Abstract

Purpose: This paper aims to provide a method to detect research communities based on research interest in researcher network, which combines the topological structure and vertex attributes in a unified manner.

Design/methodology/approach: A heterogeneous researcher network has been constructed by combining multiple relations of academic researchers. Vertex attributes and their similarities were considered and calculated. An approach has been proposed and tested to detect research community in research organizations based on this multi-relation researcher network.

Findings: Detection of topologically well-connected, semantically coherent and meaningful research community was achieved.

Research limitations: The sample size of evaluation experiments was relatively small. In the present study, a limited number of 72 researchers were analyzed for constructing researcher network and detecting research community. Therefore, a large sample size is required to give more information and reliable results.

Practical implications: The proposed multi-relation researcher network and approaches for discovering research communities of similar research interests will contribute to collective innovation behavior such as brainstorming and to promote interdisciplinary cooperation.

Originality/value: Recent researches on community detection devote most efforts to single-relation researcher networks and put the main focus on the topological structure of networks. In reality, there exist multi-relation social networks. Vertex attribute also plays an important role in community detection. The present study combined multiple single-relational researcher networks into a multi-relational network and proposed a structure-attribute clustering method for detecting research community in research organizations.

Keywords Community detection; Multi-relation social network; Semantic association
1 Introduction

A social network is a social structure made up of a set of actors (such as individuals, groups and organizations) and relationships between them. Social network analysis examines the structure of social relationships in a group to reveal informal connections between people, which facilitate members within an organization to locate expertise, identify new communities of practice, develop cross-functional knowledge sharing and improve strategic decision-making across leadership teams. A primary research area in the social network analysis is community mining. Here, the community is defined as a subset of nodes which are densely connected to each other and loosely connected to the rest of a network. Despite a lot of interdisciplinary researches, the identification of a community structure in networks remains an open problem. Some main reasons could be iterated as follows;

- Most social networks studied so far are homogeneous networks, that is, only one relationship between entities has been taken into consideration. In reality, there exist more than one relationships in a social network. For example, academic researchers may have different kinds of relations such as co-authoring one or more papers, participating in the same conference, being members of the same scientific center, joining the same project, and so on. Thus, a social network of academic researchers is multi-relational. It is a great challenge to uncover the structure in multi-relational networks.
- Most community mining techniques focused on topological structures by using various criteria such as normalized cut\(^1\), modularity\(^2\), structural density\(^3\) or flows\(^4\). Vertexes are clustered by measuring vertex closeness based on connectivity or structural similarity. These methods did not consider vertex attributes in the clustering process.

In the present study, we proposed a method combining multiple relations of academic researchers to construct a heterogeneous network. Four types of relations had been taken into consideration, including similarities, social relations, interactions and flows. Vertex attributes had been added to the heterogeneous network. Similarities between attributes were calculated, and topologically well-connected, semantically coherent and meaningful research community detection has been achieved based on our multi-relation researcher network.

2 Related studies on community detection

In general, a community in a network refers to a group of nodes with greater internal ties than the rest of the network. The strength of ties can be reflected in two ways: The edge closeness of the nodes from a topological view and the attribute similarities
of nodes. A community is expected to have a cohesive topological structure and members with semblable attributes.

A wide variety of algorithms based on structure similarity have been developed to detect network communities in social networks. Previous endeavors fall into two categories: Heuristic methods and optimization based methods[5].

Heuristic methods usually lead to algorithms based on intuitive assumptions. One of these examples is GN Algorithm[2], which assumes that edges with high edge betweenness scores are more likely to connect different communities. Therefore, a number of disjoint communities can be obtained by iteratively removing edges with the highest scores. Other heuristic algorithm examples include MFC algorithm[6], CPM algorithm[7], FEC algorithm[8] and WH algorithm[9]. The optimization based methods include spectral methods and local search methods.

The goal of spectral methods is to minimize the defined cut-function. Local search based method aims to optimize an objective function such as the function of conductance[10] and modularity. Newman[11] proposed a greedy agglomerative clustering algorithm for optimizing modularity. The vertices are merged to form a larger community if it increases the modularity.

For optimizing modularity, many strategies have been developed such as greedy techniques[12,13], simulated annealing[14], extremal optimization[15], spectral optimization[16,17] and other optimization strategies[18,19]. Modularity is by far the most popular and best known quality function for measuring the quality of a partition of a network. However, some researchers believe that modularity optimization has a resolution limit that may prevent it from detecting clusters which are comparatively small with respect to the graph as a whole, even when they are well defined communities[20]. Another drawback is that the communities detected by these approaches tend to contain topically-diverse sub-communities within each community[21].

Attributes of nodes are also important in community partition. Attributes are feature description of nodes. It could be continuous or discrete values, and most likely textual description or documents. Communities detection can be based on attribute similarities of nodes using traditional unsupervised methods such as hierarchical clustering[22] and k-means[23]. Based on textual information, researchers have applied topic models such as LDA[24,25] to find communities as well as the associated topics[26]. However, above methods did not take into account the topological structure and may result in nodes being grouped into one cluster with only few connections.

Communities are supposed to be groups of vertices that are not only well connected but similar to each other. Therefore, a challenge to improve the quality of community detection is to make full use of both relationship and attribute information.
Steinhaeuser & Chawla\cite{27} presented a simple approach which construct a network with edge weights based on node attributes and cluster nodes whose edge weight exceeds the threshold in the same community. Li et al.\cite{28} developed a scalable community discovery solution for large-scale document corpus. A set of preliminary community cores were identified through a relation topology analysis. Afterwards, attributes were used to improve community consistency. Non-relevant documents, though connected with the community, are eliminated.

