

基于节点相似度的社团划分方法研究*

甘立强^{1,2} 王旭阳² 燕楠¹ 王岚¹

(1. 中国科学院深圳先进技术研究院中国科学院人机智能协同系统重点实验室 深圳 518055)

(2. 兰州理工大学计算机与通信学院 兰州 730050)

摘要 社团划分有助于了解网络结构、分析网络特性。大多数传统的社团划分方法只关注于网络的局部信息,通过优化特定的目标函数进行社团的划分,但是这种方法往往只针对特定的网络有效,不具有普遍性。因此,论文提出了一种综合考虑网络局部信息和全局信息的算法,该算法不需要任何的关于社团的先验知识,同时在选择不同初始节点的情况下,社团及其划分结果都具有鲁棒性。通过用一些真实网络评估算法的性能,结果表明,论文算法能高效地划分复杂网络中的社团结构。

关键词 复杂网络;节点相似度;社团划分

中图分类号 TP391 **DOI:**10.3969/j.issn.1672-9722.2018.02.001

Detecting Community Structure in Complex Networks Based on Node Similarity

GAN Liqiang^{1,2} WANG Xuyang² YAN Nan¹ WANG Lan¹

(1. CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Shenzhen 518055)(2. School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050)

Abstract The detection of the community structure in networks is beneficial to understand the network structure and to analyze the network properties. Most of conventional community detection methods focused on only considering the local information of the network to optimizing a certain objective function. However, these methods just have preference for specific types of networks but are not general. In present study, a new node similarity which captures global and local structures is proposed to detecting network structure. The presented method does not need any prior knowledge about the communities and its detection results are robust on the selection of the initial node. Some real-world and networks are used to evaluate the performance of the presented method. The simulation results demonstrate that this method is efficient to detect community structure in complex networks.

Key Words complex networks, community detection, node similarity

Class Number TP391

1 引言

近年来复杂网络已经得到了广泛的应用,如万维网、社会网、蛋白质交互网以及科学家合作关系网等。这些网络具有较高的复杂性,因此被称为“复杂网络”^[1]。复杂网络已经成为当前最重要的多学科交叉研究领域之一^[2]。随着对复杂网络性

质的物理意义和数学特性的深入研究,人们发现许多复杂网络都具有一个共同的性质,即社团结构^[3]。社团内部各节点之间连接紧密,而不同社团节点之间连接稀疏。社团结构分析在生物学、物理学、计算机图形学和社会学等领域内有着广泛的应用。

目前,学者们已经提出了很多有效的社团检测

* 收稿日期:2017年8月12日,修回日期:2017年9月20日

基金项目:国家自然科学基金项目(编号:11474300,61771461);深圳市基础研究项目(编号:JCYJ20160429184226930)资助。

作者简介:甘立强,男,硕士研究生,研究方向:复杂网络。王旭阳,女,硕士,教授,研究方向:数据库理论及应用,数据挖掘。燕楠,男,博士,副研究员,研究方向:正常与病理嗓音。王岚,女,博士,研究员,研究方向:语音可视化和大规模连续语音识别。

方法,比如基于分裂思想的GN算法^[4],自包含GN算法^[5]、基于相异性的分裂算法^[6]。该类算法划分效果较好,但是时间复杂度太高,不适合在大型网络中使用,随后学者们根据上述缺点提出了基于凝聚思想的Newman快速算法^[7]、CNM算法^[8]、MSG-VM^[9],但是这些算法容易将过多的节点划分到同一社团,导致社团过大,与真实网络的划分情况不相符。因此,针对这些缺点学者们又相继提出了LPAm算法^[10]、SYM^[11]、CHM^[12]、ILM^[13]等,这些算法同样存在很多问题,诸如LPAm算法的稳定性不高,时间复杂度虽然较小,但是不容易得到最优结果,CHM算法空间复杂度高,需要较大的内存空间,不适合大型网络,ILM算法参数选择困难,算法稳定性差等。

