Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 3860083, 13 pages
https://doi.org/10.1155/2021/3860083

Research Article

Intelligent Link Prediction Management Based on Community Discovery and User Behavior Preference in Online Social Networks

Jun Ge,[1,2,3] **Lei-lei Shi**,[1,3] **Lu Liu**,[4] **Hongwei Shi**,[2] **and John Panneerselvam**[4]

[1]*School of Computer Science and Telecommunication Engineering, Jiangsu University, China*
[2]*School of Information Engineering, Suqian University, China*
[3]*Jiangsu Key Laboratory of Security Tech. for Industrial Cyberspace, Jiangsu University, China*
[4]*School of Informatics, University of Leicester, UK*

Correspondence should be addressed to Lu Liu; l.liu@leicester.ac.uk

Received 7 April 2021; Accepted 19 May 2021; Published 1 June 2021

Academic Editor: Varun Menon

Link prediction in online social networks intends to predict users who are yet to establish their network of friends, with the motivation of offering friend recommendation based on the current network structure and the attributes of nodes. However, many existing link prediction methods do not consider important information such as community characteristics, text information, and growth mechanism. In this paper, we propose an intelligent data management mechanism based on relationship strength according to the characteristics of social networks for achieving a reliable prediction in online social networks. Secondly, by considering the network structure attributes and interest preference of users as important factors affecting the link prediction process in online social networks, we propose further improvements in the prediction process by designing a friend recommendation model with a novel incorporation of the relationship information and interest preference characteristics of users into the community detection algorithm. Finally, extensive experiments conducted on a Twitter dataset demonstrate the effectiveness of our proposed models in both dynamic community detection and link prediction.

1. Introduction

With the rapid development of social networks and the wide spread application of intelligent terminals, we are facing an unprecedented volume of data generation, which in essence referred to as the big data era [1–3]. The big data era has led to various changes in the society and in our everyday lifestyle, where social networking is playing a pivotal role with great significance [4], which is a field of managing large-size datasets in a distributed computing environment. These datasets require robust network algorithms to transport huge block files efficiently. The traditional processing dataset approach involves a basic data placement technique that delivers resultant data blocks and exchanges block replicas in the cluster. And the widespread use of the Internet around the world enables people to connect with each other. People use the Internet for various purposes such as watching movies, reading news, searching information via search engines, shopping in e-commerce websites, and establish connections with friends via online social networks [5–8].

Online social networks are now flooded with various forms of data in massive volume, as people make new connections, post their updates, share and comment on other updates, etc., and such a structure of online social networks emphasizes the need to study their social relationships [9]. Besides, users can obtain information from related objects or spread information. Predicting the possible links between objects and/or information, based on the known information [10], would enable us to better understand the evolution of social networks [11, 12], also help business planners to make decisions, carry out precise services based on user connections, and achieve greater business value [4, 5, 9, 13–15].

In recent years, the essence of online social networks is largely identified to reflect the real-world networks. In the

Internet, people create their own content, tag, like, comment or join the community, and connect with other users. Better recommendation of potential friends for users can increase the connectedness of users with a better user experience. So far, recommendation systems are widely used in online social networks [16, 17], for various purposes such as new friend suggestions, and to offer efficient information retrieval services, etc., which in turn increase the overall traffic flow in the Internet.

In this context, the recommendation system used for suggesting friends introduced into the social network platform forms the basic function of social network services. The recommendation technology based on link prediction has become a research hotspot in recent years because of its high accuracy and low algorithm complexity. Link prediction is one of the basic problems in social network analysis, resolving which can provide us with an understanding of the mechanism of network evolution in theory, and further help us to optimize social networking services in accordance with the evolution of network structure. The concept of link prediction involves utilizing network structure information and node attributes to predict the possibility that users who have not yet generated link relationship in the network becoming friends in the future, to recommend the results with high possibility as "target users," so it is naturally suitable for relationship recommendation in social networks.

Although we can learn from the research on recommendation methods in traditional social network, due to the complexity and diversity of social networks, there are still many problems yet to be resolved in the field of relationship recommendation in online social networks, and the accuracy of relationship recommendation needs to be further improved [9, 18, 19]. To this end, we intend to develop a more effective relationship recommendation algorithm that is suitable for online social networks, characterizing higher prediction accuracy, low algorithm complexity, and better system integration. Herein, we propose an improved intelligent link prediction management technique based on exploiting the relationship strength information in online social networks. Moreover, considering the network structure attributes and interest preference of users as important factors affecting the links, further improvement is achieved by encompassing the community detection algorithm with the relationship information and interest preference characteristics of users. Finally, this paper designs a friend recommendation model, by integrating the intelligent link prediction algorithm and the label propagation community detection algorithm.

The main contributions of this paper are as follows:

(1) This paper proposes a community detection algorithm-based user behavior preference model (UBP), which can improve the data quality from the source of community detection. Specifically, in the calculation of community influence on nodes, the influence from nonadjacent nodes is also included. Through multiple iterations of experiments, the proportion of influence weight between adjacent nodes and nonadjacent nodes is identified to conform to real environment. Extensive experiments show that the proposed community detection results are better than the existing methods, and our community detection structure is more reasonable and accurate. Based on the UBP algorithm, the DPRank algorithm is introduced [4], where the global influence is replaced by the topology of social network and the local influence of nodes. Our approach not only ensures the accuracy of the algorithm but also improves its efficiency

(2) This paper proposes the novel link prediction algorithm based on label propagation. Firstly, we collect the attribute features and text information of users to explore their potential preferences and extract tags and then construct the user feature vector model to calculate the similarity between users. Then, based on an improved multisource label propagation community detection algorithm (multisource label propagation algorithm (MSLPA)), similar communities are mined. Finally, based on community, we use link prediction to estimate the node pairs with closest relationship strength and select the Top-k potential friend list as recommendation to users. This method not only improves the accuracy but also reduces the computational complexity of the link prediction algorithm. Performance evaluation carried out based on the real dataset shows that our algorithm achieves better performance than the state-of-the-art local index methods

(3) We conducted experiments to evaluate the performance of our proposed models. The experimental results on a Twitter dataset demonstrate the effectiveness of our proposed UBP and MSLPA models, in terms of both dynamic community detection and link prediction

The rest of this paper is organized as follows. In Section 2, we review previous studies of link prediction. In Section 3, we introduce our proposed UBP method. We present the MSLPA model in Section 4. We discuss our experimental results in Section 5, and in Section 5.1, we draw our conclusions and future work.

2. Related Work

Online social networks focus on the interaction between individuals and network topology. Internet, scientist cooperation network, power network, aviation network, biological network, and so on, all reflect the characteristics of social networks [4, 12]. It is worthy of note that most of the interconnected things can be abstracted as social networks. Typical networks, such as the cooperative network between scientists in the academic field, and the network structure of protein molecules in the biological field [18–20] resemble the topology of social networks. The ultimate purpose of studying the topological structure and properties of different types of networks is to understand the evolution principle of social networks, predict the future evolution direction and trend of social networks, estimate links [21] to better cope with the sudden changes in social networks, and apply this

knowledge in actual networks [22–24]. For example, in the field of counterterrorism, law enforcement officers can analyze social network links to identify the direct and indirect connections of suspects. "Guess what you like" in e-commerce website, recommendation of the target of interest in Twitter [25], etc., can be seen as the application of link prediction in real life. Because of its significant practical value, link prediction and recommendation systems have become the hotspot research topics in the context of online social networks [26, 27].

