IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 1, 2021, accepted September 20, 2021, date of publication September 24, 2021, date of current version October 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3115605

A Label Propagation Based Node Clustering Algorithm in Heterogeneous Information Networks

DONGJIANG LIU AND LEIXIAO LI
Inner Mongolia Autonomous Region Engineering and Technology Research Center of Big Data Based Software Service, College of Data Science and Application, Inner Mongolia University of Technology, Hohhot, Inner Mongolia 010080, China

Corresponding author: Dongjiang Liu (ldongjiang@yeah.net)

This work was supported in part by Inner Mongolia University of Technology Research Fund Key Project under Grant ZZ201908, in part by Inner Mongolia Key Technological Development Program under Grant 2019ZD015, in part by the Key Scientific and Technological Research Program of Inner Mongolia Autonomous Region under Grant 2019GG273, in part by Inner Mongolia University of Technology Research Project Doctoral Fund under Grant BS2020040, in part by the National Natural Science Foundation of China under Grant 62062054, and in part by the Natural Science Foundation of Inner Mongolia Autonomous Doctoral Fund under Grant 2020BS06007.

ABSTRACT With the fast development of network technology, great amount of data have been accumulated. Plenty of them are organized by using heterogeneous information networks (HIN). So mining heterogeneous information networks efficiently is very important. Node clustering is an essential part of this task. *And several clustering algorithms have been proposed. As all these algorithms contain complicated optimization procedure and matrix calculation procedure, complexity of these algorithms is very high. To overcome the shortage described above, in this paper, a new clustering algorithm is proposed. In this algorithm, several parameters should be inputted. These parameters include a heterogeneous information network, meta-paths that are used and the names of target types. During the clustering procedure, a homogeneous network will be built by the proposed algorithm firstly. All the target objects of HIN are treated as nodes of this network. The instances of meta-paths are edges. After the homogeneous network is constructed, label propagation procedure can be performed. Then the clustering result will be obtained. Obviously, by using the proposed algorithm to perform clustering, the complex optimization procedure and matrix calculation procedure are eliminated. As the convergence rate of label propagation procedure is fast, the proposed algorithm is very efficient. Besides, we can find that label propagation procedure can be executed in parallel. Thus, the proposed algorithm is easy to be parallelized. In this situation, it is fit for processing large scale HIN based on server cluster.* From experimental results, we can find that the proposed algorithm running faster than all the other algorithms for comparison.

INDEX TERMS Heterogeneous information networks, clustering, community detection, label propagation, graph mining.

I. INTRODUCTION

As network technology develops rapidly, it is very easy for all fields to share their data. So different types of data are accumulated, such as trajectory data, relational data, image data, text data and graph data etc. Recently, graph data becomes more and more prevalent. As graph can depict the relationship of different elements clearly, many fields try to organize their data in the form of graph. For example, social networks can be treated as a large graph [1]. In this graph, every user is a node. The relationship between two users will be treated as an edge. So, if two users are friends or share same interest, there will be an edge between them. In microbiology [2], all the microorganisms can be organized in a complex networks. In this network, all the microorganisms are treated as nodes. Potential connections between these components are edges. Computer network topology can also be viewed as a graph [3]. The network devices are nodes of the graph. If two devices are connected by a cable, there will be an edge between these two devices.

Heterogeneous information network (HIN) is a kind of special graph. In traditional graph, all nodes belong to the

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Asif.

same category. But, in HIN, nodes must belong to several different categories. And the edge meaning of HIN are also different. For example, in bibliography network, four types of nodes are contained, namely conferences, papers, terms and authors. The meaning of edges are distinct. If an edge connects an author and a paper, it means that this paper is written by the author. If an edge connects a paper and a conference, it means that the paper is published in this conference. If a term is connected to a paper, it means that this term exists in this paper. In microbiology, a complex network composed of microorganisms can also be viewed as a HIN. It is because different kinds of microorganisms are contained in this networks, including protein, virus and bacteria etc. In computer science, computer network can be viewed as a HIN as well. Because several different devices are contained in this network. These devices include routers, switches or servers etc. As HIN are very common, it is necessary to analyze HIN efficiently.

In recent years, many graph mining algorithms have been proposed. These algorithms can be divided into several different categories, including frequent subgraph mining algorithms [4], graph classification algorithms [5], link prediction algorithms [6], node clustering algorithms [7] and link based object ranking algorithms [8] etc. Node clustering, which is also called community detection, mainly tries to put nodes of a graph into different clusters. It is an important component of graph mining. Nowadays, many algorithms have been proposed to fulfill this task [9]–[14]. As large scale complex networks become common, there are also some algorithms proposed for performing clustering based on this kind of networks [15], [16]. Heterogeneous information networks is different from traditional homogeneous networks. Thus new methods are needed to analyze this kind of networks [17]–[19]. HIN clustering algorithm can be classified into two types. One kind of algorithms is mainly based on ranking method. In this kind of algorithms, objects' ranking will be considered while performing clustering. Sun *et al.* [20] proposed an algorithm named RankClus. This algorithm is a ranking based algorithm. While clustering, quality of cluster and ranking will be enhanced mutually. The most important drawback of this algorithm is that it only can put the target objects into two different categories. So if the target objects belong to more than two clusters, this algorithm will not work. Because of this, another algorithm called NetClus [21] was proposed. This algorithm constructs net-clusters by using the links across multi-typed objects. And all the target objects can be put into more than two categories. But it still has some shortcomings. Firstly, each target object are put into a cluster randomly. Secondly, in this algorithm, the impact of target object relationships is weakened. Another type of algorithms is meta-path based algorithms. This type of algorithms takes relationship of all the target objects into consideration. The most popular meta-path based clustering algorithm is named PathSelClus [22]. As instances of all the meta-paths can connect the target objects directly, the relationships of these nodes are clearly presented. Thus cluster quality will be improved effectively. When Meta-path is proposed, it is used by many algorithms to analyze HIN [23]–[25]. At the same time, Li *et al.* [26] proposed another meta-path based clustering algorithm. This algorithm considers not only the meta-paths of HIN, but also the attributes of target objects. As the attributes of target objects contains great amount of information, the clustering quality of this algorithm is improved. But it still has a shortage. If target objects don't have any attributes, the algorithm may be not better than other meta-path based algorithms. After that, several other attribute based algorithms were proposed [27], [28]. As links and attributes are very useful for clustering, a new algorithm [29] which tries to unify both link-based method and attribute-based method was proposed. Moreover, Liu *et al.* [30] proposed a new node mapping method. This method tries to map each target object into a vector. By using these vectors, clustering result can be improved efficiently. These algorithms are also based on meta-path. In these algorithms, some matrices are constructed firstly. Each matrix is corresponding to a specific meta-path of HIN. So, during the clustering process, complicated matrix computation and optimization procedure are performed. All these procedures are costly. And, at the same time, while the number of target objects increases, the matrix size will also increase. Thus the calculation process will spend more time. And, obviously, these algorithms are not fit for parallelization.