Zhou et al.\cite{29} proposed a novel graph clustering algorithm, SA-Cluster, which achieved a good balance between structural and attribute similarities through. Attribute vertices and edges were added to the original graph to get an augmented one with vertices sharing attribute values connected. The proposed graph clustering method is based on a unified random walk distance rather than a linear combination of structure and attribute distance.

A weight self-adjustment method was also used to learn the degree of contributions of different attributes in random walk distances. Dang & Viennet\cite{30} proposed two algorithms, SAC1 and SAC2, which use both structural and attribute information to extract more meaningful communities. SAC1 is based on the modification of modularity function using a composite modularity as a weighted combination of modularity structure and modularity attribute. SAC2 is a two-phase algorithm, which constructs a $k$-nearest neighbor graph with the similarity measure defined as a linear combination of attribute similarity and link strength, then finds structural communities in the graph to obtain final clustering.

Zhao et al.\cite{31} proposed a two-step topic-oriented community detection approach which combines both social objects clustering and link analysis. A subspace clustering algorithm was used to group all the social objects into topics. Members involving in those social objects were divided into topical clusters, with each corresponding to a distinct topic.

Approaches based on both structure and attribute similarities have provided more meaningful communities. However, most algorithms for community mining only consider the homogeneous network with single type of relationship between vertexes. In reality, social networks are normally heterogeneous with different types of nodes and relations. A network with different types of nodes is called multi-modes network; a network with different types of relations between the same sets of nodes is called multi-dimensional network.

Identifying communities in heterogeneous networks is very challenging. The present study focuses on the heterogeneous network with homogeneous nodes (researchers) and different types of relations. This article presents and analyzes the strategies to integrate four types of relations between researchers and exploit new ways to compute the similarities between researchers’ attributes. The aim of this
study is to divide communities with researchers in the same cluster being connected and semblable on research interest.

3 Clustering multi-relational social network

Researcher network is a typical multi-relational social network. These relationships indicate formal and informal communications between researchers, implying that they have one or more common research interests. Therefore, the assumption of the present study is that the quality of community detection might be improved if all these relations are combined into a single network.

Social network analysis provides a promising technique for exploring formal and informal communication between researchers. The aim of the present study in detecting communities is to find clusters of nodes which are both structurally and semantically close-connected in the multi-relation social network. It is necessary to integrate data of individual attributes with those of interpersonal relations.

3.1 Framework

The framework for research community detection is illustrated in Fig. 1. The whole process for detecting research community from multi-relation researcher network is based on four modules.

Fig. 1 The framework for detecting research communities from multi-relation researcher network.
Research community detection from multi-relation researcher network based on structure/attribute similarities

- Multi-relations extraction. This module aims to extract four basic types of relations and propose a model combining multiple social relations between researchers. The strength of each relation is calculated in this section.
- Attributes similarity calculation. According to the attributes of nodes, the semantic similarities are calculated between node attributes.
- Community detection. A heterogeneous network is built according to the four relations. Based on the heterogeneous network, researchers are partitioned into different research communities by using WGN algorithm.
- Community optimization. The similarity of node attributes is used to optimize the community detection results. Results are then adjusted by calculating the modularity of network and entropy for each community.

3.2 Multi-relations extraction

In theory, social relationships are viewed in terms of nodes and relations. There can be many types of relations between nodes. Borgatti et al. divided dyadic relations into four basic types including similarities, flows, interactions and social relations. There are exactly same relations in researcher networks.

Specifically, similarities include spatial and temporal proximity as well as co-membership in groups and events and share socially significant attributes, such as race or class. Similarity refers to the similar degree of the characteristic properties of associated members, rather than a kind of relationship between them. For example, they are located in the same place, interested in the same thing or are experts in a certain field. Similarities include shared recognition. Cognitive is one of the characteristics of a researcher, which can be represented as a set of keywords used by researchers in his/her publication. The cognitive similarities between two researchers can be calculated as the coupling degree of two sets of keywords.

Flows are tangible or intangible things that are transmitted through interactions, such as information and resources. In researcher networks, flows can be citations between researchers, since citations are tangible evidences that resource/idea flows.

Interactions are typically conceptualized as discrete events that can be counted over a period of time. They can be viewed as being facilitated by and occurring in the context of social relations. For example, friends (social relation) give each other advice, family members (social relation) chat with each other (interaction), colleagues (social relation) cooperate with each other (interaction). In researcher networks, it refers to collaborations (interaction) between given researchers (social relation), that is, the frequency that they join in the same projects or be co-authors of the same titles.

Social relations are the canonical types of ties that most sociological theorizing about social networks is based on. Social relations are conceptualized as continuous
properties, such as family relationships, friendships, relationships between teachers and students, colleagues, and so on. In researcher networks, social relations correspond to colleague relationships, i.e. if the researchers belong to the same affiliations.

Based on the detailed analysis of the four types of relations between researchers, a multi-relational social network has been constructed. A researcher network can be described as a graph model, \( G = (V, E_i) \), where \( V \) is the set of nodes (researcher), and \( E_i (i = 1, 2, 3, 4) \) indicates the four types of edges between two nodes. The weights of edges are separately calculated and systematically integrated.