As one of the important research directions of data mining, link prediction in online social networks has received a wider attention, and many link prediction algorithms have been developed in the recent years. Traditional research methods [22, 28, 29] have two main ideas: Firstly, from the perspectives of machine learning, link prediction modeled as a typical learning problem and witnessed the application of techniques such as supervised logistic regression, support vector machine, random forest, and unsupervised learning algorithm based on Bayesian network [24]. The other idea is to mine the properties of nodes and network structure from the perspectives of social networks and predict the connection based on the similarity of nodes. These methods attempted to mine node and network related information as much as possible. Moreover, the maximum likelihood method is also heavily researched for link prediction and witnessed to have achieved reliable prediction results.

Srinivas et al. [25] comprehensively analyzed the importance of link prediction for social network analysis along with its application in bioinformatics, information retrieval, e-commerce, and other fields and summarized various link prediction technologies based on classification and kernel function and discussed the latest progress and future research direction of probability modeling in this context. Yang et al. [28] compared the advantages and disadvantages of various link prediction algorithms and conducted quantitative research in real networks. They pointed out that the similarity method based on network topology has become a research hotspot due to its simple algorithm and low computational complexity. According to the influence of different nodes, an improved similarity link prediction algorithm is proposed.

Li et al. [23] defined a preference function as a new attribute of supervised learning considering the preference of nodes and achieved reliable results. Gupta and others [24] abstracted the link prediction problem into a binary problem and established a Bayesian model to predict the possible connections of the network. The link prediction method based on probability model has been applied to various fields of social network research, in an attempt to establish a perfect social network recommendation system [30].

Bastami et al. [27] believed that most of the link prediction algorithms only consider either global or local information, but only few of them integrate both global and local structure information. A new fusion algorithm has been proposed, which considers community properties into account, and used the clustering algorithm to evaluate the link density at the community level to adjust the eigenvalues of nonlocal features and then combined the link information and nodes of neighbor nodes. Finally, the similarity model has been integrated to predict the link. This algorithm innovatively integrated the local features of nodes and the structural features of communities.

Community detection and link prediction are two different directions of social network analysis [31], while the former is used to mine network topology, the latter works based on the network structure to predict the future evolution trend in the social network. At present, a few researchers have tried to use community detection to improve the accuracy of link prediction. Yao et al. [32] studied the significance of clustering coefficient in link prediction and proposed a new periodic evolution model. Experiments on the Enron network dataset showed that the prediction ability of this model is better than that of the classical link prediction algorithms.

Link prediction and community detection are of great significance in social network analysis, as they describe the formation, evolution, and nature of the social network from different aspects. In this paper, we propose a new link prediction method for social networks by incorporating community detection, ultimately to offer valuable friends recommendation for users.

3. Concept and Definition

3.1. Social Network. In the real world, there is a wide range of connections and interactions between various things. The components of the system can be described as nodes. Many of these systems can be modeled by social networks. Studying the formation mechanism and evolution mechanism of social networks can help us to understand the nature of the system better. For example, the discovery of six-degree separation theory in a typical social network shows the world is actually very small. Link prediction involves solving one of the most fundamental problems in network science, which is to restore and predict the missing information. When a system becomes more complex, many nodes that have not been linked at present may establish links in the future. The problem of predicting such nodes with a higher likelihood of establishing links in the future is called link prediction [33]. The interaction or connection between nodes is described as the edge between nodes. A typical social network usually characterizes frequent connections between nodes, where new links become more active.

Any network can be abstracted as a graph, which is composed of finite sets. Such a graph structure encompasses node set representing the individual in the network, edge set representing the connection in the network. Generally, a network can be represented as a node or entity set of the same type. A social network usually encompasses a specific user, a link set, and a link between nodes. If a node has a complete set of possible links, the nonexistent link instances can be represented as a prediction problem of generating links. Thus, the link prediction problem can be defined as follows: given an instance of a social network, we can predict the possibility of generating links and judge the possibility of connectedness according to the score value. Generally, such a kind of prediction problem is studied with a training set and a test set.

Figure 1(a) represents a complete network, which consists of 12 nodes and 16 edges. Four edges are extracted as test

edges as shown in Figure 1(b), and the remaining 12 edges are training datasets. Through a link prediction algorithm, four test edges are given a score value according to the possibility and compared with all other edge scores that do not exist. A higher score value reflects a more accurate prediction result.

3.2. Community Structure. Community is a subgraph structure in the network topology; as shown in Figure 2, the density of node links within the community structure is higher than that of between communities. This implies that the internal relationship within communities is closer and in line with the cognition of real-world social communities.

4. UBP Model

In this section, we describe our proposed community detection algorithm based on user behavior preference (UBP), which can improve the data quality from the source of community detection. In the calculation of community influence on nodes, the influence from nonadjacent nodes is also included. Through multiple iterations of experiments, the proportion of influence weight between adjacent nodes and nonadjacent nodes is identified to conform to the real environment. Based on the UBP algorithm, the DPRank algorithm is introduced [4], where the global influence is replaced by the topology of the social network and the local influence of nodes, which ensures the accuracy of the algorithm and improves its efficiency.

4.1. Community Detection. In this section, we explain the community detection method.

As demonstrated in Figure 3, we divide the social network into communities and determine the exact community structure. We propose the UBP algorithm to achieve the desired objectives. In the UBP algorithm, we consider the topological relationship between the adjacent users in the community, the topological relationship between indirect users, and the impact probability between the users. The UBP algorithm provides a good community structure. Then, in the divided community structure, we use the IPIP model [8] based on the LT model to determine the most effective nodes in each community in order to maximize the impact of the social networks. Finally, a situation where a possible loss of seed nodes can occur in the real society, we use the Full Preselected Search, namely, FPSS algorithm [8]. When the seed node is lost, the FPSS algorithm immediately finds a replacement node to compensate for the loss of influence diffusion caused by the loss of the original seed node.

Besides, we pursue the following steps [8], as illustrated in Figure 4.

(1) The user interest is modeled, and the Pearson coefficient is used to obtain the interest similarity matrix between users

(2) The influence probability of users is modeled based on their concerns and interactions

(3) Social network is modeled based on user interest similarity and user influence probability

(4) All the users are calculated and sorted in the social graph accordingly, where the central community and the central nodes are identified, and finally the community is detected

(5) Link prediction based on independent cascade model is achieved

(6) Link prediction in the entire network is achieved

4.2. Modeling User Interest Similarity

4.2.1. Modeling User Interest. Users share their feelings and thoughts in Twitter by posting a tweet and participating in social activities. The LDA model [4, 8] is a three-level Bayesian model of document subject word, which uses probability deduction to find the semantic structure of a given dataset to obtain the topic of the text. Hence, the LDA algorithm is used to analyze the user's document for obtaining the user-interest matrix. On this basis, the interest similarity between users in social networks is resolved.