针对上述存在的问题,通过比较研究各种网络节点相似度的构造方法,通过综合考虑网络的局部信息和全局信息,设计了新的构造节点相似度的方法,进而提出了基于节点相似度的社团划分算法。该算法在不需要知道任何关于网络的先验知识的情况下,仍然取得了较好的划分结果。

2 基本概念

2.1 节点相似度

构造节点相似度方法目前主要是基于网络的局部信息^[14],比如Salto^[15]、Jaccard^[16]、Hub Pr-omot-ed^[17]、Leicht-Holme-Newman^[18]、RA^[19]等。其中RA方法是目前效果最好的节点相似度的构造方法^[19]。

RA方法是通过模拟网络中资源的分配过程来衡量两个节点的相似度。假设有一个节点对 (i, j) ,节点 j 通过它们的共同邻居节点作为中介间接地向节点 i 发送一定资源。同时规定每个中介能拥有的资源是一定的。因此每个中介发送给节点 i 的资源取决于它的度。资源分配过程如图1,我们可以把节点 i 收到节点 j 发送的资源的多少定义为节点 i 和节点 j 的相似度:

$$S_{ij} = \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{k(z)} \quad (1)$$

其中, $\Gamma(i)$ 代表节点 i 的所有邻居节点的集合, $z \in \Gamma(i) \cap \Gamma(j)$ 代表节点 i 和节点 j 共同邻居节点的集合, $k(z)$ 代表节点 z 的度。

但是这种只是依靠网络的局部信息作为衡量相似度的标准容易过度追求局部最优,忽视了网络整体的结构,造成与实际网络不符的情况,例如将RA算法应用在经典的Zachary空手道俱乐部网络

中,按照式(1)计算节点之间的相似度,发现节点3对应的最大相似度节点是节点34,应该被分到同一个社团中,但是现实网络中它们被分到了以教练和主管为首的两个社团中。所以划分社团中应该考虑网络的全局拓扑信息。如果两个节点的距离越近,即它们之间的最短路径越短,这两个节点同属于一个社团的机会就越大,反之,则它们属于一个社团的机会就越小^[20]。因此综合两个节点间共同邻居的网络局部信息和节点间最短路径的网络全局信息,本文重新定义了节点相似度 $Sim(i, j)$:

$$Sim(i, j) = \begin{cases} \frac{n}{d(i, j)} \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{k(z)}, & \text{节点 } i, j \text{ 可达} \\ 0, & \text{节点 } i, j \text{ 不可达} \end{cases} \quad (2)$$

其中 $d(i, j)$ 表示节点 i 和节点 j 之间最短路径长度, n 代表网络中的节点数目,节点 i, j 可达是指节点 i, j 之间存在路径将其连通。当 $d(i, j)$ 越小时, $1/d(i, j)$ 越大,则 $Sim(i, j)$ 就越大,两个节点越相似。同时如果两个节点之间没有可以到达的路径,这说明这两个节点之间的资源交换主要依靠邻居节点,相对于直接相连的节点,交换的资源较少,可以忽略。因此我们把相似度设置为0。因为 S_{ij} 本身很小,再除以 $d(i, j)$ 结果可能小到超出计算机的表示范围,就会出现浮点数下溢出的情况,将该运算结果处理成机器零,这时就无法对相似度进行比较。所以乘以 n 对结果进行适当的放大,防止这种情况发生。

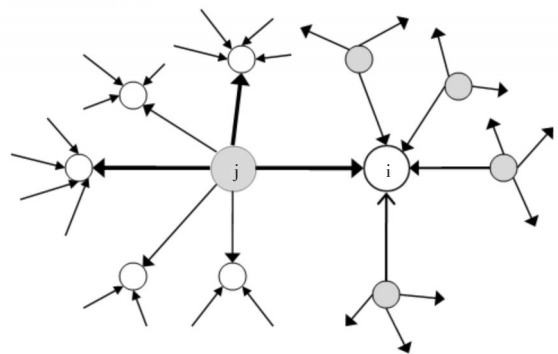


图1 网络资源流动过程图

2.2 模块度和准确率

社团模块度^[20]是Newman等引进的一个衡量网络划分质量的度量标准。假设有某种划分形式,将网络划分为 k 个社团,定义一个 $k \times k$ 维的对称矩阵 $E=(e_{ij})$,其中 e_{ij} 表示网络中连接两个不同社团的节点的边在所有边中所占的比例,这两个节点分别位于第 i 和第 j 个社团。模块度 Q 可表示为

