Application of local community detecting Algorithm in citation network

Weijiang Wu, Dan Xue, Guohe Li

Department of Computer Science and Technology, China University of Petroleum, Beijing, 102200, China

email: allan1226@163.com

Keywords: citation network; DBLP; community detecting; research group; research direction

Abstract. Citation network is a complex network composed of literature and literature citation relationships. To such a large-scale network, it is difficult to detect the global community. On the other hand, there are more concerned with the situations local information than global information. Based on DBLP citation network, after cleaning data, a local community detecting algorithm is applied to find corresponding research group. The core literature and the community including it can be obtained by this algorithm. And then, the research direction and association leader of this community can be discovered according to the titles and authors of the core literature.

Introduction

With the development of modern civilization, the research of new subject usually requires researchers coming from different disciplinary fields. These researchers and its groups play good roles in promoting the development of cross disciplines. Scientific literature is the carrier of researchers' achievement. It is an evaluation of researcher's work. Massive literature citation relationships can build a large scale citation network which is a kind of complex network. Nowadays, with the help of computer, it is not a difficult thing to acquire this sort of materials.

After detecting a community from citation network, it is easy to find a research group by the authors' information included in this community. The research direction of this group is obviously according to the keywords of literatures included in this community. And then researchers can find groups with similar research direction which are their potential partners. From a macroeconomic perspective, based on the mining results of scientific research group, it is an effective way to prediction a new research direction by comparing to the original research topic. In the citation network, each literature plays different role in knowledge flowing. Some literatures are the core and others are not. Using citation relation, the communities with different research direction can be detected from citation network [1]. Then some important literatures can be found in these communities. These literatures can help us to know the importance of special knowledge.

Community structure of complex network

Real complex network is not a random network, furthermore, it has some organizational characteristics, such as small world [2], scale-free [3], aggregation and heterogeneity of distribution of degree of node. From the visual sense, community refers to a subset of network nodes, edges between nodes of inside the subset are dense, but the edges between nodes belonging to different subset are sparse. The community structure is shown as Figure 1. The visual of network community [4] shows that it is a node set consisted of nodes with common or similar attributes.

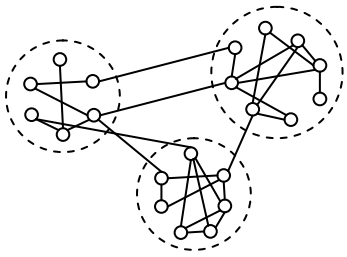

Figure 1 Community structure

How to detect community correctly is a hotspot of research in complex network. In order to find the community structure in complex network, researchers have proposed many community detecting algorithms, such as Spectral Bisection Algorithm [5-6], GN Algorithm [7], Fast Newman Algorithms [8], etc. All these algorithms reveal community structures based on global information of network. However, while the network scale is too large, it is difficult to obtain the global information, especially in the condition of dynamic network, such as Internet. In many cases, the researchers concern is local community structure in network. For example, we only care about the community that someone belongs to, and do not care the community structure of the whole network. In this case, it is no use to consume time to detect the community structure of whole network, but only to search a local community that a special node belongs to.

Some researchers have already put forward a new algorithm of local community detecting. By searching neighbor nodes, the nodes that meet requirements are joined into the local community. Finally, the local community will be obtained. The attraction of local community to a node depends on the importance of the community for it. Generally, if a neighbor node is connected to most members of the local community, we can say this community has important influence on this neighbor node, this neighbor node tends to the member of the community. The clustering coefficient is a token of connection of neighbor nodes. So the relationship between community and its neighbor nodes can be determined by investigating the change of clustering coefficient. This algorithm can also be used for global community detecting, namely after finding node's community, start from any node in the network but out of this community, repeat this process, the global community structure of network can be obtained.

Related concepts

In an undirected and unweighted network G=<V,E>, which includes n nodes and m edges, e_{ij} expresses the edge between node i and node j, $V=\{i|i\in n\}$,$E=\{e_{ij}|i,j\in n \text{ and } i\neq j\}$.

A. Neighbor node set

Neighbor node set of node i: $N_i=\{j \mid \text{Node i is connected to Node j directly}\}$,Neighbor node set of community Ψ:

$$N_\Psi = \bigcup_{i=1}^{n} N_i$$

in which n expresses the number of nodes included in community Ψ.

B. Clustering coefficient

1) Clustering coefficient of node

Clustering coefficient of node i C(i) is defined as :$C(i)=2E(i)/T(i)$, $T(i)=k_i*(k_i-1)/2$,$C(i) \in [0,1]$. In which k_i is the number of neighbor nodes of node i. E(i) expresses the number of edges existed between k neighbor nodes of node i. While C(i)=1, all of the neighbor nodes of node i are connected to each other. Node i and its neighbor nodes form a global couple network, that is to say, the connection is closely.