4.2.2. Modeling User Similarity. A substantial amount of hidden information is present in the massive data. We can use this gigantic data to extract the desired information by analyzing the data and then receiving the user score on posts and finally generating the user-post matrix. Pearson's sparse is an efficient technique for obtaining user similarity. Pearson's coefficients given in equation (1) with a modified cosine similarity and user-post scores are used to compute the similarity of users in the network.

$$w_{\text{pearson}}(i,j) = \frac{\sum_{u \in I(i,j)} \left(r_{i,u} - \overline{r_i}\right)\left(r_{j,u} - \overline{r_j}\right)}{\sqrt{\sum_{u \in I(i,j)} \left(r_{i,u} - \overline{r_i}\right)^2} \sqrt{\sum_{u \in I(i,j)} \left(r_{j,u} - \overline{r_j}\right)^2}}, \tag{1}$$

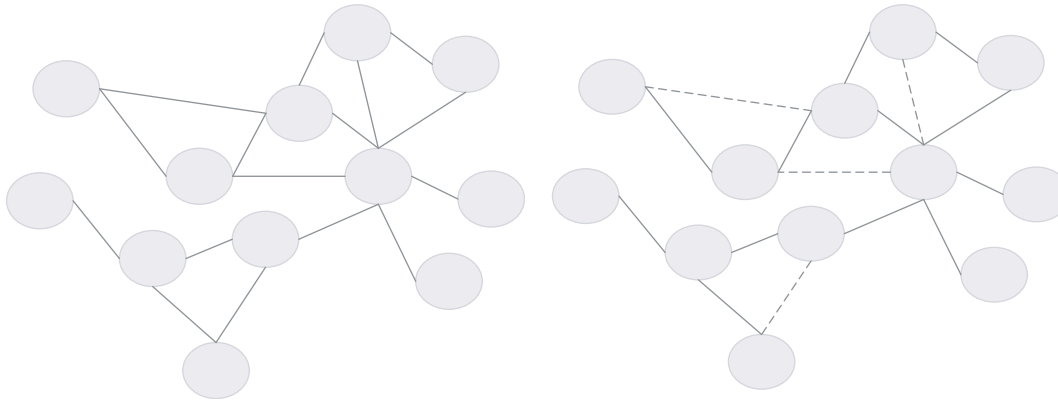
where $I(i,j)$ shows the set of common posts of user u and user v. While $r_{i,u}$ represents the score of user i on post u, and $\overline{r_i}$ denotes the average interest scores of user i and user j.

4.3. Modeling the User Influence Probability

4.3.1. Initial Influence Probability Modeling for Users. In general, user's influence is the degree of trust between each other while the community's influence on the average user is the sum of all the influence in the community.

In this paper, the influence probability of users originates from the number of interactions between users, which mostly reflects the influence of the relationship between the users. Thus, we evaluate the initial influence probability based on the user's interactions. The initial influence probability of user u on user v is calculated using

$$F(u,v) = f(u,v), \tag{2}$$



(a) Complete network (b) Prediction network

Figure 1: An example of link prediction.

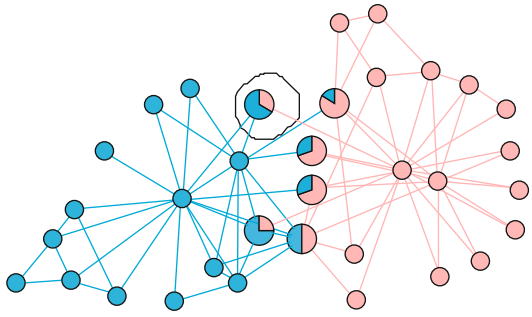


Figure 2: Community structure.

where $f(u, v)$ shows the number of interactions between users u and v.

4.3.2. Probability Prediction of Influence between Unconnected Users. In social networks, users have almost no explicit source of influence. A study on social networks [5] shows that the association between two unfamiliar users' average 4.7 hops, i.e., edges. By integrating research and reality, users will have more trust on their friends of friends. We use equation (3) to propose the probabilistic modeling for source nodes and two-hop target nodes.

$$P(A, C) = \begin{cases} F(A, C), A \to C \\ \dfrac{1}{|S|} \sum_{j \in S} P_{AB_j} P_{B_j C} + F(A, C), A \to B_j \to C \end{cases}, \tag{3}$$

where $A \to C$ indicates that user A is the follower of user C. S is a collection of common friends of users A and C, and P represents the probability of a user's influence on another user.

Figure 5 illustrates a graphical representation of the influence of a probability between the users. Moreover, it illustrates the relationship between two nodes in the network, where the weight value denotes the probability that a user may receive from another user.

Figure 6 shows the link relationship between many nodes, where black solid lines represent the concerns among the users while red dotted lines indicate the probability of predicting the influence between them. At this point, we believe that the probability of influence between users with links is obtained. We represent this social graph in the form of a matrix. For example, the processing of calculating $P(U_1, U_4)$ is given as follows:

$$P(U_1, U_4) = \frac{1}{2}(P(U_1, U_2) * P(U_2, U_4) + P(U_1, U_3) * P(U_3, U_4)) + F(U_1, U_4) \approx 0.678. \tag{4}$$

4.4. Modeling Social Networks. We model the social network based on user influence probability and user interest similarity. The weight of the side in a social network can be calculated using

$$F(i, j) = \eta P(i, j) + (1 - \eta) w_{\text{pearson}}(i, j), \tag{5}$$

where η is a tunable parameter that regulates the importance of user influence probability and user similarity in the social network model. After the experimental observations, we set $\eta = 0.4$. The result of equation (5) is the weighting of a social network. This is because the user influence is not equal and produces a realistic response.

5. MSLPA Model

Preference connection has proved to be an idea that can improve the accuracy of link prediction. On this basis, this paper proposes an integration of the label propagation and the link prediction algorithm. Firstly, we collect the attribute features and text information of users to explore their potential preferences and extract tags and then construct the user feature vector model to calculate the similarity between users. Then, based on an improved multilabel propagation community detection algorithm (multisource label propagation algorithm (MSLPA)), similar communities are mined. Finally, based on community, we use link prediction to identify the node pairs with the closest relationship strength and select

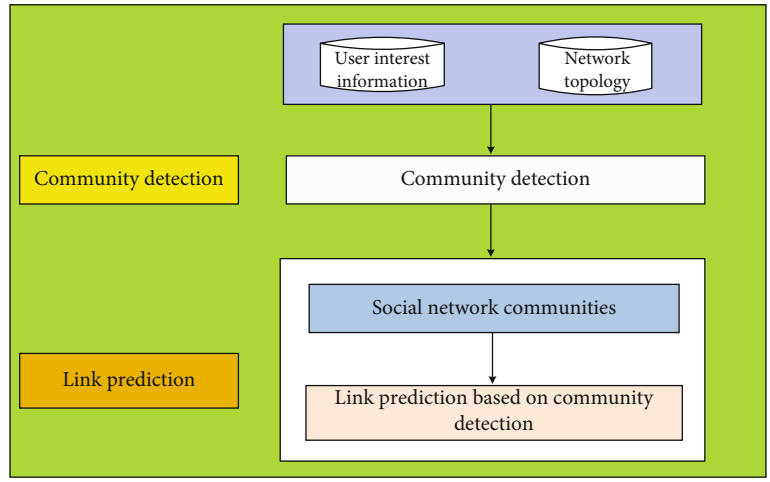


Figure 3: Link prediction method based on community detection.

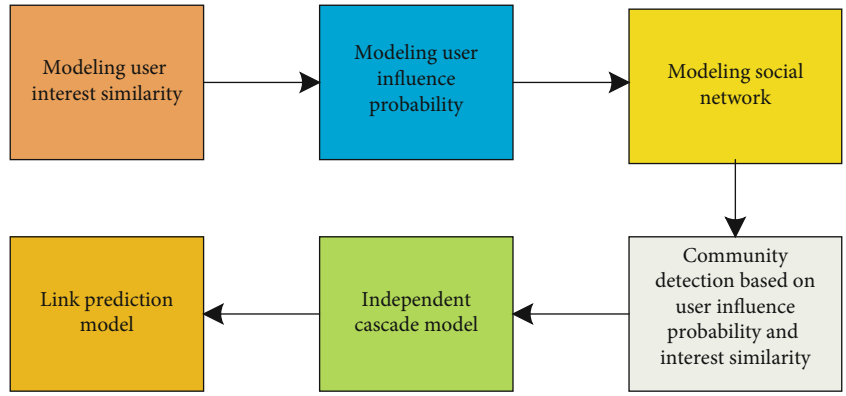


Figure 4: UBP algorithm steps.