The strength of similarities was calculated by the keywords coupling frequency in research articles written by researchers. For example, let \( K_A \) be the keywords set of articles published by researcher \( A \), and \( K_B \) the keywords set of articles published by researcher \( B \), if there exists \( x \) keywords occurrence both in \( K_A \) and \( K_B \), then the strength of the similarities between the two researchers \( (A \ and \ B) \) is \( x \).

The strength of flow was measured by the number of citation. If researcher \( A \) cites researcher \( B \)'s papers \( m \) times and \( B \) cites \( A \)'s papers \( n \) times, then the strength of the flow was \( m+n \).

The strength of interaction was calculated by the times of the collaboration between researchers. If researchers \( A \) and \( B \) worked together in \( x \) projects and wrote \( y \) books, the strength of interaction between the two nodes \( (A \ and \ B) \) was \( x+y \).

The strength of social relationship was measured by their affiliation. For example, if researcher \( A \) and researcher \( B \) are in the same research institute, the strength of social relationship between two nodes \( (A \ and \ B) \) was 1, and 0 otherwise.

The total strength of the node connection is obtained using Eq. (1).

\[
w = z_1w_1 + z_2w_2 + z_3w_3 + z_4w_4, \tag{1}
\]

where \( w_1, w_2, w_3 \) and \( w_4 \) are weights for similarity, social relationship, interaction and flow of each edge, respectively.

Different weight coefficients \( (z_1, z_2, z_3, z_4) \) are set according to the importance of each relation with respect to the researcher network. Since the cognitive similarity is the most important factor reflecting the similar research interests, \( z_1 \) should be given the maximum weight. Citation relationships can also reflect the relevance of research interests, \( z_2 \) should be set relatively high. Cooperation may only partially reflect the similarity since it is not necessary for people with different expertise to share the same research interest even they participate in the same project. So \( z_3 \) should be given relatively a low weight. Researchers in a same affiliation may have same study areas, but it does not guarantee that their research interests are the same, so \( z_4 \) should be given a minimal weight value. A constraint condition can then be obtained: \( z_1 > z_2 > z_3 > z_4 \). The most important association is the cognitive association,
followed by the citation association, cooperation association, and structure association. After testing, the weighting factors \((z_1, z_2, z_3, z_4)\) are set to 0.4, 0.3, 0.2, and 0.1, respectively.

### 3.3 Attributes similarity

Current community detection methods mainly focus on structural features and have not taken the property of nodes into account. However, attributes can provide semantic information which can be used to improve clustering process and the results. In the researcher network, attributes of nodes can be expertise or research interest of each researcher. Researchers are grouped into same clusters if they have similar research interest. Expertise or research interest of each researcher can be represented by keywords extracted from associated papers using the TF-IDF method. Since most researchers have 2 or 3 research fields (concept) and normally 3 or 4 keywords are related to one concept, the top ten keywords with the highest frequency are selected to represent individual’s research interest. Weights of keywords are normalized. Therefore, the graph model described above is modified as \(G = (V, E, X)\), where \(X\) is the set of attributes of nodes \(V\), for example the attribute of node \(V_j\) is represented as \((k_{j1}(u_{j1}), \ldots, k_{j10}(u_{j10}))\).

In the present study, a method was proposed to calculate the similarity between two nodes by considering the semantic similarity between keywords. Given two nodes \(V_i\) and \(V_j\), attributes of both nodes were expressed as: \(X_i(k_{i1}(u_{i1}), \ldots, k_{i10}(u_{i10}))\) and \(X_j(k_{j1}(u_{j1}), \ldots, k_{j10}(u_{j10}))\). In order to avoid synonyms problems, semantic similarities were calculated between two keywords \(Sim(k_{im}, k_{jn})\).

Explicit semantic analysis\(^{[33]}\) is utilized to calculate the semantic relatedness between keywords. ESA is an approach to computing semantic relatedness of natural language texts with the aid of very large scale knowledge repository such as Wikipedia. It represents the meaning of a word as a weighted vector of Wikipedia-based concepts. The semantic similarities between keywords were obtained by calculating the cosine of the angle value of the concept vectors. Then the attribute similarity between nodes \(V_i\) and \(V_j\) was calculated using Eq. (2):

\[
Sim(V_i, V_j) = \sum_{n=1}^{10} u_{im}u_{jm} \cdot Sim(k_{im}, k_{jn}).
\]  

The similarity of node attributes is used to optimize the result of the community detection, which is described in detail in the next section.

### 3.4 Community detection and optimization

The aim of community detection is to find partitions of nodes which are structurally and semantically close to each other. In the present study, a method was proposed
to detect communities by using both structural and attribute information. In this method, the modularity was calculated as an initial division. Then node attributes similarity was used to optimize the community detection result.

Many algorithms have been developed for detecting network communities in social networks. A typical algorithm (GN algorithm) was proposed by Newman and Girvan[2] for discovering community structure. The usefulness of the GN algorithm had been proved in unweighted networks. Indicated by weights information, close connection or similarity is useful in making a more accurate determination of the communities. Based on GN algorithm, the weighted GN (WGN) algorithm has been proposed[34].

WGN algorithm calculates the betweenness of all edges in weighted graph without considering edge weights. It divides betweenness according to weight of corresponding edges and removes them with the highest score. Computation of betweenness is repeated for all edges until the network is divided into a number of appropriate communities. Fig. 2 shows how the WGN works (Algorithm 1).

