复杂网络中重叠社团发现的算法探讨

□ 屈丽娟 屈宝强 / 中国科学技术信息研究所 北京 100038
王珏 / 西安电子科技大学 西安 710126

摘要：复杂网络聚类算法的研究对分析网络拓扑结构、理解其功能、发现网络中的隐藏规律以及预测网络行为具有十分重要的理论意义。目前许多寻找重叠点的算法不多，并且很多都需要比较高的时间复杂度。文章通过观察网络社团之间的相邻点与每一社团的连接边数以及定义阈值的方法对其进行了改进，最后通过期刊之间的引用关系计算期刊引用网络的相似性，构造网络图，采用基于谱的聚类算法和改进后的方法对该图进行聚类，从而验证改进算法的先进性。

关键词：复杂网络，聚类算法，相似性，重叠点
DOI：10.3772/j.issn.1673—2286.2011.05.006

现实世界中许多系统都可以用复杂网络来描述，如科研合作网、万维网、论文引用网、因特网、电力网、神经网络、新陈代谢网络以及食物链网络等。近年来，复杂网络已经成为当今科学界研究的前沿和热点，研究者来自物理学、生物学、计算机科学、管理学、社会学以及经济学等各个不同领域。从美国科学信息研究所（ISI）的统计结果可知[1]，2000-2009十年间发表的，以“复杂网络”（complex networks）为主题（topic）的SCI论文达到了4925篇之多，引用高达76834次，而且依旧呈现强劲的上涨势头。得益于陈关荣、汪小帆等工程科学学者以及方锦清、何大韧、狄增如等物理科学学者早期的推动，我国学者得以较早意识到复杂网络研究的重要性，也呈现出了迅猛发展的势头。在统计的SCI论文中，地址中标明中国的论文共有1318篇，超过了四分之一[1]。同样对复杂网络聚类方法[2]的研究也不断增多，复杂网络聚类算法旨在揭示出复杂网络中真实存在的网络簇结构，对其的研究对分析复杂网络的拓扑结构、理解复杂网络的功能、发现复杂网络中的隐藏规律以及预测复杂网络的行为，不仅具有十分重要的理论意义，而且具有广泛的应用前景。

1 现有聚类算法及其局限性

目前已存在多种复杂网络聚类算法其中杨博[2]等人在相关研究中提出，按照所采用的基本求解策略，将其归为两大类：基于优化的方法（optimization based method）和启发式方法（heuristic method）。前者将复杂网络聚类问题转化为优化问题，通过最优化预定义的目标函数来计算复杂网络的簇结构。如谱方法（spectral method）将网络聚类问题转化为二次型优化问题，通过计算特殊矩阵的特征向量来优化预定义的“截（cut）”函数。后者将复杂网络聚类问题转化为预定义启发式规则的设计问题。如被广泛引用的Girvan-Newman算法[3]的启发式规则是：簇间连接的边介数（edge betweenness）应大于簇内连接的边介数。

虽然上述算法都有自己的优点，但都仅局限在某些特定的网络结构中，不适用于所有网络结构。在现实世界复杂网络社团之间通常会存在一些重叠点，这些点属于社团间的交叉部分。现有的许多算法都只能得到标准化划分，即节点有且只属于一个社团的情况。

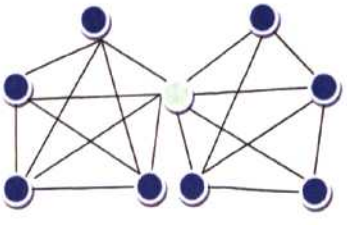


图1 社团结构图之一

如图1所示，白色节点既可以属于左边社团，也可以属于右边社团，现有的算法会将白色节点随机地归为其中一社团，导致另一社团丢失这个节点，聚类出现不准确。

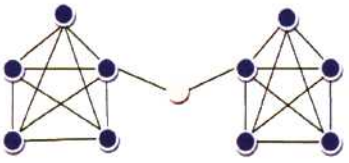


图2 社团结构图之二

如图2所示，现有一些算法将中间的白色节点划分到其中一个社团中，一些算法直接将其舍掉，但是在不同的背景下白色节点可能属于不同的社团，所以认为上述两种情况都不太合实际。

