, in the National Conference on Information Technology, Computer & Communication;2015. https://civilica.com/doc/451297/
16. Scellato S, Noulas A, Mascolo C. Exploiting Place Features in Link Prediction on Location-based Social Networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining;2011. https://doi.org/10.1145/2020408.2020575
17. Murata T, Moriyasu S. Link prediction of social networks based on weighted proximity measures. In Web Intelligence, IEEE/WIC/ACM International Conference on;2011. https://ieeexplore.ieee.org/document/4427070
18. Su Q, Chen L. A method for discovering clusters of e-commerce interest patterns using click-stream data. Electronic Commerce Research and Applications. 2015;14(1):1-13. https://doi.org/10.1016/j.elerap.2014.10.002
19. Sherkat E, Rahgozar M, Asadpour M, Structural link prediction based on ant colony approach in social networks. Physica A: Statistical Mechanics and its Applications. 2015;419(1):80-94. https://doi.org/10.1016/j.physa.2014.10.011