Algorithm 1: Weighted GN Algorithm

```plaintext
Required G, G^°
1: Q_{max}←0
2: while G is not empty
3: Q_{v}, W_{v}, B_{v}, B_{max}←0
4: for b_{v} in G do
5: b_{v}←b_{v}/ find weight of b_{v} in G^°
6: if b_{v} > B_{max} then
7: B_{max}←b_{v}
8: B_{v}←b_{v}
9: end if
10: end for
11: remove B_{v} from G
12: Q_{v}←recompute modularity of G
13: if Q_{v} > Q_{max} then
14: Q_{max}←Q_{v}
15: else
16: break
17: end if
18: end while
19: return G
```

Algorithm 2: Entropy Optimization Algorithm

```plaintext
Required C, imax,
1: H_{0}←H_{0}
2: i←0
3: while i<imax and more possible changes do
4: i←i+1
5: A←random cluster from C
6: x←random node : x \in A
7: A(x, -)
8: B←random cluster from C \{DA\}
9: B(x, +)
10: H_{i}←H_{i}
11: if H_{i} > H_{i-1} then
12: B(x, -)
13: A(x, +)
14: end if
15: end while
16: return C_{i} \{A new partition with a reduced entropy\}
```

Fig. 2 Algorithms for community detection.
Research community detection from multi-relation researcher network based on structure/attribute similarities

The modularity refers to the number of edges falling within groups minus the expected number in an equivalent network with random edges. The modularity is calculated by Eq. (3):

\[ Q_w = \sum_{s=1}^{k} \left( \frac{W_s}{W} - \left( \frac{S_s}{2W} \right)^2 \right), \tag{3} \]

where \( W_s \) and \( S_s \) are the weights of all edges and nodes of the community \( s \), respectively. \( W \) represents the total weight of all edges. The network can then be divided into a number of communities by using the modularity optimization strategy.

Entropy describes the average Shannon information content of a set. A highly disordered set always has high entropy. When it comes to data mining, the entropy of a group is used to describe the similarity of its elements. A group with similar elements has low entropy, that is, this group is more ordered.

For any two nodes \( V_i, V_j \) in the network, the attribute semantic similarity is obtained by using Eq. (2). After the preliminary community detection, a group of cluster \( C \) has been obtained. The entropy of the group \( H(C_k) \) is calculated by Eq. (4)\(^{[35]}\):

\[ H(C_k) = -\sum_{i=1}^{[C_k]} \sum_{j=i+1}^{[C_k]} \text{Sim}(V_i, V_j) \ln \text{Sim}(V_i, V_j) + (1 - \text{Sim}(V_i, V_j)) \ln(1 - \text{Sim}(V_i, V_j)), \tag{4} \]

where \( \text{Sim}(V_i, V_j) \) is the attribute semantic similarity between nodes \( V_i \) and \( V_j \). The entropy of a cluster is minimized when its nodes are semantically similar. Note that the maximum entropy can be obtained when all nodes are grouped into the same cluster. Given a partition, the entropy is optimized by using a Monte-Carlo approach\(^{[36]}\). Then, a random node \( n \) is removed from a random community \( A \) and inserted into another random community \( B \), such that the node does not belong to \( A \)'s neighbor set. If the change does not improve the entropy, the node is returned to its original community. This is repeated until all nodes have been tested and no further changes can improve the entropy or a maximum number of iterations are reached. Algorithm 2 in Fig. 2 shows how the entropy is optimized.

4 Experiments

In the present study, performance of the proposed cluster method has been tested and evaluated by investigating research communities within the School of Information Management (SIM), Wuhan University, China. There are six departments in the School and 72 staffs involving in scientific researches. These researchers may form different research communities. Three types of weighted network – cognitive
network, multi-relationships network and multi-relationships network with node attributes were built. All algorithms were implemented in java and compiled using JDK.

4.1 Datasets and preprocessing

4.1.1 Similarity association dataset

The cognitive association is reflected by word pair co-occurrence, which has been collected from a total number of 2,989 papers published by researchers since 1980s. All these papers were downloaded from CNKI and Weipu databases.

Meaningful attributes such as title, author, keywords and citations were counted. Other attributes such as PageCount, PubTime were excluded in the present experiment.

In the keywords set, there were 6,512 keywords after removing duplicated items. Each paper has two keywords on average. A total number of 3,448 articles have been analyzed, including 3,138 co-authored articles. In these papers, 346 of them were written by two or more researchers, 2,792 articles were written by a teacher and his/her students. If two researchers share a pair of keywords in one of their papers, an edge can be added. For co-authored papers, the connection has been measured by the number of keywords in a given paper, that is, if there are \( x \) keywords in this paper, the weight is set to be \( x \).

4.1.2 Social relations dataset

The social relations dataset is derived from the hierarchy structure of the organization which shows the structure relationship of its members. If two researchers belong to the same department such as Electronic Commerce or Information Science, a connection is then established and the weight is set to be 1. There are 561 links in total in this dataset, and one department contributes 93.5 links on average.

4.1.3 Interaction dataset

Two types of relationship are included in the interaction dataset. One relationship is the co-author relationship of books. During 1982 to 2010, 112 books have been written on collaboration. Authors of these books cover 65.3% of researchers in SIM. The other relationship is the collaborative relationship in research projects. A total number of 68 projects have been approved and financed in the latest 9 years. There are 217 links established in the cooperation dataset.