进一步的改进中，有些学者也提出了一些找到重叠点的方法，如中国科学院计算技术研究所陈学旗和中国科学技术大学胡茂斌在GN算法的基础上提出了挖掘具有重复节点的群落结构的高效算法[4]，通过基于聚类密度的方法来寻找这些重叠点，即在简单聚类之后，对每一个节点进行遍历，依次来计算这个节点在归为每一类当中时是否会使得该类密度变大，如果是，则认为是重叠点，反之就不是重叠点。这种方法是现在研究中比较常用的一种，但是它时间复杂度较大，运行速度比较慢。

另一种寻找重叠点的方法[1]是在聚类之后，将类与类之间比较相近的每一个点分裂成几个，然后对应到附近每一个类当中，类似于上述计算类的密度来判断，同样该方法时间复杂度也比较高，而且在不同的背景下很可能有误，这些都是算法现在出现的问题。

2 聚类算法的改进

2.1 改进算法的基本思想

基于上述问题，需要找到一个时间复杂度较小而且较准确的方法来寻找重叠点，而不是单纯地将其舍弃或归入一个社团当中。

重新定义如下：在处理大量数据，完成简单聚类过程后，分别对聚类图中类与类之间相邻的点进行分析其与每个社团之间节点的连边条数，根据连边数来确定方法。假设节点与社团（1，2，3… … n）连接的边数分别为n_1, n_2, n_3 ……n_n，比较这些数，选取其中最大的一个n_m为基准，分别计算n_m:n_1, n_m:n_2, n_m:n_3, …… n_m:n_n，如果n_m:n_n≥a（a ≥1，当a越小，该点是重叠点的可能性越大），则认为该节点是只属于m社团，它不属于m社团和x社团之间的重叠点；如果n_m:n_n≤a则认为该节点是这两个社团之间的重叠点。（n_x代表节点与社团X之间的连边条数，下面我们假设a=1.2）

如上述图1，白色节点与两边社团之间的连接边数相同，比例为1<1.2，可认为白色节点为一个重叠点。图2中，白色节点通常被舍弃，但是这样并不符合实际，根据上述的方法也认为这是一个重叠点。

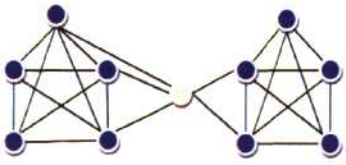


图3 社团结构图之三

如图3所示，中间节点与左边的社团的连接边数为3，与右边连接的边数为2，3:2=1.5>1.2，则认为白色节点只属于左边社团。

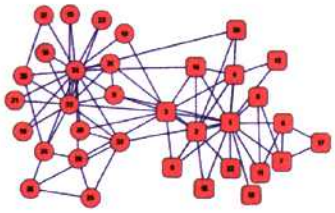


图4 Zachary研究的空手道俱乐部内部成员的关系网络[5]

如图4所示，前面提到的不同的划分算法把节点3归为不同的类，即节点3经常被分错类，从网络拓扑角度来看，可以将节点3看做是一个重叠点，如其分别与两个社团的连接边数为7条，7:7=1<1.2，则认为这个点是一个重叠点。

在现实社会网络中，节点数量和连接边数都比较多，尽管有些节点与两边的连边数不是相同的，但如果相差的边数相比所有边数来说可以忽略不计时也可以认为这个点是重叠点。

2.2 算法的流程

运行基于划分的聚类算法（上面提到的任意一种都可以，文章使用的是基于谱的聚类算法）。得到聚类社团之后继续按照下列步骤进行。

(1) 对i∈{1,2,3,…k}

(2) {对j∈{1,2,3,…n}

(3) if(n_{ij}>0)

(4) {求m，使得n_{im}最大，m=1,2,3,…,n

(5) if(n_{im}, n_{ij}≤a)

(6) 认为节点i是社团之间的重叠点；

(7) Else节点i仅为m社团；

(8) Else i++}

}

算法的第一行和第二行虽然对所有n个社团和每个社团的k个顶点都进行扫描，但第4-7行的操作是以第3行的条件来判断的，如果条件不满足就不会有后面的4步，所以真正进行操作的遍数是候选节点的个数q，第四行有 o (n)步，第5、6步都为 o (1)，一遍的时间复杂度为 o (n)，所有候选节点都遍历完的总时间复杂度是 o (n*q)。