$$Q = \sum_i (e_{ii} - a^2) = Tre - \|e^2\|$$

其中： $Tre = \sum_i e_{ii}$ 为矩阵 E 对角线上各元素之和，它表示网络中连接某一个社团内部各节点的边在所有边中占的比例； $a_i = \sum_j e_{ij}$ 为每行各元素之和，它表示与第 i 个社团中的节点相连的边在所有边中所占的比例。一般情况下，好的社团划分内部节点连接概率 Tre 应远大于社团外部节点的连接概率 $\|e^2\|$ ，即模块度越大，社团划分的质量越好。现实中的网络 Q 值一般介于 0.3~0.7 之间， Q 值最大为 1。准确率是指算法划分正确的节点占网络中所有节点的比例。以上两个指标将作为本文算法与其他算法比较的评价指标。

3 基于节点相似度的社团划分方法

3.1 基本思想

两个节点的相似度越大，属于同一个社团的概率越大^[21]，根据这一思想，本算法利用节点之间的相似性，判断两个节点是否属于同一个社团。

对于一个含有 n 个节点的网络，首先根据式(2)计算网络中所有节点对的相似度。由此我们可以得到一个 $n \times n$ 的矩阵 $Sim(n, n) = [Sim(i, j)]$ 。 $Sim(i, j)$ 代表的是两个节点之间的相似度。为了方便处理数据，我们把这个矩阵的每一行用链表存起来，头节点保存节点 i ，头节点的 next 节点保存节点 i 对应的相似度最大的节点 j 。如果有多个相似度相同的节点，则把它们都保存起来。算法具体步骤：

- 1) 首先将每个节点看作是一个社团，选择任意一个节点作为初始节点；
- 2) 然后从剩余的节点中选择和这个节点相似度最大的节点，将它们合并成一个新的社团；
- 3) 最后再将新加入的节点作为初始节点，寻找和它相似度最大的节点。如果找到的这个节点不在当前的社团中，则把这个节点合并到当前社团，如果在的话，就随机从剩余未处理的节点中选择一个作为初始节点，回到 2)；
- 4) 重复执行步骤 2) 和 3)，直到处理完所有的节点。

为了更清楚地表达本文算法的思想，我们构建了一个含有 8 个节点的小网络(如图 2)。

首先计算每个节点和它邻居节点的节点相似度，然后找出每个节点以及它对应的相似度最大的节点(结果如表 1)。

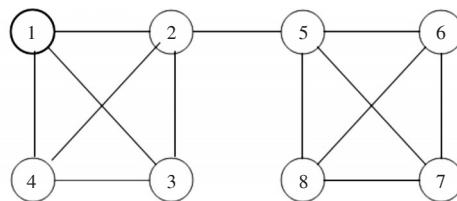


图 2 示例网络

表 1 对应的最大相似度节点列表

Node	MSNs	Node	MSNs	Node	MSNs
1	2	4	2	7	5
2	1,3,4	5	6,7,8	8	5
3	2	6	5		

注：MSNs 最大相似度节点

然后随机地选择一个节点，用带箭头的线段将当前节点和对应的相似度最大的节点连接起来，箭头指向的是对应的相似度最大的节点。同时，线段两端的节点合并到同一个网络。等到标记完网络中所有的节点，划分过程结束，得到最后的划分结果(如图 3)。

