doi:10.3772/j.issn.1006-6748.2016.04.011

Clustering approach based on hierarchical expansion for community detection of scientific collaboration network①

Li Xiaohui (李晓慧)②, Zheng Yanning
(Institute of Scientific and Technical Information of China, Beijing 100038, P. R. China)

Abstract

This study presents a clustering algorithm based on hierarchical expansion to solve the problem of community detection in scientific collaboration network. The characteristics of achievements information related to scientific and technological domains are analyzed, and then an ontology that represents their latent collaborative relations is built to detect clusters from the collaboration network. A case study is conducted to collect a data set of research achievements in the electric vehicle field and better clustering results are obtained. A hierarchical recommendation framework that enriches the domain ontologies and retrieves more relevant information resources is proposed in the last part of this paper. This work also lays out a novel insight into the exploitation of scientific collaboration network to better classify achievements information.

Key words: scientific collaboration network, clustering, achievements information, recommender systems

0 Introduction

Nowadays, scientific research and technological innovation have become a key factor to determine the competitiveness of country. Many countries have greatly increased the input of scientific and technological research. With increasing the number of scientific and technological achievements, information overload of scientific and technological achievements is becoming a ubiquitous problem in this era of big data.

Intelligent recommendation for scientific and technological achievements is to solve the problem of overload, create an information sharing environment, and maximize implement information management and service innovation. The recommender system has provided users more convenient and effective information acquisition experience from Internet. However, in order to recommend items to users who want the results, the system needs to find a recommendation candidates set from the mass of scientific and technological achievements. The work for clustering the mass of scientific and technological achievements and finding relevant results with the users needs is a crucial step. Accordingly, study of cluster mining algorithm in complex network plays a crucial role for personalized recommendation.

This study presents a clustering algorithm to divide the community into the scientific collaboration networks. An ontology that represents their collaborative relations network is built by analyzing the characteristics of achievements information related to the scientific and technological domains. Finally, a hierarchical recommendation framework of scientific and technological achievements information is proposed.

1 Related work

1.1 Scientific collaboration network

The entity is considered as nodes, relations between entities are considered as edges. Many systems in the real world can be presented in the form of network. In these networks, the relationship between entities is usually more complex, such as the small world, scale-free, self-organization and community structure characteristics, which simple network doesn't have many properties, so they are also called complex networks[1]. How to mine meaningful information efficiently in the complex network has become one of the important research content of multi-subjects fields.

In recent years, more and more scholars pay attention to the research of the complex network model in scientific collaboration network. At present, there are three aspects in the research of scientific collaboration network: influence of the individual to the global net-

① Supported by the National Social Science Foundation of China (No. 14CTQ045) and China Postdoctoral Science Foundation (No. 2015M570131).
② To whom correspondence should be addressed. E-mail: xhli@istic.ac.cn
Received on July 3, 2016

work generation model, influence of individual on local network community evolution, and prediction of the cooperative relationship in scientific research activities[2].

Scientific collaboration network is a complex network where nodes are scientists and links are co-authorships as the latter is one of the most good documented forms of scientific collaboration. It is an undirected and scale-free network where the degree distribution follows a power law with an exponential cutoff and most authors are sparsely connected while a few authors are intensively connected[3,4]. The network has an assortative nature hubs tend linking to other hubs and a low-degree nodes tend linking to low-degree nodes. Assortativity is not structural, meaning that it is not a consequence of the degree distribution, but it is generated by some process that governs the network's evolution.

Newman is one of the earliest researchers to study scientific collaboration network. He presented the most original logical connection and a paper contribution connection method to construct scientific collaboration network and make a detailed analysis[5,6]. Guimera and his research teams at the Northwestern University proposed a network evolution model for scientific collaboration network. This research found that there was a clear relationship between team diversity, cooperation, network structure, and team performance[7].

In addition, there are many other interesting studies to be measured on collaboration networks, including the number of collaborators of scientists, the number of papers they write, and the degree of "clustering", which is the probability of academic cooperation between the two scientists. Although the idea of constructing a scientific collaboration network from the academics record is not new, no detailed study has been published previously in the recommender system applications.