To overcome the low efficiency problem described above, a new algorithms is proposed in this paper. In this algorithm, label propagation method is used. This method is proposed by Raghavan *et al.* [31]. Initially, it tries to assign a label to each node. Then every node will propagate the label to its neighbors. As each node receives multiple labels, it must choose one for itself. The propagation and selection procedure will be repeated until the label of each node doesn't change. Obviously, the clustering process of this algorithm is very simple. So the running speed of label propagation method is faster than others. Besides, as each node can perform label propagation and label selection independently, this algorithm is easy to be parallelized. So running speed of this algorithm can be further accelerated by using multi-thread technology. In this paper, a new community detection algorithm is proposed for heterogeneous information networks. This algorithm is based on label propagation method. So its running speed is very fast.

As label propagation procedure is executed based on homogeneous network, a homogeneous network should be constructed based on HIN by the proposed algorithm firstly. In this network, all the target objects are contained. And the instances of meta-paths are used to connect these target objects. So these instances will be treated as edges of homogeneous network. As there may be more than one instances of meta-paths between two target objects, a weight value should be assigned to the edge that connects them. This weight value is used to represent the number of edges. As the instances of meta-paths are used while constructing a homogeneous network, the relationship of all the target objects are fully considered. After a homogeneous network is obtained, the label propagation can be executed. During label propagation process, the weight value should be considered. It is because the weight value is used to represent the number of instances between two target objects. So we can calculate closeness of target objects by using these weight values. The contribution of this paper is as below:

1 A new homogeneous network construction method is proposed. This method tries to construct a homogeneous network based on all the target objects and meta-paths of HIN. The relationships of all the target objects are fully considered by using this method.

2 A new label propagation method is proposed for performing node clustering. As weight values of the constructed homogeneous network should be considered while propagating labels, a new label propagation method should be used to perform clustering by the proposed algorithm.

3 A HIN clustering algorithm is proposed in this paper. This algorithm is based on the new label propagation method.

The rest of this paper is organized as follows. In section 2, the definition of some important concepts will be given. In section 3, the algorithm proposed in this paper is described. In section 4, the proposed algorithm will be empirically evaluated based on the real-world data sets. And the paper will be concluded in section 5.

II. DEFINITIONS

In this section, the definitions of some key concepts are introduced. These concepts are network schema, target type, target object, HIN clustering and meta-path.

Definition 1 (Network Schema): Suppose that $G(V, E)$ is a graph and $T = \{T_1, T_2, \ldots, T_t\}$ is a type set. Each element of set T represents a type. Every node of graph G at least belong to one type of set T. And each element of set T at least contains one node of graph G. A new graph $G_S(V_S, E_S)$ can be built based on set T. In this graph, node set V_S is composed of all the elements of set T. Set E_S contains all the edges that connect different elements of set T. Graph $G_S(V_S, E_S)$ is called the network schema.

In a bibliography network, four types of nodes can be found, including "author", "conference", "paper" and "term". So, according to definition 1, the network schema of bibliography network contains 4 nodes. As papers are written by authors, an edge should exist between "author" and "paper". At the same time, a paper must be submitted to a conference. So "paper" and "conference" should be connected. Moreover, as a term must be contained in a specific paper, "term" and "paper" should be connected as well. In this situation, according to the above description, the network schema of bibliography network contains 3 edges.

Definition 2 (Target Type): Suppose that graph $G(V, E)$ is a HIN. Set $T = \{T_1, \ldots, T_t\}$ is a type set. All nodes of graph G must belong to these t different types. If the clustering task is to put all nodes of type T_i into different cluster, then type T_i will be called target type.

Definition 3 (Target Object): Suppose that the clustering process is mainly performed based on all nodes of type T_i. From definition 2, we know that type T_i is called target type. Then nodes that belong to type T_i are called target objects.

Definition 4 (HIN Clustering): Suppose that $T = \{T_1, \ldots, T_t\}$ is a type set. In HIN $G(V, E)$, every node must belong to one type of set T. If clustering is performed based on all the nodes of type T_u, according to definition 2, type T_u should be called target type. Then this clustering process will be called HIN clustering.

Definition 5 (Meta-Path): Meta-path is an edge set. It consists of several edges of a network schema. The instance of a meta-path can be used to connect two nodes of HIN that are not adjacent.

Suppose that T_u, T_v and T_m are three elements of type set T. R_0 and R_1 are two edges of a network schema. In HIN, R_0 is used to connect nodes of type T_u and nodes of type T_m. R_1 is used to connect nodes of type T_m and nodes of type T_v. From definition 5, we know that $T_u \xrightarrow{R_0} T_m \xrightarrow{R_1} T_v$ is a meta-path. Thus, nodes of type T_u and nodes of type T_v are not adjacent. But they can be directly connected by using the instances of meta-path $T_u \xrightarrow{R_0} T_m \xrightarrow{R_1} T_v$. And, at the same time, we can find that two edges of network schema are contained in this meta-path. These edges are R_0 and R_1.