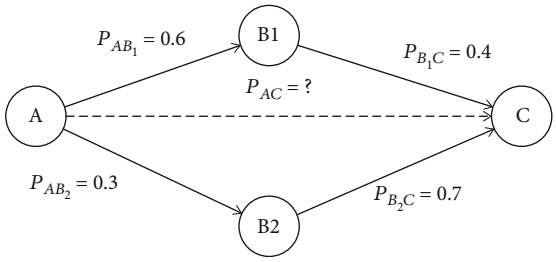


Figure 5: Schematic representation of user influence probability prediction.

the Top-k potential friend list as a recommendation to users. This method not only improves the accuracy but also reduces the computational complexity of the link prediction algorithm.

In this section, we first introduce the label propagation model and then propose a link prediction recommendation algorithm combined with label propagation. Unlike the single similarity index, in combination with the relationship strength, our model takes into account not only the local index but also the global structure information and user attribute information. In comparison with the traditional link prediction algorithm, our proposed model not only improves the accuracy of recommendation but also makes the network more compact after the label propagation. Furthermore, the required amount of calculation is considerably reduced, as our model avoids computation for all the nodes; thus, it is fast and efficient.

5.1. Link Prediction. Community detection can be used to mine the social network user information and network structure attributes and further can be applied for link prediction. User information and network structure usually complement each other well. Social network theory shows that people with similar characteristics tend to establish a relationship. Traditional link prediction methods are required to calculate information about all the nodes in the whole network. Due to its high computational complexity, this strategy incurs a large amount of calculation, which significantly affects the efficiency of the prediction algorithm in large-scale social networks. In order to solve this problem, researchers put forward the idea of dividing the large-scale social network into communities and then using link prediction to calculate the similarity within the communities. For the recommendation system, it is equivalent to the recall stage first, which not only reduces the computation scale but also makes the recommendation source more targeted. This paper introduces

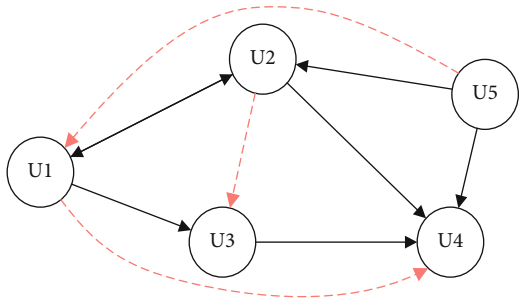


Figure 6: Schematic diagram of links between the nodes.

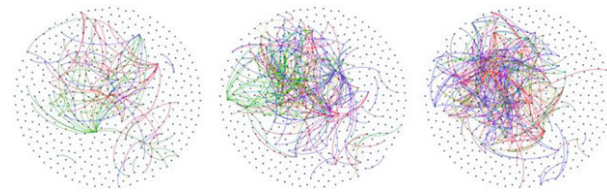


Figure 7: Community evolution.

an improved multilabel propagation algorithm. Tag extraction uses user preference features extracted from user self-set and topic model, divides the community, and then calculates the similarity through link prediction. According to the similarity ranking, it further helps users to find friends who have similar interests, or they may know. In this way, our recommendation algorithm makes use of the user's personal preference attributes and social network structure information to make the recommendation more personalized and improve the recommendation accuracy.

5.2. Community Detection. With the gradual deepening of social network research, people begin to realize the existence of locally connected node sets in the network, which have a very important impact on the topological structure of the whole graph.

Community detection involves mining the communities in the network through various algorithms, so as to analyze the communities and understand the evolution trend of communities. Figure 7 shows the evolution process of different types of communities on a social platform. The community detection algorithm based on label propagation algorithm (LPA) is efficient and simple, with only linear complexity. It is suitable for large-scale networks and widely used in industry.

Tags reflect the interests and characteristics of users, and the process of tag propagation reflects the simulation of human information exchange. The process of extracting tags from node behaviors for further propagation retains the node's personality. Based on the nodes after tag propagation, some communities based on similar interests are formed. This idea is similar to the process of community formation in social networks, where users tend to join most of the neighbor's community. With the gradual expansion of social networks, various types of communities are formed. In this paper, some local similarity indexes such as CN, Jaccard, and AA are compared based on tags. The results show that these algorithms can achieve good prediction results in the case of relatively dense data. Community structure exists objectively, but users in a certain community only interact with those users who have a direct connection with them. However, in a community, users who are not directly connected are also regarded as being "close." Friend recommendation system usually gives priority to users belonging to the same community, as birds of a feather flock together, which reflects the community detection algorithm. In fact, it divides the community in a certain way. On this basis, it can make further link prediction for each community. From the computational perspective, community detection can be regarded as equivalent to the decomposition of the network, which reduces the computational complexity.

At present, information on social networks is growing explosively. Tag propagation algorithm has become a research hotspot because of its simplicity, high efficiency, and low computational complexity. In order to facilitate label management, we establish corresponding tag systems, which can help users to retrieve users and related resources they are interested in.

The traditional label propagation algorithm assumes that each node only carries one tag, belongs to a community, and does not distinguish the importance of different tags, which is not consistent with the real-world social networks. This paper improves the tag propagation algorithm, distinguishes the tag weight, and can carry out multilabel propagation. Finally, those nodes with the maximum similarity believe to be in the same community but actually may belong to multiple communities at the same time.

5.3. Label Selection and Extraction. Users often set up some personal tags or post some blog posts and comments in social networks. The analysis and mining of this text information make the recommendation more personalized. In the Twitter network, users usually add their own tags to reflect their personality.

The main idea of label propagation algorithm is described as follows: suppose a node "a" and its neighbor nodes are $\{A_1, A_2, A_n\}$. Each node carries its own label. During the process of propagation, the label of node "a" is determined by the label of its neighbor node; that is, the label of node "a" with most neighbors is taken as the label of node "a." With the continuous spread of tags, it tends to be stable in the end.

The traditional LPA algorithm does not distinguish the importance of tags. In real networks, there are often first-class, second-class, and third-class tags. For example, in the user profile system, the tags selected by users themselves are more important than the tags that are counted out. In order to make the community detection more reasonable, we distinguish according to the importance of labels. In this paper, we mainly determine the following two types of labels with different weights and adjust the weights of various labels through experiments according to the actual situation.

(1) Labels and system tags set by users themselves

(2) There are obvious tags in the user text, such as ∗∗ school and ∗∗ location

In addition, we use the LDA topic model to extract the corresponding tags according to the blog information published by users. In the tag extraction process, we input some candidate seed words. Since the entity set is not directly represented in the user's Tweets, we need to use specific tools to find the candidate entity set and then compare the similarity with the seed vocabulary to obtain the required classification tag. The input document of the model is collected based on historical behavior data of users. Finally, it is merged with the first two types of tags as a user's tag set. Users in social networks generally have multiple tags. It is obviously not suitable for social networks to select only one tag in the propagation process, as in the case of the traditional tag propagation algorithm. In this paper, we propose an improved multisource label propagation algorithm (MSLPA).