4.1.4 Flows dataset

The flow dataset is derived from the reference database of CNKI. For every article of each researcher, the quotations can be retrieved from the CNKI. According to the
Research community detection from multi-relation researcher network based on structure/attribute similarities

quotations records, the citation relationship among the teachers can be extracted. If researcher $A$ cites researcher $B$’s papers $m$ times and $B$ cites $A$’s papers $n$ times, then the strength of the flow was $m+n$. In total, 594 citation relationships were established in this dataset.

4.2 Results

- Cognitive network. This network is generated from the similarity association dataset. And the cognitive relationship is embodied by the co-occurrence of word pair. As mentioned above, the edge between two researchers was counted in two ways. The cognitive network is a network with weight.
- Multi-relationship network. In the multi-relationship network, the edge weight between any two researchers is the sum of the weights of the corresponding entities in each single-relational networks multiplied by some coefficient, which shows the importance degree of the data source.
- Multi-relationship network with node attributes. Based on multi-network of relationships with weight, the vertex attribute is added. Top ten keywords were extracted from each keyword set of the researcher’s papers as the vertex attribute. The similarity between these attributes is calculated through the ESA method.

Taking researchers HKQ and QJP as examples (Table 1), although some keywords are literally different, there are still semantic connections between these words. ESA method is used to calculate the semantic similarity between two keywords (Table 2). Then the attribute similarity between researchers HKQ and QJP is calculated by using Eq. (2), the final result is 0.549.

<table>
<thead>
<tr>
<th>Keywords (weight)</th>
<th>HKQ</th>
<th>QJP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informatization (1)</td>
<td>Bibliometrics (1)</td>
<td></td>
</tr>
<tr>
<td>Computer network (0.67)</td>
<td>Informatics (0.444)</td>
<td></td>
</tr>
<tr>
<td>Publishing (0.67)</td>
<td>University rankings (0.869)</td>
<td></td>
</tr>
<tr>
<td>Digitizing (0.67)</td>
<td>Webometrics (0.695)</td>
<td></td>
</tr>
<tr>
<td>Electronic publishing (0.67)</td>
<td>Knowledge management (0.695)</td>
<td></td>
</tr>
<tr>
<td>Information technology (0.17)</td>
<td>Link analysis (0.608)</td>
<td></td>
</tr>
<tr>
<td>Printing press (0.17)</td>
<td>Comparative research (0.521)</td>
<td></td>
</tr>
<tr>
<td>Network Technology (0.17)</td>
<td>Citation analysis (0.478)</td>
<td></td>
</tr>
</tbody>
</table>

4.3 Clustering analysis

Initially, we analyzed the clusters divided from the cognitive network. The method is based on the weighted clustering method WGN. Ten groups have been extracted. There are still four isolated nodes (Fig. 3(a), SGX, WSS, NJ, LYJ). Then, the same
### Table 2  Semantic similarity of different keywords

<table>
<thead>
<tr>
<th></th>
<th>Informatization</th>
<th>Bibliometrics</th>
<th>Computer network</th>
<th>Informatics</th>
<th>Publishing</th>
<th>University rankings</th>
<th>Electronic publishing</th>
<th>Webometrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informatization</td>
<td>0.0012</td>
<td>0.0093</td>
<td>0.0026</td>
<td>0.8265</td>
<td>0.0169</td>
<td>0.0353</td>
<td>6.7229</td>
<td></td>
</tr>
<tr>
<td>Bibliometrics</td>
<td>0.0012</td>
<td>0.0026</td>
<td>0.0243</td>
<td>0.0060</td>
<td>0.0863</td>
<td>0.0056</td>
<td>0.2140</td>
<td></td>
</tr>
<tr>
<td>Computer network</td>
<td>0.0071</td>
<td>0.0040</td>
<td>0.0143</td>
<td>0.0270</td>
<td>0.0017</td>
<td>0.0017</td>
<td>0.2140</td>
<td></td>
</tr>
<tr>
<td>Informatics</td>
<td>0.8265</td>
<td>0.0040</td>
<td>0.0026</td>
<td>0.0143</td>
<td>0.0270</td>
<td>0.0017</td>
<td>0.0017</td>
<td></td>
</tr>
<tr>
<td>Publishing</td>
<td>0.0169</td>
<td>0.0060</td>
<td>0.0243</td>
<td>0.0143</td>
<td>0.0453</td>
<td>0.1178</td>
<td>0.0063</td>
<td></td>
</tr>
<tr>
<td>University rankings</td>
<td>0.0353</td>
<td>0.0863</td>
<td>0.0086</td>
<td>0.0270</td>
<td>0.0090</td>
<td>0.1819</td>
<td>0.0181</td>
<td></td>
</tr>
<tr>
<td>Electronic publishing</td>
<td>0</td>
<td>0.0056</td>
<td>0.0078</td>
<td>0.0017</td>
<td>0.1178</td>
<td>0.0090</td>
<td>0.0018</td>
<td></td>
</tr>
<tr>
<td>Webometrics</td>
<td>6.7229</td>
<td>0.2140</td>
<td>0.0026</td>
<td>9.8309</td>
<td>0.0063</td>
<td>0.1819</td>
<td>0.0018</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3  Clusters divided from the cognitive network (a) and multi-relationships network (b).
method is used to divide the multi-relationships network with or without node attributes. Fig. 3(a) shows the divided groups based on the cognitive network using only the similarity relations. Most researchers are connected for word co-occurrence. There are still some orphan nodes for sharing no keywords with other nodes. Fig. 3(b) shows clusters divided from the Multi-relationships network. The whole network is divided into 10 clusters. New relations were revealed in the social network and some researchers were categorized to newly identified groups. This occurred due to multiple relations between these researchers. These results demonstrated that many alternative knowledge flows in multi-relational social network are not represented in the homogeneous social network. It can also be concluded that the multi-relational social network has a better knowledge flow than the homogeneous social networks.