1.2 Clustering in collaboration network

In complex networks, a set of nodes, which are closely connected with the internal nodes and connected with the external nodes, is known as the community. In order to analyze the characteristics of community in the network, researchers have sought the community structure of complex networks through various methods. Community detection is a hot research topic in complex network. Community detection algorithm in complex network has a wide prospect of applications in many fields. For example, Internet network analysis and clustering search engine, social organization structure network analysis, personalized service and recommendation, etc.

As a typical complex network, collaboration network has some characteristics such as small world and no scale. At the same time, it also has the characteristics of community structure, that is, the whole network includes groups or clusters, within each cluster the link is relatively tight, and the link is loose between clusters. Compared with the construction of collaboration network, the study on community detection is more extensive. The existing researches mainly focus on the topology analysis, function analysis, identification of the hidden patterns and the prediction of network behavior.

Several different clustering algorithms were proposed in the past decades, such as graph partition, hierarchical clustering, and clustering in high dimensional data. Graph partition based clustering algorithms aim at dividing network graph collaboration into several subspaces. The partition condition is that the connection between nodes is tightly or closely in the subspaces. The optimal solution is found by procedure heuristic. Such as, Kernighan-Lin algorithm and Laplace eigen-matrix based Spectral Bisection algorithm[8]. Girvan and Newman proposed a groundbreaking modular algorithm that through distinction between edges of community to achieve community identification[9]. Flake analyzed community problems in the World Wide Web via maximum flow method[10].

Hierarchical clustering algorithms aim at dividing network into several clusters according to the similarity or the connection between nodes, which includes two kinds of implementation processes, a Divisive Method that removes edges from the network and an Agglomerative Method that adds edges to the network, such as GN algorithm[9] and CNM algorithm[11]. Hastie presented a hierarchical clustering algorithm to reveal the multi-level structure of graph. The hierarchical clustering algorithm in complex network analysis, biology, engineering and marketing is very common. Clustering approaches aim at the detection of clustered objects using all attributes in the full data space.

However, clusters in collaboration network are not in mutual isolation but overlapping. In other word, the nodes in the network belong to a number of clusters simultaneously. To solve this problem, researchers proposed the idea of cluster detection based on the module of cluster for the overlapping cluster[12].

2 Characteristics analysis of achievements information

To cluster the achievements information, the characteristics of information resources must be studied firstly. The definition of scientific and technological

achievements is the knowledge products that people have some kind of knowledge products, which have been recognized in the scientific and technological activities. It is a reflection of various kinds of information collection, which is a description of the connection and interaction between various research activities. Achievements information resources have the characteristics of diversity, heterogeneity, timeliness and semantic ambiguity.

2.1 Diversity and randomness of achievements information

With the rapid development of science and technology and the popularization of network, the achievements information published on the network shows a massive increase trend, and the source of achievements information is diverse. Since Internet is a non-central and decentralized management of interconnected institutions, it leads to the lack of effective management and organization for these achievements information. Therefore, the source of information contains a variety of sources, also the disorder. It is very difficult to make effective utilization for these information resources without effective organization and aggregation.

2.2 Heterogeneity of achievements information

Internet contains massive Web sites. The achievements information formations of these Web sites are different, such as structured data, semi-structured data and unstructured data. Because the achievements information resources from Internet in the matter of the data structure definition, syntax, semantic description, heterogeneity of access language in different systems result in that a large number of achievements information resources cannot be effectively used.

2.3 Dynamicity of achievements information

As publishers of achievements information according to their requirement and user demand, the publication is of no periodic and updates the achievements information, which adds to the dynamic and increases difficulty of information retrieval.

2.4 Semantic fuzziness of achievements information

Because of the different application fields of achievements information, the semantic representation of achievements information is different. Different users have different semantic understanding on the same achievements information. They have resulted in the ambiguity of the semantic of information resources. Therefore, it is necessary to solve the problem of semantic ambiguity in the same achievements information to improve the aggregation and sharing of achievements information resources. It will provide support for achievements recommendation.