5.4. Improved Label Propagation Algorithm. Considering the interaction characteristics of real social networks, we improve the classic label propagation algorithm and apply it to the whole recommendation system. The improved communication process can be divided into the following three steps:

Step 1. Initialize labels instead of community numbers for all nodes, and nodes carry multiple labels, and assign weights to labels at the same time.

Step 2. Refresh the labels of all nodes iteratively. The label of all neighbor nodes is investigated, and the weight is calculated; then, the labels are assigned with the largest number to the current node. When the number of labels with the largest number is not unique, select one randomly.

Step 3. After n iterations, convergence is reached, and the algorithm is completed. In the final community, the nodes with the greatest similarity degree belong to the same community.

Considering the computational complexity, the most widely used text-matching model in the field of text analysis is vector space model (VSM) [13]. In the concept of VSM, a document is represented in vector form, and its relevance is measured by the similarity between vectors. Each dimension of the vector corresponds to a term, and the weight of each component element of the vector represents the importance of the term in the document. This paper studies the label words, which can be abstracted as document vectors. The way to calculate semantic similarity using cosine similarity can be expressed as follows:

$$\mathrm{RSV}(A,B) = \frac{\sum_{i=1}^{n}(a_i \times b_i)}{\sqrt{\sum_{i=1}^{n}(a_i)^2} \times \sqrt{\sum_{i=1}^{n}(b_i)^2}}. \tag{6}$$

When the value of $\mathrm{RSV}(A,B)$is 1, it means that the labels between the two nodes are the same; when the value is 0, there is no overlapping label between the two nodes. All neighbor nodes whose values exceed the predetermined propagation threshold are selected, and the most selected labels are passed to the corresponding nodes. Finally, we use the improved link prediction algorithm to predict the possible edges in the community.

The propagation process is shown in Figure 8. The label propagation process simulates the information exchange process and behavior of people in social networks. During initialization, each node has a fixed label, which is uploaded by the user and extracted from the previous LDA model. Then, some nodes are randomly selected to interact with other nodes. The $\{A, B, C, D\}$ in Figure 8(a) represents the label currently owned by the node. We randomly select the node {4, 2, 1} as the receiving node. First, node 4 receives the label from neighbor 1 and the label from neighbor 2, and the label of node 4 becomes $\{D, A, B\}$.Then, the label of node 2 is updated. Since its neighbor node 3 has multiple labels, a label is randomly selected and passed on to node 2. Here, suppose the propagation label is selected, and the label of node 2 becomes $\{B, A, C, D\}$. Finally, the label of node 1 is updated. The neighbor nodes {2, 3, 4} are randomly selected to propagate. If there are repeated labels, the corresponding label weight is increased by one in turn, for example, after the propagation in Figure 8(b) is finished. If there are two "a" labels in node 1, the weight is two, and the weight of $\{B, C\}$ labels is one. The whole propagation process is asynchronous. Some nodes will receive the label information of neighbor nodes first and can end the propagation process according to the label propagation. Finally, according to the different labels of stable nodes, the similarity is calculated by formula and divided into multiple communities. At the same time, a node can belong to multiple communities.

The time complexity of the algorithm is very low. In the process of label propagation, nodes are randomly reordered to ensure the convergence of the algorithm. Because the formation of community only depends on the local information of the network, the algorithm is very suitable for community detection and partition in large-scale social networks.

Firstly, the initial seed node is set, and the label is propagated to the surrounding nodes according to the label weight. After each round, the similarity between nodes is calculated according to formula (1), and then, the label is updated.

Finally, according to the specified number of iterations, the program ends after the community becomes stable.

Label classification and link prediction can promote each other and have homogeneity in social relations; that is to say, people with similar attribute characteristics are more likely to establish friendship relationship; therefore, the closer the relationship is, the more likely they are to have the same label.

The nodes in the circle in Figure 9 are communities formed based on a certain relationship. It can be seen that nodes I and E, M, and K are equally likely to generate links through calculation. However, since I and E belong to a certain community relationship, there are similarities between them. Therefore, link prediction based on label division is more likely to recommend friends within a community, hence can be more targeted. On the other hand, users' interests are

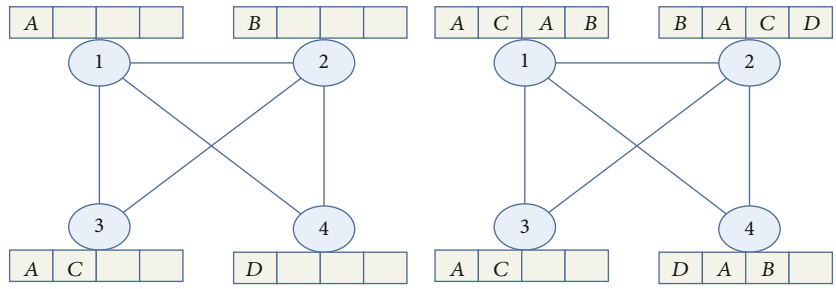


Figure 8: Schematic diagram of multisource label propagation process.

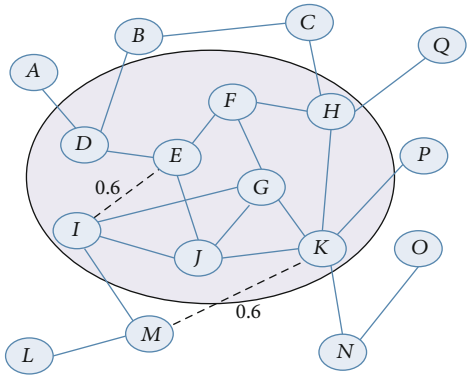


Figure 9: Schematic diagrams of link prediction differences among different nodes.

diverse, and users tend to add topics of interest and find similar people. Through the way of label propagation, more potential friends can appear in the recommendation list, which also reduces the cold start problem of new users to a certain extent.

Here, the first stage uses the multisource label propagation algorithm, as in algorithm one. The algorithm inputs the preprocessed network data, including adjacency matrix and label set. In the second stage, the MSLPA algorithm is used to calculate the similarity of the corresponding users, and the Top-k potential users are selected to generate the recommendation list.

5.5. Recommendation Model Based on Combination Algorithm. In this section, this paper proposes a friend recommendation system model by integrating label propagation and link prediction. The whole recommendation process is divided into four modules: data cleaning, community detection, link calculation, and user recommendation. The whole system model framework is shown in Figure 10.

Firstly, user's home page information and blog text information are collected, and a single microblogging is regarded as a short text. Then, the label is generated by extracting the subject words of the blog post content using the LDA model. Then, the ID information of the followers and their followers is obtained, and the labels are saved in the corresponding table after being extracted for further data cleaning.

Secondly, we read the obtained label and link relationship data, use the community algorithm to divide the community, and stop the iteration when the community is relatively stable to the set conditions.

Finally, we traverse all the communities formed in the previous step, use the improved link prediction algorithm to calculate the similarity within the community, and select the Top-k users as the recommendation list for each user according to the score ranking. For the friend recommendation for a given user, the community to which the user belongs to is first obtained, and the potential friends with the highest similarity are ranked according to the final total score.

6. Experiments

In this section, we present the results obtained in our experiments conducted on real-world short-text data collections in order to demonstrate the effectiveness of our proposed method. We consider four typical algorithms as our benchmark methods, namely, CPM [9], COPRA [18], LFM [19], and GCE [20]. We also introduce the collection of the dataset, experimental setup, analysis, the baseline approach, and the model evaluation.