III. HINLPCLUS ALGORITHM

In this section, the proposed algorithm is introduced. To describe the clustering procedure of the proposed algorithm more clearly, a flow chart is very necessary. It is presented in figure 1. From the flow chart, we can find that two steps are contained in the proposed algorithm. In the first step, a homogeneous network is constructed. In this network, all the target objects are treated as nodes. And the instances of meta-paths are treated as edges. It is because meta-paths can depict the relationship of target objects very well. In the second step, the label propagation process is performed based on the constructed homogeneous network. The detailed introduction of each steps is presented in the following sections.

A. HOMOGENEOUS NETWORK GENERATION

In this part, the homogeneous network construction process will be introduced firstly. Let's take bibliography network as an example.

In bibliography networks, plenty of publication information can be found. Generally, nodes of bibliography networks belong to four categories, namely author, paper, conference and term. So the network schema of bibliography networks contain four nodes. At the same time, as an author can write a paper, a paper should be published in a conference, and a term is contained in a paper, the network schema of bibliography networks must contains three edges, namely (*author*, *paper*), (*paper*, *conference*) and (*term*, *paper*). Suppose that the target type is author. Then the clustering task is to put all the authors into different clusters. Thus, while generating

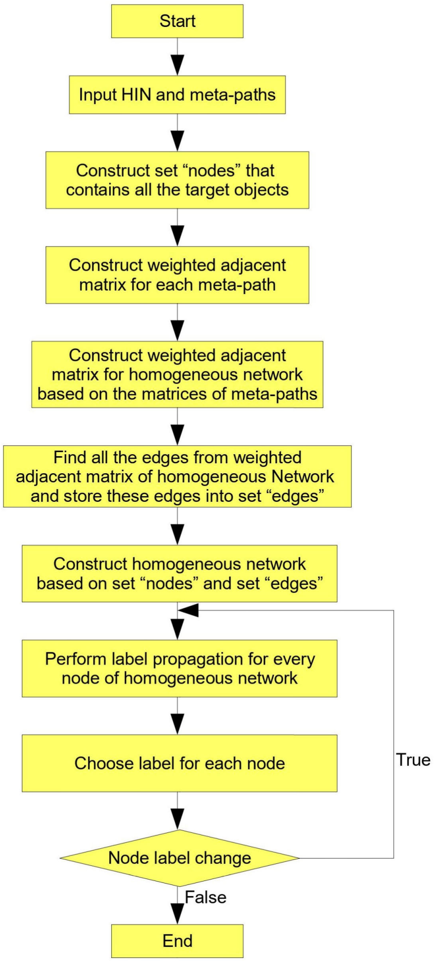

Start
Input HIN and meta-paths
Construct set "nodes" that contains all the target objects
Construct weighted adjacent matrix for each meta-path
Construct weighted adjacent matrix for homogeneous network based on the matrices of meta-paths
Find all the edges from weighted adjacent matrix of homogeneous Network and store these edges into set "edges"
Construct homogeneous network based on set "nodes" and set "edges"
Perform label propagation for every node of homogeneous network
Choose label for each node
True
Node label change
False
End

FIGURE 1. **Flow chart of the proposed algorithm.**

homogeneous network, all the authors are treated as nodes and instances of the meta-paths that connect authors are treated as edges. It is mainly because that, if two target objects are connected by an instance of a specific meta-path, they always share same characteristics. In this situation, these two target objects are more likely to be put into same cluster. In bibliography networks, three meta-paths that connect target objects can be found, including $author - paper - author$, $author - paper - conference - paper - author$ and $author - paper - term - paper - author$. The authors connected by the first meta-path are co-authorship of a paper. These authors may focus on the same research direction. The second meta-path tries to connect the authors that publish papers in the same conference. As a conference usually focus on a specific field, the authors who submit paper to the same conference probably have same research direction. Similarly, the authors who are connected by instances of the third meta-path may also focus on the same research direction. It is because that different papers concentrated on the same field always contain same keywords. For example, the papers concerning deep learning may contain keywords like neural networks, CNN, LSTM and RNN etc. But the paper concerning image processing always contains keywords like image segmentation, image reconstruction and resolution etc.

Obviously, the target objects connected by instances of the meta-paths described above are more similar. So, instances of meta-paths are treated as edges of homogeneous network. And a weight value should be assigned to each edge. It is used to represent the number of instances found between two target objects. As there may exists more than one instances, the number of instances should be checked while generating homogeneous networks. If there is only one instance, weight value of the corresponding edge will be set to 1. If n instances of meta-paths are found between two target objects, the weight value will be set to n. According to the above description, we can conclude that the larger the weight value is, the more identical characteristics these two target objects share. So it is very likely that these two target objects will be put into same cluster. Thus, while performing label propagation, the weight values of the edges should be considered.

The features of homogeneous network generated based on HIN have been described clearly. Now the homogeneous network construction procedure will be introduced detailedly. During this procedure, we should construct a weighted adjacent matrix for each meta-path at first. Then the weighted adjacent matrix of homogeneous network will be calculated based on the matrix of all the meta-paths.

In this situation, the matrix construction process of each meta-path should be introduced firstly. Suppose that we tend to construct a weighted adjacent matrix for meta-path $A - P - C - P - A$. A, P and C are three different node types. And each node type contains more than one nodes. As type A is target type, the nodes belonged to this type are target objects. Obviously, in HIN, nodes of type A are connected to the nodes of type P. The nodes of type P are connected to the nodes of type C. M_{AP} is weighted adjacent matrix constructed based on nodes of type A and nodes of type P. This matrix can be directly constructed based on HIN. In M_{AP}, if element $e_{i,j}$ is equal to 1, it means that the $i - th$ node of type A and $j - th$ node of type P are connected by an edge of HIN. If element $e_{i,j}$ of M_{AP} is equal to 0, it means that the $i - th$ node of type A and $j - th$ node of type P are not connected. Similarly, weighted adjacent matrices M_{PC}, M_{CP} and M_{PA} can be constructed in the same way. After matrices M_{AP},M_{PC}, M_{CP} and M_{PA} are obtained, the weighted adjacent matrix M_{APCPA} of meta-path $A - P - C - P - A$ can be calculated. The calculation method is as below:

$$M_{APCPA} = M_{AP}M_{PC}M_{CP}M_{PA} \tag{1}$$

Obviously, some elements of matrix M_{APCPA} are larger than 1. If an element $e_{u,v}$ of matrix M_{APCPA} is equal to w, it means that there are w instances of meta-path $A - P - C - P - A$ between $u - th$ target object and $v - th$ target object. Accordingly, the matrix of other meta-paths can be calculated by using the same method.