2) Clustering coefficient of edge

$C(i,j)=|N(i,j)|/(k_i+k_j-|N(i,j)|-2)$,

in which $N(i,j)=N(i)\cap N(j)$ expresses the public neighbor set of node i and node j, $|N(i,j)|$ expresses the number of triangles which takes e_{ij} as edges. Clustering coefficient of edge $C(i,j) \in [0,1]$ expresses connection strength of two nodes connected by this edge. The greater the value, the stronger the strength, the two nodes are in the same community more possible.

3) Clustering coefficient of community

$$C(\Psi) = \frac{\sum_{i=1}^{n} C(i)}{n}$$

in which n expresses the number of nodes included in community Ψ.

For Clustering coefficient of community $C(\Psi)\in[0,1]$, while $C(\Psi)=1$,the clustering coefficient of all nodes in community are 1. Nodes in the community are connected to neighbor nodes out of

the community closely.

C. Community modularity

Community modularity index Q is a parameter used to depict the characteristics of community, defined as follows [9]:

$$Q = \frac{1}{2m}\sum (A_{ij} - \frac{k_i k_j}{2m})\delta(C_i, C_j)$$

in which K_i and K_j express the degrees of node, C_i expresses the community that node i belongs to, m expresses the total number of edges of network. While $C_i=C_j$, $\delta(C_i,C_j)=1$, otherwise $\delta(C_i,C_j)=0$. The value of Q is between 0~1. Generally, Q=0.3 is looked as the lower bound of network with obvious community structure. The closer the value of Q is to 1, the more obvious the community structure is.

Algorithm descriptions

Regard an unknown node as initial node to start this algorithm, and finally detect all communities. Of course, if there is not an unknown node, any node can be regarded as initial node. Firstly, an initial node is added to an empty community, and then a local community is obtained. By comparing the degrees and clustering coefficients of neighbor nodes, the neighbor node that meets agreed will be added to community, until there is not any node that meets agreed. At this time a complete local community is obtained. Select a node from the remained nodes randomly as initial node, and repeat the above process, until the whole network is divided into a plurality of community. While confirm the node in community, there is agreed as follows.

(1)If more than half connections of node i being out of this community are connected with this community, then this node i is in this community.

(2)If the clustering coefficient of community is 1, then all the neighbor nodes of this community should be in this community.

(3)If the clustering coefficient of a neighbor node of this community is 1, then this node and all of its neighbor nodes are in this community.

(4)Among all the edges that connected to community, if the clustering coefficient of edge e_{ij} is the largest non negative, Node i is one node in community, Node j is a neighbor node of community, and the clustering coefficient of other edges that connected to node j is less than the clustering coefficient of the edge e_{ij}, then node j should be in community.

The algorithm is described as follows:

Input: network G=<V,E>, V is a node set of network, E is an edge set of network, X is the target node, N_i is a neighbor node set of node i.

Output: community set of network.

1. Initial community Ψ is empty

2. If there is a target node, then select this target node as initial community. If there is not a target node, then select a node with the largest degree in remaining network as a initial node of community Ψ. Update the neighbor node set of community Ψ.

3. Calculate the degree of every node in neighbor node set $N\Psi_1$ of community Ψ_1. The node that meets the above agreed will be added to this community. While adding a node, its neighbor node set $N\Psi_1$ and its degree will be updated. Detecting the local community will stop unit there is not any neighbor node meeting this agreed.

4. Select a node from remaining nodes randomly as target node, and select neighbor nodes from remaining nodes that meeting the agreed to add to this community.

5. Repeat above process until all nodes have been added to the corresponding community. If several communities such as $\Psi_1, \Psi_2, \ldots, \Psi_n$, have been detected finally, calculate the community modularity Q in this condition.

6. Merge any two communities coming from detected n little communities, and calculate the modularity of new communities being merged. Compare this modularity with Q. If it is greater than Q, this merge is effective. Otherwise, the merge is invalid. Repeat this process until all merges are

invalid in one round, and then community detecting is finished.

Data preprocessing

DBLP is an integrated database system of research achievements in the field of computer. These achievements are English literatures of computer published in international journals and conferences. Papers included in DBLP are with high qualities, the updating speed of documents is very fast. This system is a good response of front direction of international academic research. In our research, 2555 papers and their citation relations from the year of 1990 to 2005 are crawled from DBLP database. Citation relation of literatures is shown in matrix in figure 2. This matrix is the adjacency matrix of the citation network.