6.1. Dataset.

6.2. Experimental Setup. The experiments are conducted in a machine with Intel I5 2.5 GHz CPU and 4G memory. The experiments use standardized mutual information, named NMI to measure the correlation between the community structure generated by the community detection algorithm and the standard community structure to evaluate the accuracy of the algorithm using equation (7). We use the overlapping modularity Qov to evaluate the network structure of overlapping communities in order to measure the quality of community detection as expressed in equations (7)–(9).

$$\mathrm{NMI}(X|Y) = 1 - \frac{1}{2}\left(H(X|Y)_{\mathrm{norm})} + H(Y|X)_{\mathrm{norm})}\right), \quad (7)$$

where X and Y represent the experimental community structure and the standard community structure, respectively. The higher the NMI value, the more similar the partition result is to the standard network structure and the higher the accuracy of the algorithm to partition the community.

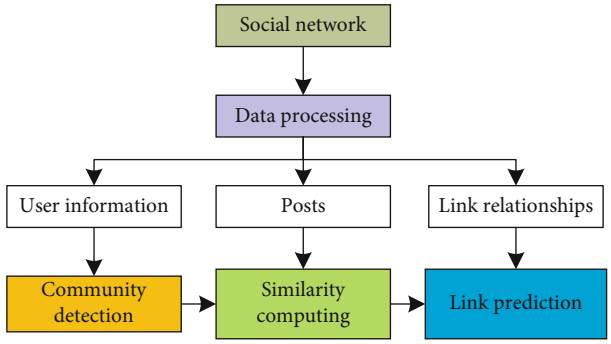


Figure 10: Framework of recommendation system.

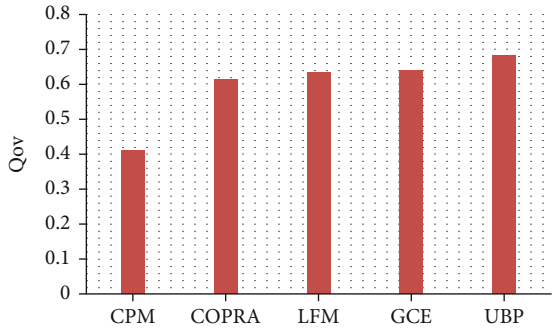


Figure 11: Comparison of Q values of algorithms for Twitter datasets.

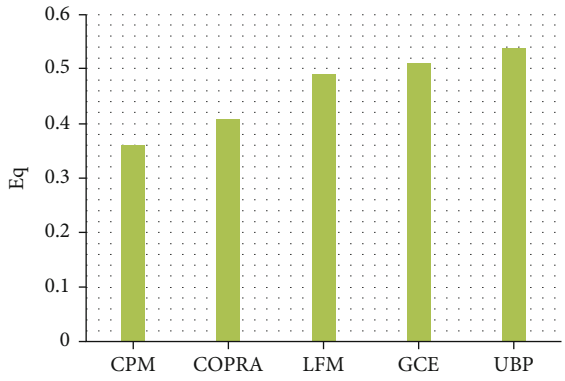


Figure 12: Comparison of EQ values.

Table 1: Test dataset detail table.

Network	Nodes	Edges	Coefficient	Degree
USAir	332	2126	0.749	12.81
NS	379	914	0.798	4.82
PB	1222	16714	0.360	27.36
Slavko	334	2218	0.488	13.28
Email	1133	5451	0.254	9.620
Router	5022	6258	0.033	2.49
Jazz	198	2742	0.618	27.70
Twitter	11241	732193	0.162	65.14

Table 2: UBP comparisons with other approaches.

		CPM	COPRA	LFM	GCE	UBP
Twitter	EQ	0.3545	0.4084	0.4902	0.5076	0.5387
	Q_{ov}	0.4122	0.6089	0.6298	0.6425	0.6856

$$\mathrm{Qov} = \frac{1}{m}\sum_{c\in C}\sum_{i,j\in V}\left[r_{ijc}A_{ij} - \omega_{ijc}\frac{k_i^{\mathrm{out}}k_j^{\mathrm{in}}}{m}\right], \tag{8}$$

where $A$is the adjacency matrix, K shows the degree of users, m indicates the number of edges, r_{ijc} denotes the probability that users i and j belong to community c, $r_{ijc} = \iota(P_{i,c}, P_{j,c})$, $P_{i,c}$ represents the probability that i belongs to community c, and ω_{ijc} denotes the probability that node i or node j belongs to community c.

$$\tau\left(P_{i,c}, P_{j,c}\right) = \frac{1}{\left(1+e^{-f\left(P_{i,c}\right)}\right)\left(1+e^{-f\left(P_{j,c}\right)}\right)}, \tag{9}$$

$$\omega_{ijc} = \frac{\sum_{j\in V}\tau\left(P_{i,c}, P_{j,c}\right)}{|V|} \times \frac{\sum_{i\in V}\tau\left(P_{i,c}, P_{j,c}\right)}{|V|}, \tag{10}$$

where f is defined as $f(x) = 60x - 30$ and Qov ranges from 0 to 1. The larger the value of Qov, the better the overlapping community structure will be.

COPRA, LFM, GCE, and CPM methods are selected comparative evaluation in the experiment. In order to avoid the influence of randomness of the algorithm in the experiments, we conduct 20 experiments and obtain the average results. Our model only needs one experiment because of the stability of the algorithm. The NMI and Qov of each algorithm are obtained. Figures 11 and 12 illustrates the changes in the NMI and Qov values in the social network with the mixing parameter.

6.3. Experimental Result. We test five algorithms on the Twitter dataset as shown in Table 1, which includes eight networks and their detailed information. Table 2 and Figure 11 demonstrate the comparison of community modules obtained after experiments on Twitter dataset, which terms the EQ and Qov with these algorithms, in which we can observed our proposed UBP algorithm performs better than the CPM, COPRA, LFM, and GCE algorithms due to the fact that our method considers both the user interest similarity and user influence probability with regard to Qov. And in the same way, Figure 12 shows the comparison of overlapping community modules attained from experiments on Twitter dataset. It can be observed that in the Twitter dataset, our algorithm also performs better than the CPM, COPRA, LFM, and GCE algorithms because our method considers both the user interest similarity and user influence probability.

Figures 11–14 demonstrate the output of our experimental analysis. Figures 13 and 14 show the effect of the variable parameter η and variable parameter λ, respectively.

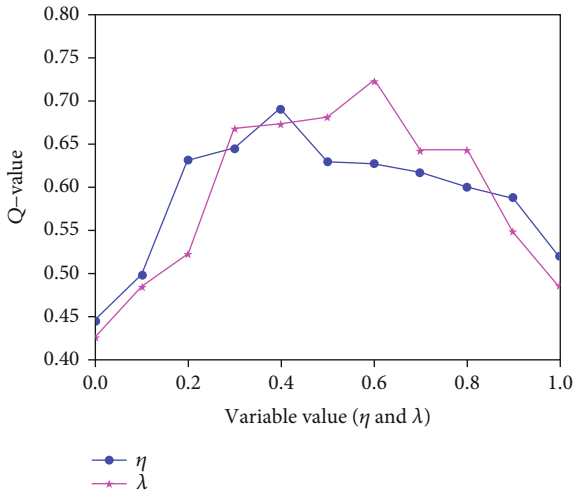


Figure 13: The effect of variable parameters.