After obtaining weighted adjacent matrix of all the meta-paths, the weighted adjacent matrix of homogeneous network can be calculated. Suppose that there are three meta-paths contained in HIN, that are $A-P-A$, $A-P-C-P-A$ and $A-P-T-P-A$. The matrix of meta-path $A-P-A$ is M_{APA}. The matrix of meta-path $A-P-C-P-A$ is M_{APCPA}. And the matrix of meta-path $A-P-T-P-A$ is M_{APTPA}. Now we can compute the weighted adjacent matrix M_{homo} of homogeneous network as below:

$$M_{homo} = M_{APCPA} + M_{APA} + M_{APTPA} \tag{2}$$

With weighted adjacent matrix M_{homo}, the homogeneous network can be constructed. All the target objects of HIN will be treated as nodes. And all the edges should be got from matrix M_{homo}. If element $e_{p,q}$ of matrix M_{homo} is equal to k, it means that there is an edge between the $p-th$ target object and the $q-th$ target object. And the weight value of this edge is equal to k.

B. LABEL PROPAGATION PROCESS

After the homogeneous network is constructed, the label propagation method can be performed. This method will be introduced in this section. At the beginning, every node of homogeneous network will get a unique label. Then, label propagation and selection should be performed and repeated for several rounds. In each round, all the nodes will transmit its own label to direct neighbors. And every node will be assigned a new label, which is got from labels of its neighbors. When the label of each node is not changed, label propagation and selection procedure will be terminated. Now all the target objects of HIN have got its final label. So, at last, the target objects that share same labels should be put into same cluster.

Obviously, during the label propagation procedure, each node will receive several labels from its neighbors. Deciding which label should be adopted is a very critical step. Thus the label selection process is introduced here. While performing label selection, every label will get a value. Then label selection task can be fulfilled by using these values. As noted above, each edge of the homogeneous network is assigned a weight value. This weight value can present how close two target objects are. So, while we try to calculate a value for each neighbor's label of a target object, the weight value of the edge connected them should be considered. In order to provide a better description, let's take bibliography network as an example. In this network, four types of nodes must be contained, that are authors, papers, conferences and terms. Suppose that all the authors are selected as target objects. Then the meta-paths needed to be considered are $author-paper-author$, $author-paper-conference-paper-author$ and $author-paper-term-paper-author$. If two authors are connected by n instances of the first meta-path, it means that these two authors are coauthors of n different papers. While two authors are coauthors of many papers, they are very likely to focus on same research field. If two authors are connected by the instances of the second meta-path, it means that these two authors have submitted papers to the same conference. Generally, conferences are concentrated in a specific field. Thus the authors who have submitted paper to the same conference must focus on identical research field. If two authors are connected by the instances of the third meta-path, it means that the papers written by these two authors have same keywords. As keyword can represent the features of a specific research field, it is likely that the authors who are connected by the instances of the third meta-path have similar research direction. From above description, we can find that, the more instances of these three meta-paths exist between two authors, the more similar research directions of these two authors are. So, in the homogeneous network, if two authors are connected by an edge and the weight value of the edge is very large, these two authors are very likely to be put into same cluster. So, while a node calculates a value for the label of its neighbor, the weight value of the edge that connect these two nodes should be considered.

While calculating value for each label, two cases should be considered. In the first case, a neighbor transmits a unique label to the target object. It means that this target object doesn't receive identical label from other neighbors. Suppose that target object A has t different neighbors. These neighbors are $A_1, \ldots, A_t$. And the i-th neighbor A_i transmits a unique label $label_i$ to target objects A. $weight\_value_i$ is weight value of the edge between A and A_i. In this situation, the value of $label_i$ should equal to $weight\_value_i$. In the second case, more than one neighbors of a target object transmits same label to this target object. Suppose that m neighbors of target object A transmit label $label_{use}$ to A. These neighbors are $A_j, A_{j+1}, \ldots, A_{j+m-1}$. The weight values of edges between A and these neighbors are $weight\_value_j, weight\_value_{j+1}, \ldots, weight\_value_{j+m-1}$. Then the value owned by label $label_{use}$ is computed as below:

$$value_{label_{use}} = \sum_{t=j}^{j+m-1} weight\_value_t. \tag{3}$$

After obtaining labels of all the neighbors and the corresponding values of these labels, each target object should make a choice. It depends on the value of the label. If the value of $label_j$ is greater than the value of all the others, $label_j$ will be selected as the new label of the target object, that is, the label with largest value will be chosen.

From above description, we can find that label selection is executed after label transmission. During the label transmission procedure, each node tries to transmit its label to its neighbors. Apparently, this process can be executed in a concurrent manner. While label transmission procedure is finished, each node will get several labels. Now they should select its new label. As the label selection procedure is isolated, this process also can be executed simultaneously. Thus, the label transmission and label selection can be fulfilled concurrently by using multi-thread technique.

The suedo-code of the proposed algorithm is as following:

Algorithm 1 NodeVecClus Algorithm

```
Input: HIN: $G(V, E)$, Meta-path Set: $P^r_{m=1}$
Output: Clustering Result Set: $res$
Construct a homogeneous network based on heterogeneous information network $G(V, E)$ and all the meta-paths by using the method described above
count = 1
while count > 0 do
    for each node of homogeneous network do
        The node transmits label to its neighbors
    end for
    for each node of homogeneous network do
        Calculate a value for every obtained label by using the method stated above
        Select the most suitable label as its new label based on the value of all the labels
    end for
    Count number of the nodes that label has been changed and assign this number to variable $count$
end while
Target objects with same label will be put into same cluster
Store the clustering result into set $res$ and return set $res$
```

C. TIME COMPLEXITY ANALYSIS

Obviously, running time of the proposed algorithm mainly depends on the label propagation process. The more times label propagation process repeats, the longer the proposed algorithm runs. As every node of homogeneous network need to collect its neighbors' labels in each cycle, the running time of each cycle mainly depends on the number of nodes contained in the homogeneous network. Suppose that the label propagation process will repeat d times and m nodes are contained in the network. Time complexity of the proposed algorithm is $O(d \cdot m)$.