3 实例验证

3.1 实验数据

本文选用中国知网引文数据库中引用次数在前36的管理类期刊和前12的情报类期刊，即管理工程学报、管理科学学报、预测、系统工程、运筹学报、中国机械工业、科学学研究、工业工程与管理、计算机集成制造系统、中国人口资源与环境、工业工程、经济与管理研究、科学进步与对策、经济管理、生态经济、资源科学、管理科学、软科学、情报学报、中国图书馆学报、图书情报工作、情报科学、情报理论与实践、现代图书情报技术、图书情报知识、情报资料工作、情报杂志、现代情报、图书与情报、中国信息导报、南开管理评论、中国管理科学、管理评论、研究与发展管理、科研管理、系统工程理论与实践、管理世界、战略与管理、中国软科学、经济体制改革、外国经济与管理、科技管理研究、科学学与科学技术管理、中国人力资源开发、宏观经济管理、中国科技论坛、商业经济与管理、管理现代化。

选择年限在2007年到2010年，利用中国期刊网CNKI，统计期刊之间的引用次数，并且摒弃引用次数小于3（阈值为3），且不在统计之内的期刊。由于其中有跨学科的期刊，还有近四年中没有被引用到的期刊，数据有重复，或引用次数为0，所以最终选用有效期刊数46个。

根据采集到的数据，构建期刊间的引用关系网络。利用聚类算法对上述期刊聚类，将引用有向图转变为无向图，本文通过数据矩阵计算期刊之间的相似性来构造引用网络结构图。

3.2 网络构建

3.2.1 相似性定义

相似性的计算方法有很多，本文将定义建立在期刊间的相互引用关系上。一个期刊甲引用了丙和丁两种期刊，同时乙也引用了这两种期刊，则甲和乙就非常相似；如果乙只引用了丙和丁中的一种，则他们之间的相似性就没有前一种情况高。如图5所示，期刊A1和A引用了B期刊，期刊A2和A引用了C期刊，不考虑引用的次数，就可以说期刊A分别与A1、A2同等相似。当考虑一步间接引用时，A直接引用了C，而且A引用了B后B引用了C，这就说明C对于A来说更为重要，则A与A2的相似性就比A与A1的相似性要大一点，也就是说A2比A1与A更为相似。当然这只是考虑了一步间接引用，以此类推，当考虑的步数越多，所得到的结果就越精确。

首先以矩阵的形式表示前面得到的引用数据，矩阵的行表示该期刊引用情况，列表示该期刊被引用情况，于是得到如下矩阵：

$$A_{ij}=\begin{bmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{146} \\ a_{21} & a_{22} & \cdots & \cdots & a_{246} \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ a_{461} & a_{462} & \cdots & \cdots & a_{4646} \end{bmatrix}$$

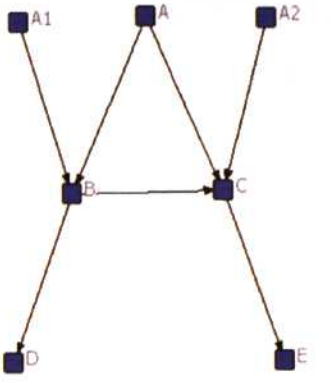
其中，aij表示期刊i引用j的次数，i，j=1，2，...，46，将A_{ij}转置为A^T_{ij}，

$$A^T_{ij}=\begin{bmatrix} a_{11} & a_{21} & & a_{461} \\ a_{12} & a_{22} & & a_{462} \\ & & & \\ & & & \\ a_{146} & a_{246} & & a_{4646} \end{bmatrix}$$

然后进行“归一”，即以列向量为单位，算出其和，再用列向量中的每一项除去它们的和，得到一个新的列向量，接着按此法算出下面的所有列向量。接着求出转置矩阵，然后算出它们之间的内积，运用一下公式：