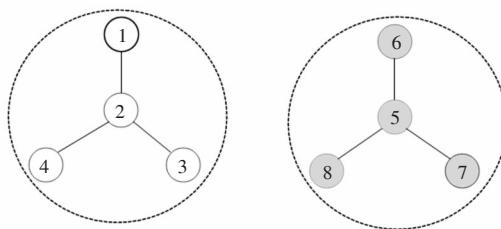


图 3 本文算法划分结果

针对上述步骤，有一种特殊情况需要说明，如果一个节点和它所有的邻居节点都没有共同的邻居节点，则这个节点和其它所有节点的相似度都等于 0，因此按照上述算法无法完成划分，这种情况下我们把这个节点划分到它邻居节点中度最大的那个节点所在的社团中。因为一个节点的度越大，它就越重要，新来的节点就倾向于和这个节点靠拢^[22]。同时为了提高划分社团的准确率，我们引进一个相似度阈值 ϵ ，当一个节点和它所有邻居节点的最大相似度小于 ϵ 时，我们就把这个具有最大相似度的节点划分到它所有邻居节点中度最大节点所在的社团中，因为如果相似度很小，依然使用节点的这个属性作为划分标准的话，容易过拟合。针对不同的网络这个 ϵ 也不同，我们可以通过调整 ϵ 的值来获得更好的划分效果。

3.2 算法性能分析

本算法需要时间的步骤主要包括三部分：计算节点相似度，寻找下一个需要处理的节点，合并社团。由于下一个要处理的节点是当前节点对应的相似度最大的节点，并且之前我们已经把这两个节

点存在一个以当前节点为头节点的链表中,因此可以直接获取到下一个需要处理的节点,因此这部分的计算时间可以忽略。计算节点相似度只需要知道它最近的邻居节点的信息,因此计算量非常小。计算式(2)的时间复杂度是 $O(1)$, 计算一个节点和它的 k 个邻居节点相似度的时间复杂度就是 $O(k)$, k 是网络中节点的平均的度。因此计算有 n 个节点的网络所需要的时间复杂度为 $O(kn)$ 。合并一个节点和它相似度的节点所需要的时间复杂度明显的是 $O(1)$ 。所以将网络中所有 n 个节点都合并花费的时间复杂度是 $O(n)$ 。使用 Dijkstra 算法计算网络中节点之间的最短路径的时间复杂度为 $O(n^2)$, 所以算法时间复杂度为 $O(k+n)n \approx O(n^2)$ ($k \ll n$)。而 CHM、ILM 算法的时间复杂度也同样为 $O(n^2)$ 。

4 实验结果与分析

4.1 Zachary 空手道俱乐部网络

空手道俱乐部网络是描述美国一所大学里空手道俱乐部里俱乐部成员之间关系的网络^[23], 该网络由 Zachary 耗时两年研究出来的。此网络含有 34 个节点和 78 条边, 节点表示俱乐部中的成员, 边表示俱乐部成员之间的关系。通过研究发现, 由于俱乐部主管 JohnA (node 34) 和俱乐部教练员 Mr.Hi (node 1) 之间产生分歧, 俱乐部被划分成了以主管和教练为首的两个团体。首先通过式(2)计算每个节点的和邻居节点的相似度, 然后找出最大相似度的节点, 结果如表 2。然后用本文算法对其进行划分, 图 4 描述了此次划分的过程 ($\varepsilon=0.88$)。由结果可以看出本文算法划分的结果与现实中的划分结果一致。

表 2 Zachary 网络中对应的最大相似度节点列表

Node	MSNs	Node	MSNs	Node	MSNs
1	2	13	1	24	34
2	1	14	1	25	32
3	1	15	34	26	32
4	1	16	34	27	34
5	1	17	6,7	28	34
6	1	18	1	29	34
7	1	19	34	30	34
8	1	20	1	31	34
9	33	21	34	32	34
10	33	22	1	33	34
11	1	23	34	34	33
12	1				

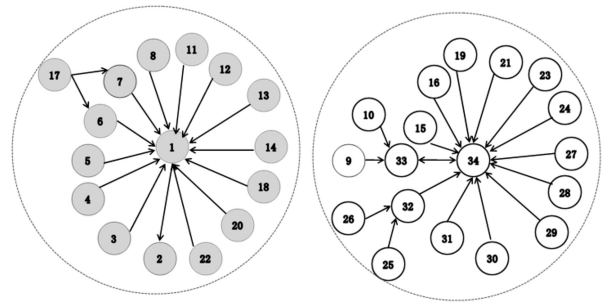


图 4 Zachary 网络的划分过程

表 3 表示将本文算法与 Newman 快速算法和 CNM 算法进行比较, 通过结果可以看出本文算法模块度更大, 准确率更高, 因此本文算法更适合划分 Zachary 空手道俱乐部网络。