IV. EXPERIMENTS

A. DATA DESCRIPTION

In this experiment, four data sets are used to evaluate the proposed algorithm. These four data sets are DBLP data set, Yelp-b data set, Yelp-r data set and Yelp-s data set. All the data sets will be introduced in detail.

DBLP computer science bibliography is an open bibliographic information service website. This website contains plenty of papers that are published in major computer science journals and proceedings. DBLP data set is extracted from this website, in which all the papers of the website can be found. As data volume of the data set is very large, it should be preprocessed. During the procedure, papers that are published in eight top conferences will be kept. These conferences are SIGMOD, VLDB, ICDM, SIGIR, TREC, ACML, COLT, SIGKDD. Then all the authors of these papers can be obtained. These authors are sorted in descending order based on the number of papers they published in these conferences. After that the data of top 1000 authors should be used to form the preprocessed data set. Now, based on the preprocessed DBLP data set, a heterogeneous information network can be generated. During the procedure, author names, conference names, paper titles and key terms contained in the paper should be extracted. These objects will be treated as nodes of HIN. And all the authors are treated as target objects. As eight conferences selected in this experiment mainly focus on four different fields, including data mining, machine learning, information retrieval and database, the tested clustering algorithms should try to put all the authors into four different clusters. Moreover, in this experiment, 3 meta-paths are used, including $A-P-A$, $A-P-C-P-A$ and $A-P-T-P-A$. A represents authors. P represents papers. C represents conferences. And T represents terms.

Yelp is the largest review website of the United States. It contains information of stores that are located in different places. And these stores belong to different industries. In this experiment, three data sets that are extracted from this website will be used. These data sets are Yelp-b, Yelp-r and Yelp-s.

- **Yelp-b:** Stores contained in this data set belong to four different industries, which are Food, Shopping, Health & Medical and Beauty & Spas.
- **Yelp-r:** This data set contains great amount of restaurant information. The stores of this data set can be classified into four categories, including Chinese, Thai, Mexican and Italian.
- **Yelp-s:** Plenty of shopping markets can be found in this data set. The stores contained in this data set also belong to four categories, which are Eyewear & Opticians, Books, Mags, Music & Video, Sporting Good and Home & Garden.

We can construct heterogeneous information networks based on Yelp-b, Yelp-r and Yelp-s separately. During the process, four kinds of objects will be extracted, which are store names, comments of the stores, customers who wrote the comments and the key terms contained in the comments. All these objects are treated as nodes. As a store name can be used to represent a specific store, all the store names will be selected as target objects. According to the above introduction, each data set contains four kinds of stores. So the tested algorithm should try to put target objects of each data set into four different clusters. Moreover, two meta-paths of these three HIN are used in the experiment. These two meta-paths are $S-R-U-R-S$ and $S-R-T-R-S$. S represents stores. R represents reviews. U represents visitors of stores. And T represents terms of reviews.

The information of HIN constructed based on data sets DBLP, Yelp-b, Yelp-s and Yelp-r are presented in Table 1. Table 1 contains the node number, edge number, average degree and target object number of each HIN.

The original data obtained from DBLP website and Yelp website are not in the form as we expected. So some data preprocessing are needed. The data set collected from DBLP website are in xml format. It contains great amount of

TABLE 1. Information of data sets DBLP, Yelp-s and Yelp-r.

dataset	Node Number	Edge Number	Average Degree	Target Object Number
DBLP	15098	81451	11	1000
Yelp-s	28232	367596	26	749
Yelp-r	19332	242016	25	164
Yelp-b	42391	475150	18	1600

article information. So, these data should be filtered and some important objects should be extracted. During the process, papers of specific authors that are published in eight designated conferences will be kept. Objects and edges related to these papers will be extracted. The extracted objects are author names, paper titles, conferences that publish these papers. The extracted edges are (paper,author) and (paper, conference). To fulfil the task, module named minidom should be used. It is provided by Python. After paper titles are obtained, word segmentation should be executed. Then the key terms contained in the title and all the edges (paper, term) can be extracted. These objects and edges should be stored in different files. The date set collected from Yelp website are in JSON format. So the JSON package provided by Python will be used to extract information. While we try to construct data set Yelp-r, shop names of four kinds of restaurants will be extracted firstly. Then we can extract related comments, visitors who wrote the comments and key terms contained in the comments based on the shop names. The edges should be extracted at the same time. These edges are (shop,review), (shop,visitor) and (shop,term). Then all the objects and edges will be stored in different files. Data sets Yelp-b and Yelp-s can be constructed in the same way. After these four data sets are obtained, the tested algorithms can be executed. These algorithms try to put target objects of each data set into several different clusters. This experiment is performed by using a PC which contains Core-i7 CPU. The memory of the PC is 16G. The version of Python is 3.8.0. Several python packages are used in the experiment, including networkx, igraph, numpy and scipy etc.

B. BASELINES AND EVALUATION METHODS

In this experiment, the proposed algorithm will be compared with three other algorithms. These three algorithms are NetClus algorithm [21], PathSelClus algorithm [22] and SCHAIN algorithm [26]. In NetClus algorithm, object ranking is considered. This algorithm tries to generate model for target objects by using ranking distributions of other nodes. PathSelClus algorithm is proposed based on meta-path. It takes the structural relationship of target objects into consideration. SCHAIN algorithm is also proposed based on meta-path. But, it takes attributes of all the target objects into consideration at the same time.

While performing experiment, three metrics are used, including accuracy value, Rand Index(RI) value and running time. Accuracy value is used to measure how much instances are put into the correct cluster. The more instances are put into the correct cluster, the higher accuracy value this algorithm will get. RI value is used to measure how close the calculated clustering result and the real clustering result are. Obviously, these two metrics are very important for measuring clustering quality of the tested algorithms.