3 Clustering approach of achievements information

As mentioned above, compared with the general information resources, it can be seen that achievements information resource is more complex and more difficult to handle. To cluster achievements information, the achievements information ontology is to try to build via analysis cooperation between researchers or institutions, construct collaboration network and clustering achievements based on the correlation degree of nodes. Then, clustering problem of achievements is transformed into community partition problem of network through the network node correlation degree. This study attempts to present a scientific collaboration network model to study the problem of community detection. The entire network is consisted of a number of groups or clusters. The internal nodes of each group are relatively close, and the connections between the groups are sparse.

3.1 Construction of scientific collaboration network model

The institutions or researchers that participate in scientific research cooperation are abstracted as the nodes of the network. The cooperative relationship between the institutions or researchers is abstracted as the network connection between the nodes, where edge means that the cooperation relationship, the weight of the edge representes the number of times of cooperation between institutions or researchers. Thus, research collaboration network model domain is preliminary constructed.

As shown in Fig. 1, scientific collaboration network is dynamic and the connection lines between the nodes are not fixed. With the passage of time gradually changing, these nodes continuing to accumulate will cause changes in the community structure of the network. Such as disappearance of community, the merger or community detection, and large or small the community size. The unpredictable change of community detection in the dynamic network is a huge challenge.

3.2 Class library

A scientific collaboration network including the modeling, analysis, visualization and other common abstract class library is established, which provides

Fig. 1 An example of collaboration network

operation, analysis and visualization for scientific collaboration network. The main types of research cooperation network include: individual, institutions, network, simulation, collaboration, statistics, and visualization.

Each class contains basic properties and functions, specific description as follows:

- Individual: ID is the unique identity of the individual.
- Institutions: ID is the research team's unique identifier. Contain the basic information of research team.
- Network: create a collaboration network.
- Simulator: connect all the individual or institutions nodes that have a cooperative relationship.
- Collaboration: record cooperative relationship between partners and their number of cooperative times.
- Statistics: the calculation of each sub collaboration network.
- Visualization: display network on the computer screen.

3.3 Clustering model in scientific collaboration network

The definition clustering model for achievements is explained in detail as follows:

1) Node degree D_i: Node degree D_i indicates the number of times of cooperation between nodes i.

2) Contribution of nodes to cluster $T_{(i,C)}$: $T_{(i,C)}$ indicates the correlation degree between node i and cluster C. The calculation formula is as follows:

$$T_{(i,C)} = \frac{D_{i,C}}{D_i} \tag{1}$$

Note that $D_{i,C}$ denotes the correlation degree between node i and cluster C. That is, the association strength of node i and all the nodes in cluster C.

3) Modularity(Q): Modularity is used to measure the quality of the Cluster structure produced by different cluster partition and it can be considered as the objective function of optimization. While $Q < 0$, it shows that the number of node connections in the cluster is less, so the effect is poor. When $Q \to 1$, it shows that the number of nodes connections in the cluster is good. The calculation formula of Modularity is

$$Q = e_c - a_c^2 \tag{2}$$

Note that e_c represents the ratio of the total number of edges in cluster C and the total number of edges in network. The calculation formula is as follows:

$$e_c = \frac{1}{2m}\sum_{ij} A_{ij}\delta(c_i, c)\delta(c_j, c) \tag{3}$$

Note that a_c represents the ratio of the total number of interconnected edges in cluster C and the total number of edges in network. The calculation formula is

$$a_c = \frac{1}{2m}\sum_{i} d_i\delta(c_i, c) \tag{4}$$

In Eqs(3) and (4), c_i indicates that node i belongs to cluster; m indicates the total number of edges in network; A_{ij} represents the adjacency matrix of graph; d_i represents the degree of node i. The calculation formula is explained as

$$A_{ij} = \begin{cases} S_{ij}, & \text{correlation between node } i \text{ and } j \\ 0, & otherwise \end{cases} \tag{5}$$