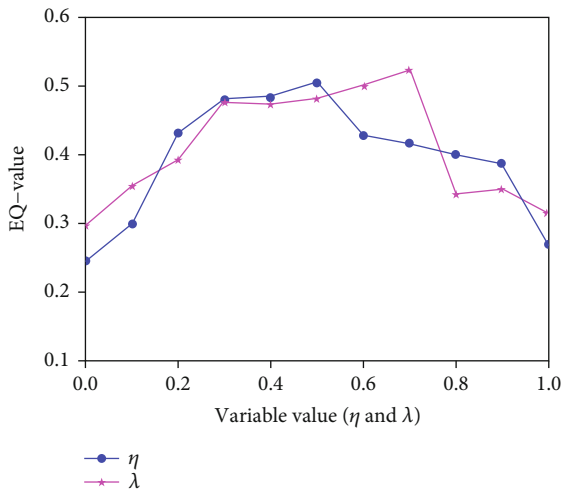


Figure 14: The effect of variable parameters.

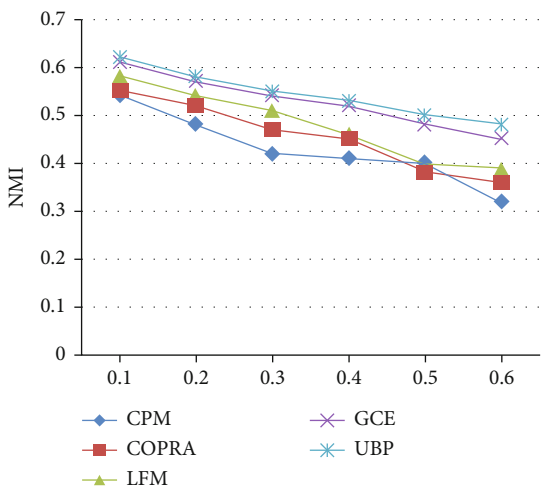


Figure 15: NMI comparisons of five algorithms.

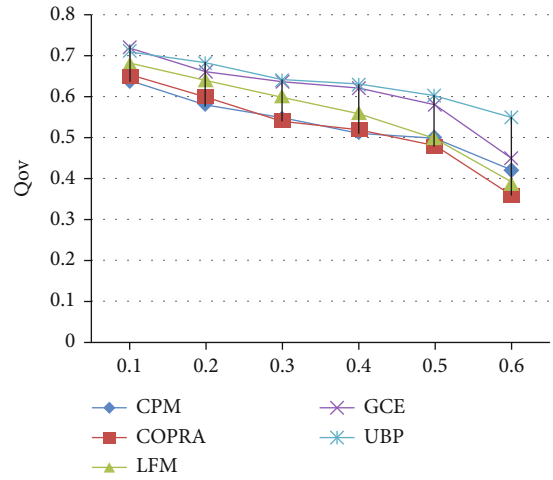


Figure 16: Qov value comparisons of five algorithms.

For variable weights and real weighted networks, we performed several experiments. The output values are not a variate during experimental analysis, and our algorithm shows an exceptional stability.

From Figure 15, when mixing parameter value is small, then the community structure of the network is more obvious, and the boundary between communities is clear. GCE characterizes a higher NMI value than UBP, but with an increase in the mixing parameter value, UBP algorithm exhibits a higher NMI value than the GCE algorithm. This shows the higher accuracy of the UBP algorithm in a large-scale network.

Figure 16 shows that the community modularity Qov divided by the GCE and UBP algorithms has more advantages than the other three algorithms regardless of the mixing parameter value. This is attributed to the improvement in the random strategy of the two algorithms, as it reduces the difference in the results and plays an important role for the users with higher potential influence to a certain extent. UBP has more potential influence, which can divide communities with high quality either in the network with obvious community structure or in fuzzy networks. The UBP algorithm improves the stability of community detection process and the quality of community generation to a certain extent, when compared with the existing community detection algorithms.

In order to verify the effectiveness of the proposed algorithm, link prediction algorithms such as CN, AA, RA, PA, JACCARD, HPI, LP, and Katz are selected for comparison. By comparing the results in Tables 3 and 4, it can be found that when compared with CN, JC, PA, AA, RA, HPI, LP, and KATZ algorithms, our proposed MSLPA algorithm delivers better prediction accuracy and AUC in most networks. In Table 3, the performance value of our MSLPA algorithm is improved by 10% to 20%, when compared with that of the traditional local index algorithms such as CN and RA. Furthermore, the average performance of PATH-based algorithms such as HPI and LP is also improved by 8.9%. In particular, our algorithm has more obvious advantages in the network with obvious social network properties, and its

Table 3: Precision comparison.

Networks	CN	AA	PA	RA	Jaccard	HPI	LP	Katz	MSLPA
USAir	0.628	0.722	0.657	0.690	0.494	0.677	0.742	0.641	0.755
NS	0.868	0.887	0.864	0.875	0.635	0.769	0.761	0.876	0.889
PB	0.816	0.833	0.858	0.827	0.667	0.850	0.739	0.733	0.904
Slavko	0.814	0.858	0.823	0.853	0.641	0.745	0.750	0.736	0.802
Email	0.845	0.867	0.859	0.811	0.643	0.813	0.816	0.819	0.844
Router	0.649	0.673	0.743	0.658	0.352	0.852	0.856	0.874	0.841
Jazz	0.775	0.802	0.796	0.772	0.361	0.798	0.811	0.802	0.829
Twitter	0.877	0.854	0.867	0.839	0.857	0.851	0.753	0.855	0.938

Table 4: AUC comparison.

Networks	CN	AA	PA	RA	Jaccard	HPI	LP	Katz	MSLPA
USAir	0.941	0.952	0.960	0.968	0.919	0.877	0.952	0.945	0.970
NS	0.968	0.980	0.964	0.975	0.976	0.979	0.981	0.983	0.988
PB	0.916	0.933	0.899	0.927	0.855	0.870	0.919	0.931	0.942
Slavko	0.954	0.948	0.939	0.953	0.946	0.945	0.950	0.936	0.959
Email	0.845	0.876	0.869	0.891	0.842	0.856	0.866	0.898	0.899
Router	0.649	0.677	0.943	0.658	0.651	0.652	0.946	0.957	0.660
Jazz	0.955	0.958	0.969	0.972	0.961	0.949	0.953	0.963	0.975
Twitter	0.971	0.966	0.977	0.963	0.957	0.951	0.952	0.965	0.982

prediction accuracy in PB network, Slavko network, and Twitter network is greatly improved. The clustering coefficient and average network degree of these networks are relatively large, the community structure is more obvious, and they are more likely to be connected by some relationship attributes. For example, the Jazz network, which consists of small groups of music, has a close relationship with a class of students on the Twitter social network. For Router, USAir, and other networks with weak social properties and sparse data, our MSLPA algorithm plays a very limited role in strengthening the relationship, when compared with other algorithms, and the prediction accuracy is not significantly improved.

In social networks such as Twitter, different mutual friends represent different relationship, which can be observed through the closeness of their neighbors. For this reason, our improved predictors have a good predictive effect on social networks. However, there is a lack of applicability for new users due to the lack of link information.

7. Conclusions and Future Work

In this paper, we presented a method called UBP based on relationship strength according to the characteristics of social networks and improved the prediction accuracy of existing link prediction algorithms based on this mechanism. The method uses both the topology structure and information content in the social network. Unlike the traditional community detection algorithm experiencing issues such as randomization of community center user selection and data sparsity of user's interest, we proposed a method based on the LPA algorithm, named MSLPA. We optimized the expansion of community structure and reduced the redundancy in the community. Extensive experimental analysis on real-world datasets showed that our UBP method performs considerably better than the existing state-of-the-art methods.