Table 3 compares members in multi-relation researcher network with or without node attributes. There is no significant difference between the two networks. A small number of researchers are underlined because they are grouped differently. For example, researcher ZYF was classified into the same community (Cluster 8) with researcher SGX in multi-relation researcher network without node attributes because the two researchers collaborated in two projects.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Members in network without node attributes</th>
<th>Members in network with node attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WYG, YYC, ZMJ, XLF, FQ, ZJW, WQ, LZC, HXR</td>
<td>WYG, YYC, ZMJ, XLF, FQ, ZJW, WQ, LZC, HXR</td>
</tr>
<tr>
<td>2</td>
<td>SEM, ZXJ, WXG, ZRY, QJP, CY, MFC</td>
<td>SEM, ZXJ, WXG, ZRY, QJP, CY, MFC</td>
</tr>
<tr>
<td>3</td>
<td>MJ, ZM, ZLY, ZZM, LFL, WL, TXQ, ZY, <strong>YSY</strong></td>
<td>MJ, ZM, ZLY, ZZM, LFL, WL, TXQ, ZY</td>
</tr>
<tr>
<td>4</td>
<td>WJX, ZN, XCR, NJ, CCY, <strong>YSY</strong></td>
<td>WJX, ZN, XCR, NJ, CCY, YSY</td>
</tr>
<tr>
<td>5</td>
<td>HSH, <strong>LJZ</strong>, LG, JYY, DH, LW, LQ</td>
<td>HSH, LG, JYY, LL, DH, LW, LQ</td>
</tr>
<tr>
<td>6</td>
<td>DJ, TXB, SL, DZH, MDC, <strong>ZYM, CH</strong></td>
<td>DJ, TXB, SL, DZH, MDC, LP, KJH</td>
</tr>
<tr>
<td>7</td>
<td>SL2, PFZ, XXM, YL, CCF, HRH, <strong>SMCJ</strong></td>
<td>SL2, PFZ, XXM, YL, CCF, HRH, ZYF</td>
</tr>
<tr>
<td>8</td>
<td>WD, SGX, CZ, QXL, LYJ, HKQ, <strong>ZF</strong></td>
<td>WD, SGX, CZ, QXL, LYJ, HKQ, SMCJ</td>
</tr>
<tr>
<td>9</td>
<td>ZYL, ZJX, ZYY, WXJ, XQH, YH</td>
<td>ZYL, ZJX, ZYY, WXJ, XQH, YH, <strong>WSS</strong></td>
</tr>
<tr>
<td>10</td>
<td>LR, ZYF2, HCP, DSL, <strong>LP, KJH, LL</strong></td>
<td>LR, ZYF2, HCP, DSL, LJZ, ZYM, CH</td>
</tr>
</tbody>
</table>

It is interesting to find that researcher ZYF was classed into same community (Cluster 7) with researcher SL2 in multi-relation researcher network with node attributes. Although researchers ZYF and SL2 have only one cooperated project, the keywords “information organization” and “information service” are found in the attribute information of these two researchers. In fact, their research interests are more similar.

A variety of quality functions or measures have been used to capture the goodness of a division of a graph into clusters[37]. In our experiment modularity and entropy were chosen to evaluate the quality of the community structure. Modularity is widely used as a measure for how good a clustering is. One of the advantages of
modularity is that it is independent of the number of clusters that the graph is divided into. The intuition behind the definition of modularity is that the farther the subgraph corresponding to each community from a random subgraph, the better or more significant the discovered community structure. The modularity $Q$ for a division of the graph into $k$ clusters $\{C_1, \ldots, C_k\}$ is given by Eq. (3). Entropy was chosen because it could measure the average Shannon information content of a set, which indicates how similar the elements are within each cluster. The entropy of the group is given by Eq. (4).

Three types of researcher network are compared in terms of modularity and entropy (Table 4). The multi-relation researcher network without attributes has the highest modularity, while that with node attributes has the lowest entropy. Since the modularity and the entropy behave oppositely, the optimizing process can be a multi-objective problem. It is necessary to tune values to obtain best results for both measures.

<table>
<thead>
<tr>
<th>Index of cluster quality</th>
<th>Cognitive network</th>
<th>Multi-relation researcher network without attributes</th>
<th>Multi-relation researcher network with node attributes (optimized)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modularity</td>
<td>0.144</td>
<td>0.465</td>
<td>0.323</td>
</tr>
<tr>
<td>Entropy</td>
<td>94.8646</td>
<td>73.8746</td>
<td>35.3649</td>
</tr>
</tbody>
</table>

Table 5 shows the primary topic for each cluster of the optimized multi-relation researcher network with node attributes. The cluster contain 10 groups of researchers working on “Publishing”, “Informetrics and evaluation”, “E-commerce”, “Information visualization”, “Information retrieval”, “Information system”, “Library information organization”, “Documentation & digital library”, “Archival science”, and “Information service”. Researchers are clustered by keywords, co-author/co-membership, and shared research interest. It should be pointed out that broad research interests of some researchers may result in scattering research topic of the entire community and subsequently affect the accuracy of results.