Note that S_{ij} represents the cooperation intensity between each side of the two nodes in the network, which indicates the number of times of cooperation between nodes i and nodes j in the last five years.

$$\delta(c_i, c) = \begin{cases} 1, & c_i = c \\ 0, & c_i \neq c \end{cases} \tag{6}$$

$$d_i = \sum_j A_{ij} \tag{7}$$

3.4 Detecting clusters in scientific collaboration network

The clusters detecting procedure is explained in detail as follows:

1) Initialize the nodes in cluster C: Compute correlation degree of all nodes in scientific collaboration network. Select the initial node in cluster and set it the minimum degree node $\min_{i \in N} D_i$. N is the set of remaining nodes in the network.

2) Calculate the cooperation intensity S_{ij} between each side of the two nodes in the network, that is, the strength of the connected.

3) Add nodes to cluster C: Compute the degree of contribution $T_{(i,C)}$ that all nodes are directly associated with the cluster C for clusters. Select the nodes that the minimum degree of contribution of $\max_i T_{(i,C)}$ to join

cluster C.

4) Dynamic extraction of clusters: Compute Modularity Q of cluster C. If Q reaches the maximum value, delete nodes and edges of cluster C from network; otherwise, return step 2) until Q reaches the maximum value. If isolated nodes are produced when cluster C is removed from scientific collaboration network, add the isolated nodes to cluster C, and remove it from the scientific collaboration network.

5) End: if the set of remaining nodes in network is not null, return step 1); otherwise, the clusters detecting process ends.

4 Case study

A collaboration network for institutions and researchers in electric vehicle field is constructed. The data come from SNAD. The cataloging information of 36,424 research achievements was collected in the dataset of electric vehicle area from 2005 to 2015. Each of the research achievements, including correlated its title, keywords, abstracts, project introduction, authors, patents, status, teams and their affiliated institutions, papers and publish year is obtained. The 1,070 distinct institutions as well as 7,821 cooperative relationships are extracted from the dataset. Collaboration network for institutions has formed a large cluster that means some close relationship between the core nodes. Also, there are few independent nodes in this network.

Some nodes are randomly selected as initial nodes from the test set, which are extended level by level. The algorithm presented in Section 3 is used to detect clusters in the collaboration network. For each of the institutions clusters, achievements clusters can be obtained from these institutions clusters. In other word, the correlation of achievements was found from the cooperative relationship between institutions in collaboration network.

In order to clearly observe the clustering results generated by the proposed algorithm, the visual D3 programming tool is used to simulate the process of hierarchical expansion. As shown in Fig. 2, the clustering results obtained by the two-level expansion are the