$$\frac{(x,y)}{\sqrt{(x,x)}\bullet\sqrt{(y,y)}} \qquad (3\text{-}1)$$

其中，x、y均为上述我们得到的转置矩阵的列向量。



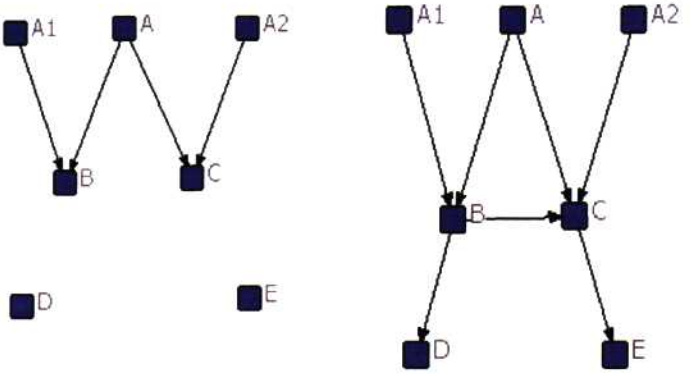


图5 ABCDE的引用关系

3.2.2 相似性结果

（1）把数据整理成矩阵A，在MATLAB（以下整个实验是在MATLAB下进行编程的）中，用Journal_citation=[数据]，对已知矩阵转置A^T，求出每列的和，形成对角矩阵，对A^T求逆，得到对角线上依次是每列数值和的倒数的矩阵，乘以矩阵A，得出状态转移矩阵P。

（2）M=P+P*P，其中P*P是一步间接引用带来的影响，加上直接引用影响矩阵P后，M就表示这两种影响的叠加。

（3）用公式（3-1）计算出相似性S。其中，x和y都是矩阵M中的列向量。

表1 随机三个期刊的相似性

	管理科学学报	情报学报	系统工程理论与实践
管理工程学报	0.717433	0.088071	0.805245
管理科学学报	1	0.061353	0.562793
预测	0.620565	0.122344	0.696519
系统工程	0.554459	0.089776	0.917138
运筹学学报	0.339951	0.054276	0.795926
中国机械工业	0.024503	0.008232	0.06623
科学学研究	0.258563	0.216111	0.250253
工业工程与管理	0.528863	0.090038	0.845579
计算机集成制造系统	0.147102	0.046662	0.287398
中国人口资源与环境	0.084715	0.025276	0.159448
工业工程	0.512908	0.088123	0.799724
经济与管理研究	0.259028	0.069589	0.289559
科技进步与对策	0.280682	0.182671	0.336822
经济管理	0.311112	0.077089	0.298651
生态经济	0.018426	0.008028	0.02338
资源科学	0.034668	0.007851	0.061424
管理科学	0.658288	0.091529	0.718345
软科学	0.507086	0.147953	0.539455
情报学报	0.061353	1	0.07899

中国图书馆学报	0.014631	0.752528	0.016553
图书情报工作	0.042066	0.819333	0.055332
情报科学	0.067201	0.864775	0.081395
情报理论与实践	0.048019	0.885956	0.052289
现代图书情报技术	0.02388	0.812508	0.038273
图书情报知识	0.037225	0.818419	0.040531
情报资料工作	0.017288	0.750734	0.021621
情报杂志	0.128354	0.794122	0.162006
现代情报	0.031048	0.740485	0.043887
图书与情报	0.024486	0.74827	0.031573
中国信息导报	0.029087	0.895801	0.03574
南开管理评论	0.284261	0.050623	0.220885
中国管理科学	0.52924	0.074282	0.73171
管理评论	0.46222	0.092725	0.51294
研究与发展管理	0.230214	0.15771	0.27323
科研管理	0.356666	0.180978	0.354368
系统工程理论与实践	0.562793	0.07899	1
管理世界	0.21179	0.039907	0.14169
中国软科学	0.330529	0.142607	0.382479
经济体制改革	0.223677	0.04105	0.155228
外国经济与管理	0.183956	0.01538	0.068195
科技管理研究	0.28762	0.218364	0.35608
科学学与科学技术管理	0.324383	0.186131	0.321579
中国人力资源开发	0.112819	0.058005	0.08596
中国科技论坛	0.199124	0.160438	0.221048
商业经济与管理	0.36464	0.064696	0.341735
管理现代化	0.385448	0.133544	0.307864

通过计算相似性，可以了解哪些期刊之间存在着内容上的相关或相近，在以后进一步研究时，可以方便我们全面准确地找到相关课题的所有资料，进行更为细致深入的研究。同时，还为以谱算法为基础的聚类方法提供数据基础。