As a future work, we plan to consider more real-world networks. The UBP and MSLPA methods will be evaluated on the social network for event topic detection and propagation. The UBP and MSLPA methods will be also deployed to dynamically discover and self-configure the hot events in a dynamic social network environment. The proposed algorithms will also be implemented for various different types of networks to address the existing problems in different domains.

Data Availability

The data used to support the findings of this study have not been made available because some other papers will also use this data, which is not published yet.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The work reported in this paper has been supported by the National Natural Science Foundation of China Program (61502209 and 61502207) and the Suqian Municipal Science and Technology Plan Project in 2020 (S202015).

References

[1] H. Lu, S. Liu, H. Wei, and J. Tu, "Multi-kernel fuzzy clustering based on auto-encoder for fMRI functional network," *Expert Systems with Applications*, vol. 159, article 113513, 2020.

[2] A. Monney, Y. Zhan, J. Zhen, and B.-B. Benuwa, "A multi-kernel method of measuring adaptive similarity for spectral clustering," *Expert Systems with Applications*, vol. 159, pp. 1135–1147, 2020.

[3] S. Tang, S. Yuan, and Y. Zhu, "Data preprocessing techniques in convolutional neural network based on fault diagnosis towards rotating machinery," *IEEE Access*, vol. 8, pp. 149487–149496, 2020.

[4] L. L. Shi, L. Liu, Y. Wu, L. Jiang, J. Panneerselvam, and R. Crole, "A social sensing model for event detection and user influence discovering in social media data streams," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 1, pp. 141–150, 2020.

[5] S. Y. Shih, M. Lee, and C. C. Chen, "An effective friend recommendation method using learning to rank and social influence," PACIS, 2015.

[6] H. Peng, J. Li, S. Wang et al., "Hierarchical taxonomy-aware and attentional graph capsule RCNNs for large-scale multi-label text classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2505–2519, 2021.

[7] H. Peng, R. Yang, Z. Wang, J. Li, and R. Ranjan, "LIME: low-cost incremental learning for dynamic heterogeneous information networks," *IEEE Transactions on Computers*, vol. 99, p. 1, 2021.

[8] S. C. Yang, L. L. Shi, and L. Liu, "Community detection method based on user influence probability and similarity," in *2018 Sixth International Conference on Advanced Cloud and Big Data (CBD)*, pp. 183–190, Lanzhou, 2018.

[9] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[10] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "A unified deep model for joint facial expression recognition, face synthesis, and face alignment," *IEEE Transactions on Image Processing*, vol. 29, pp. 6574–6589, 2020.

[11] S. Tang, S. Yuan, and Y. Zhu, "Convolutional neural network in intelligent fault diagnosis toward rotatory machinery," *IEEE Access*, vol. 8, pp. 86510–86519, 2020.

[12] J. Ge, L. -L. Shi, Y. Wu, and J. Liu, "Human-driven dynamic community influence maximization in social media data streams," *IEEE Access*, vol. 8, pp. 162238–162251, 2020.

[13] G. Salton and C. S. Yang, "On the specification of term values in automatic indexing," *Journal of Documentation*, vol. 29, no. 4, pp. 351–372, 1973.

[14] Z. Li, X. Fang, and O. R. L. Sheng, "A survey of link recommendation for social networks: methods, theoretical foundations, and future research directions," *ACM Transactions on Management Information Systems (TMIS)*, vol. 9, no. 1, pp. 1–26, 2017.

[15] M. Slokom and R. Ayachi, "A new social recommender system based on link prediction across heterogeneous networks," in *International Conference on Intelligent Decision Technologies*, pp. 330–340, Springer, Cham, 2017.

[16] S. Khusro, Z. Ali, and I. Ullah, "Recommender systems: issues, challenges, and research opportunities," in *Information Science and Applications (ICISA) 2016*, pp. 1179–1189, Springer, Singapore, 2016.

[17] T. Ha and S. Lee, "Item-network-based collaborative filtering: a personalized recommendation method based on a user's item network," *Information Processing & Management*, vol. 53, no. 5, pp. 1171–1184, 2017.

[18] S. Gregory, "Finding overlapping communities in networks by label propagation," *New Journal of Physics*, vol. 12, no. 10, p. 103018, 2010.

[19] P. Kim and S. Kim, "Detecting overlapping and hierarchical communities in complex network using interaction-based edge clustering," *Physica A: Statistical Mechanics and its Applications*, vol. 417, pp. 46–56, 2015.

[20] C. Jia, J. Ma, Q. Liu, Y. Zhang, and H. Han, "Linkboost: a link prediction algorithm to solve the problem of network vulnerability in cases involving incomplete information," *Complexity*, vol. 2020, Article ID 7348281, 14 pages, 2020.

[21] B. K. Nagra, B. Chhabra, and D. Sharma, *"Recommendation and Interest of Users", Intelligent Communication, Control and Devices*, Springer, Singapore, 2018.

[22] Y. P. Xiao, "3-HBP: a three-level hidden Bayesian link prediction model in social networks," *IEEE Transactions on Computational Social Systems*, vol. 5, pp. 430–443, 2018.

[23] Y. Li, P. Luo, Z. P. Fan, K. Chen, and J. Liu, "A utility-based link prediction method in social networks," *European Journal of Operational Research*, vol. 260, no. 2, pp. 693–705, 2017.

[24] A. K. Gupta and N. Sardana, "Naïve Bayes approach for predicting missing links in ego networks," in *2016 IEEE international symposium on Nanoelectronic and information systems (iNIS)*, pp. 161–165, Gwalior, India, 2016.

[25] V. Srinivas and P. Mitra, "Link prediction using thresholding nodes based on their degree," in *Link Prediction in Social Networks*, pp. 15–25, Springer, Cham, 2016.

[26] W. Kai, S. Liu, H. Chen, and X. Li, "A new link prediction method for complex networks based on resources carrying capacity between nodes," *Journal of Electronics & Information Technology*, vol. 41, pp. 1225–1234, 2019.

[27] E. Bastami, A. Mahabadi, and E. Taghizadeh, "A gravitation-based link prediction approach in social networks," *Swarm and Evolutionary Computation*, vol. 44, pp. 176–186, 2019.

[28] Y. Yang, J. Zhang, X. Zhu, and L. Tian, "Link prediction via significant influence," *Physica A: Statal Mechanics and its Applications*, vol. 492, pp. 1523–1530, 2018.

[29] T. S. Li, "Deep dynamic network embedding for link prediction," *IEEE Access*, vol. 6, no. 5, pp. 29219–29230, 2018.

[30] Z. X. Guo, Z. Ma, and Z. Zhang, "A novel recommendation system in location-based social networks using distributed ELM," *Memetic Computing*, vol. 10, no. 3, pp. 321–331, 2018.

[31] Z. L. Liao, L. Liu, and Y. Chen, "A novel link prediction method for opportunistic networks based on random walk and a deep belief network," *IEEE Access*, vol. 8, no. 5, pp. 16236–16247, 2020.

[32] L. Yao, L. Wang, L. Pan, and K. Yao, "Link prediction based on common-neighbors for dynamic social network," *Procedia Computer Science*, vol. 83, pp. 82–89, 2016.

[33] V. Martínez, F. Berzal, and J. C. Cubero, "A survey of link prediction in complex networks," *ACM Computing Surveys (CSUR)*, vol. 49, no. 4, pp. 1–33, 2017.