<table>
<thead>
<tr>
<th>Cluster details</th>
<th>Members in clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1: Publishing</td>
<td>WYG, YYC, ZMJ, XLF, FQ, ZJW, WQ, LZF, HXR</td>
</tr>
<tr>
<td>Cluster 2: Informetrics &amp; evaluation</td>
<td>SEM, ZXJ, WXG, ZRY, QJP, CY, MFC</td>
</tr>
<tr>
<td>Cluster 3: E-commerce</td>
<td>MJ, ZM, ZLY, ZMM, LFL, WL, TXQ, ZY</td>
</tr>
<tr>
<td>Cluster 4: Information visualization</td>
<td>WJX, ZN, XCR, NJ, CCY, YSY</td>
</tr>
<tr>
<td>Cluster 5: Information retrieval</td>
<td>HSH, LG, JYY, LL, DH, LW, LQ</td>
</tr>
<tr>
<td>Cluster 6: Information system</td>
<td>DJ, TXB, SL, DZH, MDC, LP, KJH</td>
</tr>
<tr>
<td>Cluster 7: Library information organization</td>
<td>SL2, PFZ, XXM,YL, CCF, HRH, ZYF</td>
</tr>
<tr>
<td>Cluster 8: Documentation &amp; digital library</td>
<td>WD, SGX, CZ, QXL, LYJ, HKQ, SMCJ</td>
</tr>
<tr>
<td>Cluster 9: Archival science</td>
<td>ZYL, ZXJ, ZYY, WXC, XQH, YH, WSS</td>
</tr>
<tr>
<td>Cluster 10: Information service</td>
<td>LR, ZYF2, HCP, DSL, LJZ, ZYM, CH</td>
</tr>
</tbody>
</table>
5 Conclusion

Community detection is extremely useful in finding communities with implicit knowledge. Most reports were based on homogeneous network with single-relationship and only utilized the structural information in clustering process. A variety of social relations between people and the semantic information of vertex have been ignored. In the present study, a model was proposed combining multiple social relations between researchers and a heterogeneous network was built. A community detection method was also developed to couple topological structure as well as attribute information in detecting the community. Experimental results demonstrated that the proposed approach has achieved flexibility in combining structural and attribute similarities, hence be able to bring in more meaningful communities. The proposed approach can improve the efficiency of collaborative learning, experts finding and knowledge sharing for each topic, and make full use of collective intelligence. It can also be applied to different types of social networks containing social objects.

Although the proposed method has been proved to be functional, there are still some limitations. Firstly, the sample size of evaluation experiments was relatively small. In the present study, a network made up of 72 researchers was analyzed for detecting research community. A large sample size is required to give more information and reliable results. Secondly, in reality, relationships between researchers are more complex and diversified. Although the four types of relationships were extracted between researchers to build a network in this study and obtained reasonable results, it is not sufficiently comprehensive and to complete summary of all relationships between the researchers. Thirdly, the associated strength is a variable which is very hard to quantify. In this study, the definitions and measurement of the four types of relations correspond to common keywords, scientific collaboration, citations, etc. Because of the large time span, there are difficulties in collecting data, such as changing jobs and personnel changes, which can influence the accuracy of results. Therefore, future work will focus on better quality evaluation of the divided communities and research topic detection of each community. It is also necessary to develop new strategy to integrate semantic information into a social network.

References


Research community detection from multi-relation researcher network based on structure/attribute similarities


Reference Citation Format

The format for citations in text and for bibliographic references follows GB/T 7714—2005. The citation should be ordered in number as it appears in the text of the submitted article.

- **For journal article**

- **For book**

- **For proceedings**

- **For electronic journal article**

- **For electronic book**

- **For thesis or dissertation**
Chinese Journal of Library and Information Science

(CJLIS, Chinese Title: 《中国文献情报》)

ISSN 1674-3393, CN 11570/G2;  Price: $US 199/Volume

Chinese Journal of Library and Information Science (CJLIS) is the first and only English-language academic journal in the field of Library and Information Science (LIS) that has been published in Mainland China so far. It aims to establish a platform for students, researchers and professionals in LIS and archive science to engage in intellectual dialogs and to share best professional practices of their experience.

CJLIS adopts a double-blind peer-review process for each published article. The Editorial Board of CJLIS is composed of well-known scholars both from national and international LIS fields. The main aim of CJLIS is to serve librarians, information professionals, researchers and practitioners in the LIS field for their various purposes and keep them well informed of the latest development of LIS field both in China and the world.

CJLIS mainly publishes original research articles, including research papers, library practice papers, progress reports, research reviews and perspectives. For subscription, you can use the following channels:

1) Via EBSCO International, Inc (USA)

2) To order at any post office inside China, with a distribution code 82-563.