3.2.3 构建网络图

根据相似矩阵，使用软件Netdraw画出图形，以相似系数>=0.4为界限，构造期刊之间的引用关系网络图，图中期刊越相似，两者之间的连线就越粗。

3.3 实验结果及其分析

3.3.1 现有聚类算法的实现

选用基于谱的聚类算法实现：

第一步：根据上面得到的相似矩阵A，定义一个阈值，找出对应的邻接矩阵S；

第二步：计算邻接矩阵的特征值和特征向量；（[V，E]=eig(S)%V表示矩阵S的N个特征向量，E表示矩阵S的N个特征值）

第三步：定义一个K值，选取前k个特征值对应的特征向量构成新的向量空间F，利用K均值聚类算法进行聚类，结果为图7。

3.3.2 本文改进后聚类算法的实现

基于上述提出的改进方法，运行后的聚类效果如图8所示。

图6 期刊引文网络图

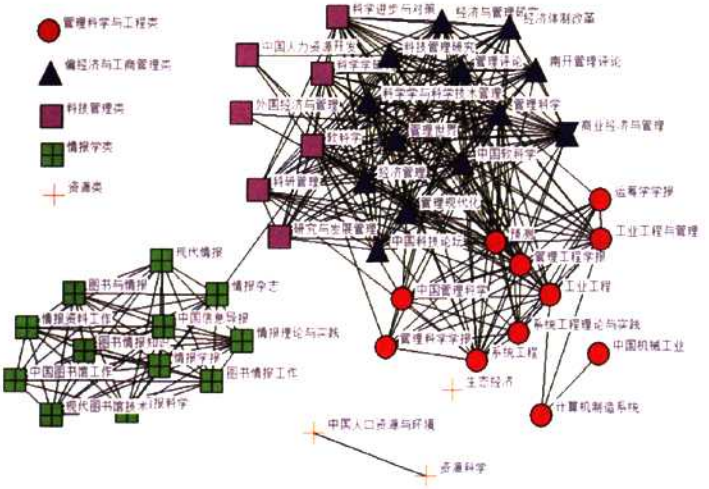


图7 聚类效果图

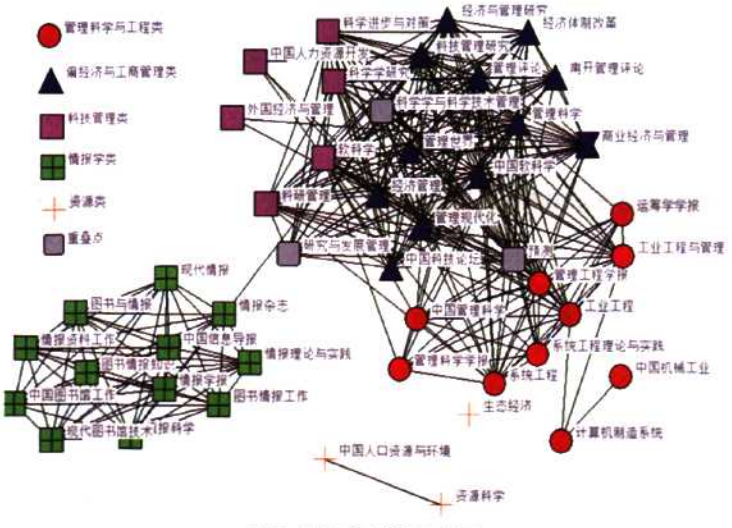


图8 改进后聚类效果图

3.3.3 分析结果

如图7所示，一共46个期刊主要分成了五大类，情报类的12种期刊聚类稳定，内部关系紧密，相似程度极高，聚类效果很好。《中国人口资源与环境》、《资源科学》、《生态经济》这三种资源类期刊的文章，聚类比较独立，效果也比较好。

其次，管理类期刊共被分成三类，圆圈类是管理科学与工程类，三角形为经济和工商管理类，正方形为科技管理类。

最后，从图中发现，有些期刊虽然被分到了同一个类，但是它与其他类的关系也非常紧密，引用也比较频繁，并不应该仅属于一个类当中，比如，《预测》这个期刊分别于三角形和圆圈这两类的连接边数都是8，却被强制分到了红色这一类当中。

图8所示，《预测》、《科学学与科学技术管理》、《研究与发展管理》被标出为重叠点，分析可以知道《预测》这个期刊既可以属于管理科学与工程类也可属于企业管理类，为一重叠点。同样，期刊《科学学与科学技术管理》、《研究与发展管理》是企业管理类与科技管理类之间的重叠点也是较合理的。可以看出，改进之后的算法聚类的效果更接近于实际的分类情况。