表 3 不同算法的性能比较

算法	模块度	准确率
CHM 快速算法	0.498	0.97
ILM 算法	0.521	0.99
本文算法	0.563	0.99

4.2 美国大学足球赛网络

美国大学足球赛网络^[24]由 115 个节点 613 条边组成。每个节点代表一支球队, 每条边代表端点上的两支队伍进行过比赛。这些球队被划分成若干小组, 每个小组有 8~12 名球员, 同组内球队进行比赛的次数要比组间球队之间比赛的次数多。如图 5 ($\varepsilon=0.7$) 所示本文算法将网络划分 11 个社团, 其中 7 个社团和现实网络中的社团一致, 但是 CNM 算法只划分出 7 个社团, 其中划分完全正确的社团只有 4 个。所以本文提出的算法划分社团的准确率更高。表 4 表示的是本文算法与 Newman 快速和 CNM 算法的从模块度和准确率上对比的结果, 从数据上看, 本文算法要优于 Newman 快速算法和 CNM 算法。

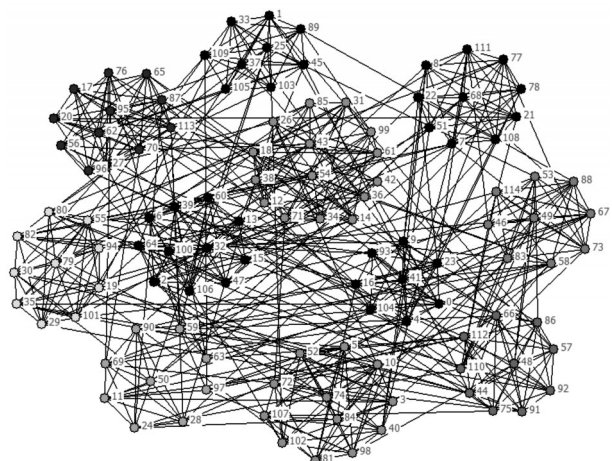


图 5 美国大学足球赛网络

表4 不同算法性能比较

算法	模块度	准确率
CHM算法	0.562	0.363
ILM算法	0.603	0.571
本文算法	0.617	0.636

4.3 海豚社会网络

Lusseau 等在新西兰对 62 只宽吻海豚的生活习性进行了长时间的观察,他们研究发现这些海豚的交往呈现出特定的模式,并构造了包含有 62 个结点的社会网络^[25]。如果某两只海豚经常一起频繁活动,那么网络中相应的两个结点之间就会有一条边存在。利用本算法划分的结果如图 6 ($\varepsilon=0.75$)所示。从图 6 可以看出,本文算法划分的结果与现实中的划分较为符合。

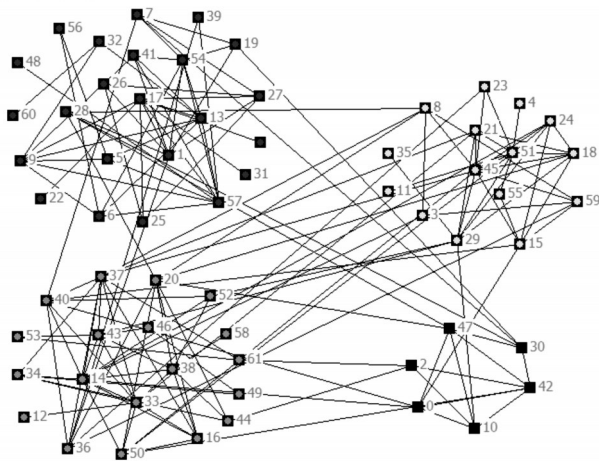


图6 海豚社会网络

5 结语

目前节点相似度的构造方法大多数是基于网络的局部信息,使用这种构造方法的节点相似度容易使算法陷入局部最优,使得划分结果和真实网络差异较大,针对这一缺点,本文综合网络局部信息和全局信息,重新定义了节点相似度,并设计了基于凝聚思想的划分算法,通过与 CHM 算法、ILM 算法的对比,在时间复杂度相当的情况下,经过仿真实验证明,本文算法可以更好地对复杂网络进行划分。因此,可以将本文算法用于复杂网络中的社团划分。