There always exist some irrelevant nodes which are no use in detecting community. On the other hand they will occupy large storage space and cause interference on the results. After cleaning these data, it is found that there are 582 discrete nodes in the network. Getting rid of these nodes, 1973 nodes are remained. With DFS traversal, 418 sub graphs based on these remained nodes are found. The number of sub graphs appears too much. So according to the scale of network, set a threshold of 10. Finally there are 27 sub graphs meeting this threshold to be remained. All of the nodes that meet this threshold compose a new network. Mapping table is shown in table 1.

$$A_{2555\times2555}=\begin{pmatrix}0 & 1 & \cdots & 0\\ \vdots & & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{pmatrix}$$

Figure 2 Literature citation relations

Table 1 Mapping table

Original node	Mapping node
4	0
6	1
……	……
2552	841

Application of algorithm

For example, set a target literature “Workflow Verification: Finding Control-Flow Errors Using Petri-Net-Based Techniques”, it is looked as a node and its number is 72. In mapping table, the number 72 is mapped as number 13. Through the local community algorithm, the community {0,1,2,3,4,13,14} which includes node 13 is detected. According to the map table, this result maps back to original number is {4,6,81,5,30,72,80}. So the literature numbered 72 is in the community Ψ={4,5,6,30,72,80,81}. That is to say the literatures of which number in this set are the corresponding field literatures. Continuing to detect community in the whole network system, another 46 scientific research groups are detected. Every scientific research group has its own research field and the authors in these groups can carry out exchange and cooperation in scientific research. If there is not a target node, we can select any node as a initial node to start detecting community in this network.

After detecting community, many node sets can be obtained. Every node set is a scientific research group with the similar research direction. According to the result, it is very easy to answer the problems below, such as who are in these scientific research groups, who is the leader of the group, what do they research. By scanning the global network, 51 core literatures in this citation system are found which is shown in table 2. For example, select the lecture numbered 280 as the target node, by local community detecting algorithm, the community including the lecture numbered 280 is detected. The scale of this community is 37. Parts of the nodes in the community are shown in table 3. All the authors of literatures in the community form a scientific research group. The authors of these 37 literatures appear 96 times. After merging repeated authors, 64 people are remained. That is to say, there are 64 researchers in this scientific research group. Sort these

researchers according to the times they appear in the community. The result of sort is shown in Table 4. From Table 4 we can see that these 5 researchers, Jiawei Han, Philip S. Yu, Rakesh Agrawal, Tomasz Imielinski and Arun N. Swami are the leaders of this research field. By reading the title of the literature, It is known that the literature numbered 280 is an article about association rules mining. After reading the title of literature in Table 3, the research direction “association rules mining” of this group is further confirmed.

By local community detecting algorithm, the community with core literature can be detected, and then the research direction of research group can be known. This method can save more time than the method of comparing key words and mining abstract. Furthermore, the result of the former is more exact.

Table 2 Core literature

Number	Literature
280	Fast Algorithms for Mining Association Rules in Large Databases
414	Description Logics in Data Management
600	Using Segmented Right-Deep Trees for the Execution of Pipelined Hash Joins
2450	On the Propagation of Errors in the Size of Join Results

Table 3 Literature and its author

Literature Number	Title of literature	Author of literature
280	Fast Algorithms for Mining Association Rules in Large Databases	Rakesh Agrawal, Ramakrishnan Srikant
558	Description Logics in Data Management	Gregory Piatetsky-Shapiro
564	Using Segmented Right-Deep Trees for the Execution of Pipelined Hash Joins	Jong Soo Park, Ming-Syan Chen, Philip S. Yu

Table 4 Citation frequency

Author	Appear times
Jiawei Han	84
Philip S. Yu	5
Rakesh Agrawal	4
Jennifer Widom	3

Conclusion

Clustering coefficients that indicate the extent of connections between neighbor nodes are the main parameters in local community detecting algorithm. Start from a single target node, other node that meets agreed are added to the community one by one, and then the scale of community expands continuously. As a result, a community consisted of nodes with the same or similar character is obtained. Data used in this paper come from DBLP citation system. After analyzing and cleaning, these data are used in local community detecting algorithm. In the experiment, many scientific research groups are found. The core literature can be obtained according to the citation frequency of the literatures in the group, at the same, the research direction and the leader can be obtained too. If there is no target node, by local community detecting algorithm, the potential patterns, the frontier issues, future research focus can be found easily. If data set is changed, the local community detecting algorithm can be used in other field, such as the social network. It has broad prospect for application.

References

[1] Garfield E. Citation Indexes for Science: a new dimension in documentation through association of ideas[J]. Science, 1955, 122: 108-111.

[2] Watts D J, Strogatz S H. Collective dynamics of 'small world' networks [J]. Nature, 1998, 393(6684): 440-442.

[3] Barabási A L,Albert R Emergence of scaling in random networks[J].Science,1999,286:509-512

[4] WANG Xiaofan, LI Xiang, CHEN Guanrong. Theory of complex networks and its application[M]. Beijing: Tsinghua University Press, 2006.

[5] Fildler M. Algebraic Connectivity of Graphs[J]. Czech Math J.1973,23(98):298-305.

[6] Phothen A,Simon H.Partitioning Sparse Matrices with Eigenvectors of Graphs[J].SIAM J.Matrix Anal.Appl.,1990,11(3):430-452.

[7] Newman M E J;Girvan M Finding and evaluating community structure in networks [J]. Physical review E, 2004, 69(2): 26113.

[8] Newman M E J Fast algorithm for detecting community structure in networks[J].Physical Review E,2004,69(6):066133.

[9] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Phys Rev E, 2004, 69(2): 026113.