TABLE 2. Experimental results based on DBLP dataset.

Metrics	NetClus	PathSelClus	SCHAIN	HINLPClus
RI	0.479	0.705	0.405	0.693
accuracy	0.127	0.738	0.525	0.702

TABLE 3. Experimental results based on Yelp-s dataset.

Metrics	NetClus	PathSelClus	SCHAIN	HINLPClus
RI	0.451	0.559	0.366	0.479
accuracy	0.171	0.360	0.529	0.551

C. CLUSTERING QUALITY STUDY

In this section, the clustering quality of the proposed algorithm will be tested and compared with other clustering algorithms which are introduced above. This experiment will be performed based on four data set, including DBLP, Yelp-b, Yelp-r and Yelp-s. Table 2 presents the experimental results based on DBLP data set. Table 3 presents the experimental results based on Yelp-r. Table 4 presents the experimental results based on Yelp-s. And Table 5 presents the experimental results based on Yelp-b. In these tables, HINLPClus represents the proposed algorithm. Every algorithm of this experiment should run five times based on each data set. And final result of an algorithm is the average value of five different result values.

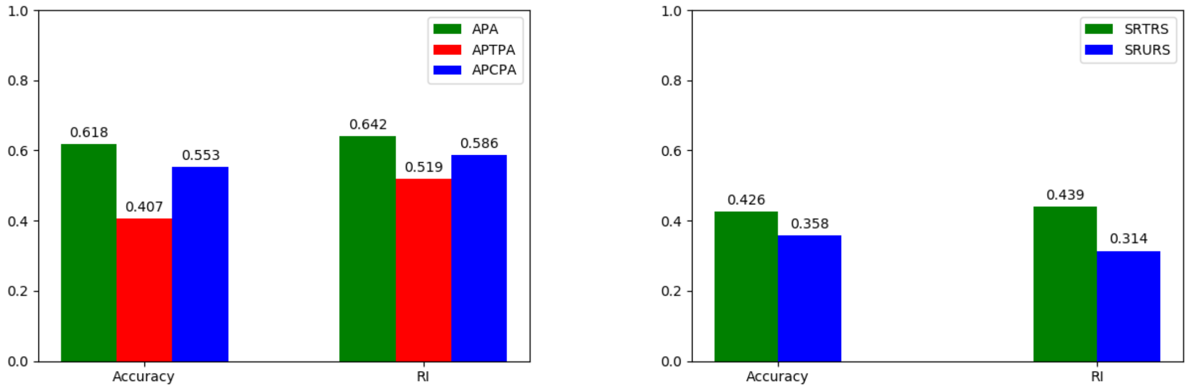
TABLE 4. Experimental results based on Yelp-r dataset.

Metrics	NetClus	PathSelClus	SCHAIN	HINLPClus
RI	0.375	0.562	0.356	0.512
accuracy	0.441	0.518	0.475	0.493

TABLE 5. Experimental results based on Yelp-b dataset.

Metrics	NetClus	PathSelClus	SCHAIN	HINLPClus
RI	0.318	0.536	0.342	0.508
accuracy	0.401	0.573	0.487	0.560

Obviously, the proposed algorithm outperforms NetClus algorithm and SCHAIN algorithm in all cases. But, in the experimental results based on DBLP data set, Yelp-b data set and Yelp-r data set, clustering effect of PathSelClus algorithm is better than the proposed algorithm. It is probably because different meta-paths can impact the clustering result in different level. In PathSelClus algorithm, each meta-path will get its weight value. If a meta-path can depict the relationship of target objects better than other meta-paths, the weight value of this meta-path is higher. Thus it will have more influence on final clustering result. In the proposed algorithm, even though the homogeneous network is also generated based on



(a) Results based on DBLP

(b) Results based on Yelp-r

FIGURE 2. Meta-path importance comparison.

the meta-paths, the difference of meta-paths is not considered. Thus all the meta-paths are equally important to the final clustering result. In this situation, the meta-path that can depict the relationship of target objects better will not have more influence on final clustering results. Moreover, from the experimental results, we can find that SCHAIN algorithm is not better than other algorithms. It is because SCHAIN algorithm takes attributes of target objects into consideration, but number of attributes contained in these three data sets are limited. Besides, from the experimental result, we can find that both RI value and accuracy value of the proposed algorithm are very high. It means that the proposed algorithm not only has high clustering accuracy, but the clustering result calculated based on the proposed algorithm is very close to the real clustering result.

D. META-PATH IMPORTANCE STUDY

In this section, the importance of different meta-paths will be studied. By doing so, several homogeneous networks should be generated. Each one is corresponding to a specific meta-path. While constructing a homogeneous network based on a meta-path t, all the target objects will be treated as nodes of homogeneous network. Then all the instances of meta-path t will be treated as edges of the homogeneous network. At last, every edge need to be assigned a weight value. The weight value is used to indicate how many instances exist between two target objects. So, if m different meta-paths are contained in HIN, m homogeneous networks will be generated. After getting these m homogeneous networks, the label propagation process will be performed based on each of them. Then, m clustering results will be got. At this time, the influence of all the meta-paths can be compared based on these clustering results. If a clustering result is better than others, the meta-path that related to this clustering result is more important than other meta-paths. Obviously, this meta-path can better reflect the features of HIN than others.

The experimental results presented in figure 2 are obtained based on two data sets, which are DBLP data set and Yelp-r data set. The metrics used in this experiment are accuracy value and RI value. Obviously, in the experimental results based on DBLP, we can find that the clustering results obtained based on the homogeneous network of meta-path $A-P-A$ is better than others. It means that homogeneous network constructed based on meta-path $A-P-A$ can depict the relationship of target objects better than others. Thus the importance of meta-path $A-P-A$ is the highest. So authors that appear in the same paper are more likely to focus in the same research field. As the accuracy value and RI value of meta-path $A-P-C-P-A$ is higher than $A-P-T-P-A$, the importance of meta-path $A-P-C-P-A$ is higher than meta-path $A-P-T-P-A$.