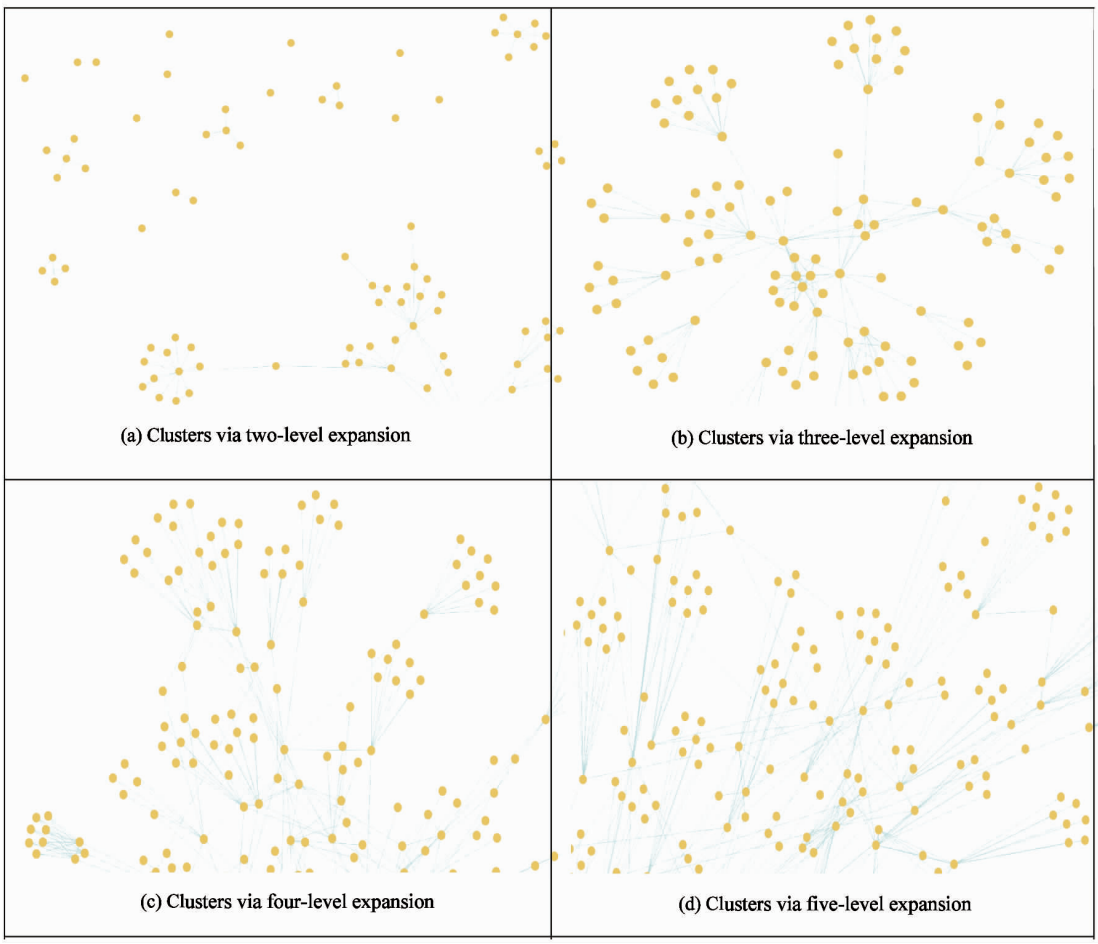

Fig. 2 An instance of institutions clusters in collaboration network

worst. The clustering results obtained by the three-level expansion are better than the results of the two-level expansion, but the number of clusters is small and the elements in each cluster are less. The clustering results obtained via the four-level and five-level expansion are significantly better than the results of the three-level expansion. However, through the computing time comparison, the computing time is significantly increased when the number of expansion level is greater than four-level. Therefore, the experimental results show that the four-level expansion is feasible and effective.

The cluster of institutions reflects the relatively stable cooperative relationship between them. Once these clusters are obtained, each institution has a number of achievements; the achievements clusters can be generated. Accordingly, the users can be also clustered via the user's access behavior. Each user cluster represents their common needs. For the recommendation service, the user's preferences and decisions are often influenced by the neighboring users. If clustering the user needs through the user's access behavior in the collaboration network, the right achievements information can be easily chosen and recommended to user. Thus, it will reduce the computational complexity of recommendation.

5 Achievements information recommendation framework

A scientific and technological achievements information recommendation framework, including three layers, the user and item retrieval layer, user and item management layer, recommendation service layer is designed, as shown in Fig. 3.

The user and item information retrieval layer is used to obtain the user information and the demand through the intelligent human-computer interface and obtain the related information from the Internet by the Web crawler technology.

The user and item information management layer is used to store the obtaining information of users and items. Construct user profile by classifying and clustering the user's information and build an ontology based on items information clusters. It will provide data support for the recommendation engine to produce the recommendation results.