3) To contact with the Editorial Office
   Address: No.33 Beisihuan Xilu, Zhongguancun, Haidian District, Beijing 100190, P.R. China
   Tel: 86-10-82624454 or 86-10-82626611 ext. 6628
   Fax: 86-10-82624454
   Email: chinalibraries@mail.las.ac.cn
   Website: http://www.chinalibraries.net

4) Via Beijing China Science Journal Publishing Co., Ltd.
   Address: Room 301, Building B, Desheng International Center, No. 83 Deshengmenwai Street, Beijing 100088, P.R. China
   Tel: 86-10-59790736 ext. 8028
   Fax: 86-10-59790736 ext. 8092
   E-mail: sunzhipeng@mail.sciencep.com
Submission Guidelines

Aims
Chinese journal of Library and Information Science (CJLIS), being sponsored by the Chinese Academy of Sciences (CAS) and published quarterly by the National Science Library of CAS, is a scholarly journal in the field of library and information science (LIS). Its aim is to provide an international communication link between researchers, educators, administrators, and information professionals.

With the publication of the research results both from China and from other foreign countries, the Journal CJLIS strikes a balance between theory and practice. With its goal to provide an open forum for Chinese and international scholars in this field to exchange their research results, CJLIS also offers new possibilities in the advancement of Chinese library operations. The CJLIS tries to establish a platform for LIS students, researchers and library staff all over the world to engage in intellectual dialog and also to improve library services so as to promote even more quickened and substantial development of LIS in China.

Scope
Striving toward academic excellence, innovation, and practicality, the CJLIS mainly includes research papers both on the theoretical as well as on the practical fronts in all aspects of the field. More specifically, it includes but not limited to informatics, library management, information technology application, knowledge organization system, knowledge management, archives, permanent preservation of library resources, LIS education, and so on.

Refereeing Process
Articles and papers covering the topics or themes mentioned above will be refereed through a double-blind peer review process.

Editorial Advisory Board
The Editorial Board is composed of the nationally and internationally well-known scholars and researchers in the LIS field and the high quality of this Journal is thus reasonably assured.

Manuscripts Categories
As the first English-language academic journal on LIS published in Mainland China, the CJLIS will take a proactive attitude to trace and report the prevailing hot issues in the field around the globe as well as the more serious scholarly communications. As such, the submitted manuscripts are classified into constant categories and unfixed categories. In the former category, research papers, library practice and progress reports are the essential components. In the latter, book reviews, biographical sketches, anecdotes, reminiscence of prominent librarians and brief communications will appear occasionally.

Research papers represent original research work or a comprehensive and in-depth analysis of a topic. More than 3,000 words are considered as a proper length for such manuscripts, with a structured abstract ca. 200 words.

Library practice covers the latest development and application in any segment of library field work and information service. The length of the manuscript is preferred to be more than 3,000 words, with a structured abstract ca. 200 words.

Progress reports reflect the projects result or research progress on the key topics of the library and information science. Submissions of articles to this section are expected to be comprehensive and analytical, which may deepen the understanding of the discussed issue and stimulate further researches on the topics, or give a new perspective on future technological applications. The manuscript length should be within 5,000 words, with a structured abstract ca. 200 words.

Manuscripts Requirements
All papers can be submitted either in English or in Chinese (or both) with a double-line space. For the assurance that all the materials of the to-be-submitted are included, please check the following:

Title. Please give a brief biographical introduction to all contributing authors and their research background on a separated paper. For a better organization of the paper, please use the headings and subheadings.

Authors and affiliations. Please do not forget to write down the mailing address of each and every article contributors.

References. Be sure all the references used should be cited properly in both in-text and in bibliography. Particular attention should be paid to the proceedings. Do not forget adding the name(s) of editors of the compilation, as well as the name of the publishers. For the detailed information, please request a copy of Reference Citation Format.

Copyright
All submitted papers normally should not have been previously published nor be currently under consideration for publication elsewhere. For all the materials translated or obtained from other published resources, they should be properly acknowledged. All copyright problems should be cleared without any legal entanglements prior to the publication.

Notes for Intending Submissions
A guide for authors and other relevant information, including submitting papers online, is available at the website of the Editorial Office of the CJLIS (http://www.chinalibraries.net). For any questions, you can e-mail the Office or directly to:

Prof. ZHANG Xiaolin
Editor-in-Chief of CJLIS
The CJLIS Editorial Office
National Science Library, Chinese Academy of Sciences
No.33 Beishuan Xilu, Zhongguancun, Haidian District,
Beijing 100190, P.R. China
Tel: 86-10-82624454 or 86-10-82626611 ext. 6628
Fax: 86-10-82624454
E-mail: chinalibraries@mail.las.ac.cn
Website: http://www.chinalibraries.net

Subscription
For single-copy subscription in China: RMB ¥ 200/Issue. For subscription outside of China, US $ 199/Volume yearly (including air shipping)
CONTENTS

Research Papers

1 Research on overlapping structures and evolution properties of co-citation network
Shiji CHEN & Xiaolin ZHANG

14 Research community detection from multi-relation researcher network based on structure/attribute similarities
Ping LIU, Fenglin CHEN, Yunlu MA, Yuehong HU, Kai FANG & Rui MENG

33 Dynamics and interactions of blog-based science communication: A case study of blogging activities in China related to 2011 Japanese seismic and nuclear power crisis
Yuxian LIU, Li LIU, Ronald ROUSSEAU & Yang YANG

52 Implementation of a weblog extraction system with an improved template extraction technique
E CHANG

64 An integrated document retrieval method combining entity annotation and keyword index: A KIM platform implementation
Xin XU, Jialong GUO, Yunjia HONG & Biyi JIN

78 Developing the information-seeking behavior scale for undergraduates
Dehua HU & Dan CHE

97 Reference Citation Format

98 Enlightenment to Subscribe