从聚类的整体效果来看，比现有算法有了改进，但是还是有一些不尽如人意的地方，例如《商业经济与管理》并没有归到这里面任何一类；《中国科技论坛》和《科技管理研究》应该更偏向于科技管理这一类，而图中将其分到了企业管理这一类当中。

造成上面这些现象的原因有以下几种：

第一，这三类期刊都是管理类的，互相之间都有很大的联系，宽泛地来讲都可以认为属于一类，严格划分并不能很肯定地说某一期刊仅属于某一类，文中仅根据期刊间的引用次数来判断相似性有一定的局限性。比如，某个期刊，我们假定它是属于情报学的期刊，但是它引用的大部分是其他学科的期刊，所以按定义相似性的方法，就会把它分到另外那个学科去。

第二，文中只选用了2007年到2010年的数据构建网络，3年半的统计数据量太少，不能完全说明问题。

第三，本文选取的聚类方法是基于谱的聚类算法，是根据邻接矩阵来判断期刊之间是否有关系，而邻接矩阵取决于自定义的阈值，只要是大于等于这个阈值都被视为是有同样的关系，矩阵表示都为1，例如定义阈值为0.4，期刊之间的相似性0.5、0.6、0.7、0.8、0.9表示的意义是不一样的，值越大表示越相似，此处被视为是相同，所以会出现大部分相似的期刊都被聚为一类，并没有很清楚地分开，所以网络本身不能完全准确地说明期刊之间的聚类关系。

4 总结和展望

寻找重叠点仍是当今复杂网络聚类算法研究中最主要的突破口，虽然现在出现了许多含有寻找重叠点的算法，但是无论是从聚类精度还是时间复杂度上都达不到很好的权衡，大部分都是注重了计算速度却是以牺牲精度为条件，或者是为了达到很好的聚类精度，导致时间复杂度特别大，这是一个需要大量学者不断努力研究的问题。

本文只是提出比较简单的重叠点寻找方法，还有改进的余地，例如单单以考虑边的条数来判断是否为重叠点有点片面，没有考虑边的权重，造成网络图构造方面出现偏差，这些也是今后研究急需解决的问题。

参考文献

[1] 吕琳媛,陆君安,周涛. 复杂网络观察[EB/OL]. [2010-11-01]. http://www.sciencenet.cn/m/user_content.aspx?id=359994.
[2] 杨博,刘大有,LIU Jiming,等. 复杂网络聚类算法[J]. 软件学报,2009,20(1):54-66.
[3] GIRVAN M, NEWMAN M E J. Community Structure in Social and Biological Networks[J]. Proc. of the National Academy of Science, 2002,9(12):7821-7826.
[4] SHEN H-W, CHENG X-Q, CAI K, et al. Detect Overlapping and Hierarchical Community Structure in Networks[J]. Physica A 388, 2009:1706-1712.
[5] 汪小帆,李翔,陈关荣. 复杂网络理论及其应用[M]. 北京:清华大学出版社,2006:21-28.

作者简介

屈丽娟（1988-），情报学专业，中国科学技术信息技术研究所硕士研究生。E-mail: qulijuan1117@126.com

Discussion of the Algorithm Finding Overlapping Communities in Complex Networks

Qu Lijuan, Qu Baoqiang / Institute of Scientific and Technical Information of China, Beijing, 100038
Wang Yu / Xi'an University of Electronic Science and Technology, Xi'an, 710126

Abstract: The study of the network clustering algorithms which aims to discover all natural network communities from given complex networks is fundamentally important for both theoretical researches and practical applications, and can be used to analyze the topological structures, understand the functions, recognize the hidden patterns, and predict the behaviors of complex networks including social networks, biological networks, World Wide Webs and so on. This paper reviews some algorithms, which finds out the overlapping points between the networks have very high time complexity. Addressing this lack, we propose an improvement method which can find some overlapping point by observing the number of edges between the neighboring community 's point and every community. Finally, in order to prove the improved algorithm, this paper selects an application of journal references and draws a network map based on the similarity calculated by the references among the journals. Then, clustering the graph with the spectral clustering algorithm and the improved method, the paper analyzes and compares the result.

Keywords: Complex network, Clustering algorithm, Similarity, Overlapping point

（收稿日期：2010-11-08）