The personalized service layer is used to recommend information for users. It is composed of the recommendation engine and the intelligent human-computer interface. The recommendation engine has two functions: first, analyzing the recommendation problem and applying the scene. The user's request can be transformed into the problem that can be solved directly by

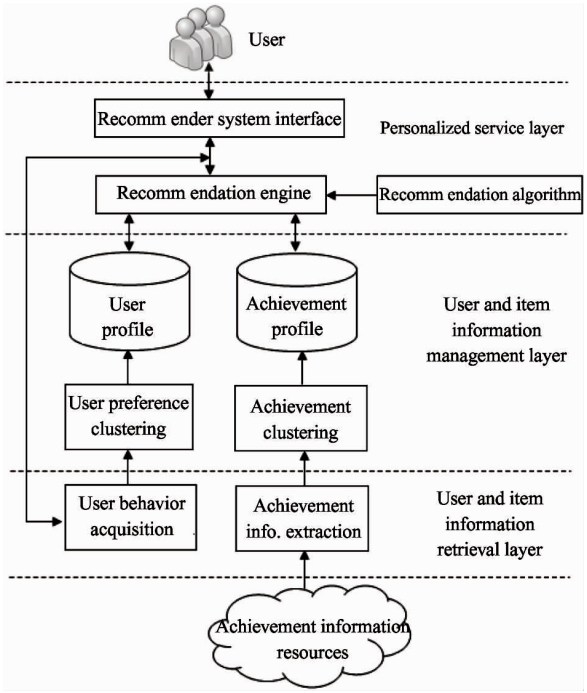

Fig. 3 Achievements recommendation framework

the recommendation algorithm. Second, returning the TOP-*N* information resources to the user from the results of recommendation algorithm via the intelligent human-computer interface. The intelligent human-computer interface is used to interact with the user and their response, including the collection of user's needs and preferences, and display the recommendation results.

Here, ontology is used to extract the institutions or researchers cooperative relationship from achievements information via automatically identifying and describing the information of all achievements information, and store them in the achievements profile database. Further more, the algorithm that introduced by previous section is used to cluster the different types of information achievements, and analyze their hierarchical structure and feature attributes. This construction of ontology will directly affect the result of the achievements recommendation.

6 Conclusion

A novel approach is deseribed which exploits domain knowledge regarding features of achievements. A hierarchical recommendation framework based on scientific collaboration network is presented to achieve a more rapid start-up to guide optimal choices for users. To address this, the collaboration networks of institutions were analyzed from achievements data of electric vehicle field in the last ten years. And the proposed

approach is used to detect clusters in the constructed collaboration networks. Taking institutions in electric vehicle field as a test area, preliminary experiments shows that the proposed algorithm generates results of the community divided by calculating node degree, the cooperation intensity between the nodes, degree of contribution and modularity, within the community closely linked, loose contact between the community so as to guarantee the accuracy of community detection.

This work is an on-going research project for detecting community structure in scientific collaboration network. Further experiments will improve the proposed algorithm to achieve better prediction accuracy with a lower empirical training time. Moreover, the obtained clusters from scientific collaboration network will provide interpretable results which are useful for achievements recommendation.

References

[1] Guimera R, Amaral L A N. Functional cartography of complex metabolic networks. *Nature*, 2005, 433 (7028): 895-900

[2] Ding Y. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, 2011, 5(1): 187-203

[3] Newman M E J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2000, 98(2): 404-9

[4] Barabasi A L, Jeong H, Neda Z, et al. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 2002, 311: 590-614

[5] Newman M E J. Detecting community structure in networks. *The European Physical Journal B*, 2004, 38(2): 321-330

[6] Newman M E J. Fast algorithm for detecting community structure in networks. *Physical Review E*, 2004

[7] Guimera R, Uzzi B, Spiro J, et al. Team assembly mechanisms determine collaboration network structure and team performa. *Science*, 2005, 308: 697-702

[8] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 1970, 291-307

[9] Girvan M, Newman M E J. Community structure in social and biological networks. *Proceedings of the National Academy Sciences*, 2002, 99(12): 6-15

[10] Flake G W, Giles C L, Coetzee F M. Self-organization and identification of Web communities. *IEEE Computer*, 2002, 35(3): 6-14

[11] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. *Physical Review E*, 2004

[12] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure of complex networks. *New Journal of Physics*, 2009, 11(3): 19-44

Li Xiaohui, is a postdoctoral fellow in Institute of Scientific and Technical Information of China. She received Ph. D. degree in System Engineering and Software Engineering from Waseda University, Japan, in 2013. Her research interests include intelligent recommender systems, science and technology information resource management.