In the experimental results based on data set Yelp-r, the accuracy value and RI value corresponding to the meta-path $S-R-T-R-S$ is higher than the accuracy value and RI value corresponding to the meta-path $S-R-U-R-S$. So the importance of meta-path $S-R-T-R-S$ is higher than $S-R-U-R-S$. It means that stores that share common terms are more likely to be put into same cluster.

E. EFFICIENCY STUDY

In this section, the running time of the proposed algorithm will be tested and compared. The results are presented in Figure 3. HINLPClus represents the proposed algorithm. Every algorithm should run five times based on each data set. Then five running time values will be got. Thus every final running time value presented in Figure 3 are average value of five running time values.

In the results presented in figure 3, we can find that the running time of the proposed algorithm is shorter than the running time of the other three compared algorithms. It is because running time of the proposed algorithm is mainly decided by convergence speed of label propagation procedure. As the convergence speed of label propagation procedure is very fast, the proposed algorithm runs faster than other algorithms. Obviously, the running time of the proposed

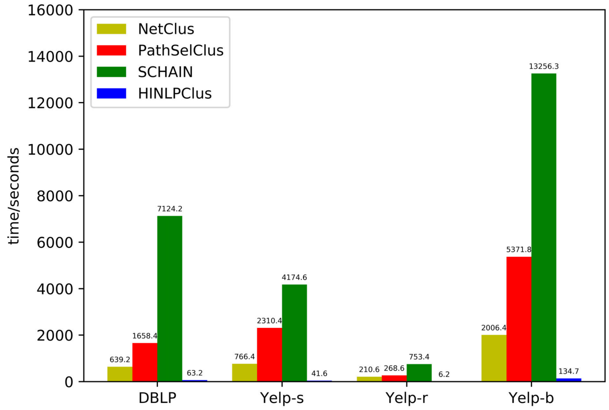


FIGURE 3. Running time of tested algorithms.

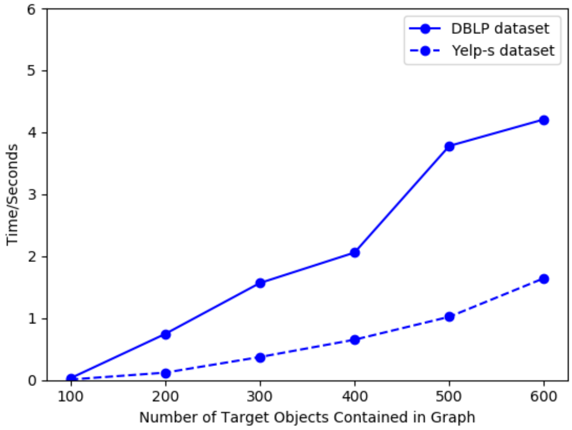


FIGURE 4. Scalability of the proposed algorithm.

algorithm is also influenced by the number of target objects. So running time of the proposed algorithm based on Yelp-b data set is longer than others. Running time based on DBLP is the second longest. Running time based on Yelp-s is longer than the running time based on Yelp-r.

F. SCALABILITY STUDY

In this section, the scalability of the proposed algorithm will be tested and presented. DBLP data set and Yelp-s data set are used. In this experiment, the data volume of each data set will be increased gradually. The target object number of each data set is set to 100, 200, 300, 400 and 500 successively. The experimental results are presented in figure 4. And each running time value presented in figure 4 is also average value of five different running time values. So the proposed algorithm should runs five times based on each specific data set with designated number of target objects. In figure 4, different curves are used to represent different data sets. The solid line represents the results obtained based on DBLP data set. The dotted line represents the results obtained based on Yelp-s.

From the experimental results presented in figure 4, we can find that the running time of the proposed algorithm increases linearly while the sample size is becoming larger. And, obviously, increasing rate is very slow. Because the slope values of these two curves are less than 1. From the results, we can find that the running time increasing rate of the proposed algorithm based on DBLP data set is larger than the increasing rate based on Yelp-s data set. It is because the edge number of homogeneous network constructed based on DBLP is larger, while the target object numbers of these two data sets are equal.

V. CONCLUSION

A community detection algorithm is proposed in this paper for heterogeneous information networks. As this algorithm is based on label propagation, the complicated optimization procedure and matrix computation procedure contained in the traditional community detection algorithms are eliminated. Thus the running time of the proposed algorithm is shorter than the compared algorithms. But the proposed algorithm needs to be further improved. One shortage of this algorithm is that it doesn't take the relative importance of meta-paths into consideration. As different meta-path has different level of influence on final clustering results, weight value should be assigned to each meta-path. In this situation, the relative importance of each meta-path can be presented.

REFERENCES

[1] D. L. Banks and N. Hengartner, "Social networks," in *Encyclopedia of Quantitative Risk Analysis and Assessment*, vol. 4. 2008.

[2] W. Gao, H. Wu, M. K. Siddiqui, and A. Q. Baig, "Study of biological networks using graph theory," *Saudi J. Biol. Sci.*, vol. 25, no. 6, pp. 1212–1219, 2018.

[3] E. W. Zegura, K. L. Calvert, and M. J. Donahoo, "A quantitative comparison of graph-based models for internet topology," *IEEE/ACM Trans. Netw.*, vol. 5, no. 6, pp. 770–783, Dec. 1997.

[4] T. Ramraj and R. Prabhakar, "Frequent subgraph mining algorithms–A survey," *Proc. Comput. Sci.*, vol. 47, pp. 197–204. Jan. 2015.

[5] J. B. Lee, R. Rossi, and X. Kong, "Graph classification using structural attention," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, London, U.K., Jul. 2018, pp. 1666–1674.

[6] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Washington, DC, USA, 2010, pp. 243–252.

[7] M. A. Javed, M. S. Younis, S. Latif, J. Qadir, and A. Baig, "Community detection in networks: A multidisciplinary review," *J. Netw. Comput. Appl.*, vol. 108, pp. 87–111, Apr. 2018.

[8] N. Duhan, A. K. Sharma, and K. K. Bhatia, "Page ranking algorithms: A survey," presented at the IEEE Int. Adv. Comput. Conf., Patiala, India, 2009.

[9] H.-J. Li, L. Wang, Y. Zhang, and M. Perc, "Optimization of identifiability for efficient community detection," *New J. Phys.*, vol. 22, no. 6, Jun. 2020, Art. no. 063035.

[10] X. Zeng, W. Wang, C. Chen, and G. G. Yen, "A consensus community-based particle swarm optimization for dynamic community detection," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2502–2513, Jun. 2020.

[11] L. Wu, Q. Zhang, C.-H. Chen, K. Guo, and D. Wang, "Deep learning techniques for community detection in social networks," *IEEE Access*, vol. 8, pp. 96016–96026, 2020.

[12] L.-E. Martinet, M. A. Kramer, W. Viles, L. N. Perkins, E. Spencer, C. J. Chu, S. S. Cash, and E. D. Kolaczyk, "Robust dynamic community detection with applications to human brain functional networks," *Nature Commun.*, vol. 11, no. 1, pp. 1–13, Jun. 2020.

[13] V. Moscato, A. Picariello, and G. Sperlí, "Community detection based on game theory," *Eng. Appl. Artif. Intell.*, vol. 85, pp. 773–782, Oct. 2019.

[14] K. Berahmand and A. Bouyer, "LP-LPA: A link influence-based label propagation algorithm for discovering community structures in networks," *Int. J. Mod. Phys. B*, vol. 32, no. 6, Mar. 2018, Art. no. 1850062.

[15] K. Berahmand, A. Bouyer, and M. Vasighi, "Community detection in complex networks by detecting and expanding core nodes through extended local similarity of nodes," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 4, pp. 1021–1033, Dec. 2018.

[16] K. Berahmand, S. Haghani, M. Rostami, and Y. Li, "A new attributed graph clustering by using label propagation in complex networks," *J. King Saud Univ.-Comput. Inf. Sci.*, to be published.

[17] C. Shi, B. Hu, W. X. Zhao, and P. S. Yu, "Heterogeneous information network embedding for recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 2, pp. 357–370, Feb. 2019.

[18] H. Wang, F. Zhang, M. Hou, X. Xie, M. Guo, and Q. Liu, "SHINE: Signed heterogeneous information network embedding for sentiment link prediction," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, Marina Del Rey, CA, USA, 2018, pp. 592–600.

[19] L. Hu, Y. Wang, Z. Xie, and F. Wang, "Semantic preference-based personalized recommendation on heterogeneous information network," *IEEE Access*, vol. 5, pp. 19773–19781, 2017.

[20] Y. Sun, J. Han, P. Zhao, Z. Yin, and H. Cheng, "Rankclus: Integrating clustering with ranking for heterogeneous information network analysis," in *Proc. 12th Int. Conf. Extending Database Technol., Adv. Database Technol.*, 2009, pp. 565–576.

[21] Y. Sun, Y. Yu, and J. Han, "Ranking-based clustering of heterogeneous information networks with star network schema," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2009, pp. 797–806.

[22] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu, "Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks," *ACM Trans. Knowl. Discovery Data*, vol. 7, no. 3, p. 11, Aug. 2013.

[23] T. Fu, W. C. Lee, and Z. Lei, "HIN2Vec: Explore meta-paths in heterogeneous information networks for representation learning," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Singapore, 2017, pp. 1797–1806.

[24] Y. Du, W. Guo, J. Liu, and C. Yao, "Classification by multi-semantic meta path and active weight learning in heterogeneous information networks," *Expert Syst. Appl.*, vol. 123, pp. 227–236, Jun. 2019.

[25] G. Wan, B. Du, S. Pan, and G. Haffari, "Reinforcement learning based meta-path discovery in large-scale heterogeneous information networks," in *Proc. AAAI Conf. Artif. Intell.*, New York, NY, USA, 2020, pp. 6094–6101.

[26] X. Li, Y. Wu, M. Ester, B. Kao, X. Wang, and Y. Zheng, "Semi-supervised clustering in attributed heterogeneous information networks," presented at the Int. Conf. Int. World Wide Web Conf. Steering Committee, Apr. 2017. [Online]. Available: http://papers.www2017.com.au.s3-website-ap-southeast-2.amazonaws.com/proceedings/p1621.pdf

[27] Y. Zhou and L. Liu, "Social influence based clustering of heterogeneous information networks," presented at the ACM SIGKDD Conf. Knowl. Discovery Data Mining, Aug. 2013. [Online]. Available: https://www.cc.gatech.edu/ lingliu/papers/2013/ACMKDD-SI-Cluster.pdf

[28] Y. Sun, C. C. Aggarwal, and J. Han, "Relation strength-aware clustering of heterogeneous information networks with incomplete attributes," in *Proc. VLDB Endowment*, Istanbul, Turkey, vol. 2012, pp. 394–405.

[29] S. Zhou, J. Bu, Z. Zhang, C. Wang, L. Ma, and J. Zhang, "Cross multi-type objects clustering in attributed heterogeneous information network," *Knowl.-Based Syst.*, vol. 194, Apr. 2020, Art. no. 105458.

[30] D. Liu, L. Li, and Z. Ma, "A community detection algorithm for heterogeneous information networks," *IEEE Access*, vol. 8, pp. 195655–195663, 2020.

[31] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 76, no. 3, 2007, Art. no. 036106.

DONGJIANG LIU received the B.S. degree from the College of Computer and Information Engineering, Inner Mongolia Agricultural University, in 2011, the M.S. degree from the School of Computer and Information Technology, Beijing Jiaotong University, in 2014, and the Ph.D. degree from the Computer Network Information Center, Chinese Academy of Sciences, in 2019.

He is currently working with the College of Data Science and Applications, Inner Mongolia University of Technology. His current research interests include data mining, graph mining, and machine learning.

LEIXIAO LI received the B.S. and M.S. degrees from the College of Information Engineering, in 2004 and 2007, respectively, and the Ph.D. degree from the College of Computer and Information Engineering, Inner Mongolia Agricultural University, in 2019.

He is currently working with the College of Data Science and Applications, Inner Mongolia University of Technology. His current research interests include big data mining and cloud computing.