参考文献

[1] 骆志刚,丁凡,蒋晓舟. 复杂网络社团发现算法研究新进展[J]. 国防科技大学学报,2011,33(1):47-52.
LUO Zhigang, DING Fan, JIANG Xiaozhou. New Progress

on Community Detection in Complex Networks[J]. Journal of National University of Defense Technology, 2011, 33(1):47-52.

- [2] 解岱,汪小帆. 复杂网络中的社团结构分析算法研究综述[J]. 复杂系统与复杂性科学,2005,2(3):1-12.
XIE Zhou, WANG Xiaofan. An Overview of Algorithm for Analyzing Community Structure in Complex Networks[J]. Complex Systems and Complexity Science, 2005, 2(3): 1-12.
- [3] Chen Y, Wang X, Bu J, et al. Network structure exploration in networks with node attributes[J]. Physica A: Statistical Mechanics and its Applications, 2016, 449: 240-253.
- [4] Guimera R, Danon L, Diaz-Guilera A, et al. Self-similar community structure in a network of human interactions [J]. Physical review E, 2003, 68(6): 065103.
- [5] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(9): 2658-2663.
- [6] Zhou H. Distance, dissimilarity index, and network community structure [J]. Physical review e, 2003, 67(6): 061901.
- [7] Newman M E J. Detecting community structure in networks [J]. European Physical Journal B, 2004, 38(2): 321-330.
- [8] Clauset A. Finding local community structure in networks [J]. Physical Review E, 2005, 72(2): 26-132.
- [9] Schuetz P, Caflisch A. Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement [J]. Physical Review E, 2008, 77(4): 046112.
- [10] Frey B J, Dueck D. Clustering by passing messages between data points [J]. Science, 2007, 315: 972-976.
- [11] Symeonidis P, Tiakas E, Manolopoulos Y. Transitive node similarity for link prediction in social networks with positive and negative links [C]// Proceedings of the fourth ACM conference on Recommendersystems. ACM, 2010: 183-190.
- [12] H. Chen, L. Guo, X.L. Zhang, C.L. Giles, Proceedings of the 27th Annual ACM Symposium on Applied Computing [J]. 2012, 1: 138-143.
- [13] Li Y, Luo P, Wu C. Information loss method to measure node similarity in networks [J]. Physica A: Statistical Mechanics and its Applications, 2014, 410: 439-449.
- [14] Chen Z, Xie Z, Zhang Q. Community detection based on local topological information and its application in power grid [J]. Neurocomputing, 2015, 170: 384-392.
- [15] Salton G, McGill M J. Introduction to modern information retrieval [J]. 1986.

(下转第 240 页)

- SHI Wei, GAO Jun. An Improved Weighted Centroid Localisation Algorithm Based On Rssi For Wireless Sensor Networks [J]. Computer Applications and Software, 2015, 32(12):68-70, 104.
- [13] 张正华,戴磊,黎家文,等. 基于RSSI的优化加权质心定位算法研究[J]. 电子设计工程, 2013, 21(7): 171-173.
- ZHANG Zhenghua, DAI Lei, LI Jiawen, et al. Research on optimization weighted centroid localization algorithm based on RSSI[J]. Electronic Design Engineering, 2013, 21(7):171-173.
- [14] 崔法毅,邵冠兰. 基于RSSI多边形定位误差的加权质心定位算法[J]. 红外与激光工程, 2015, 44(7): 2162-2168.
- CUI Fayi, SHAO Guanlan. Weighted centroid localization algorithm based on multilateral localization error of received signal strength indicator[J]. Infrared and Laser Engineering, 2015, 44(7):2162-2168.
- [15] 王琦. 基于RSSI测距的室内定位技术[J]. 电子科技, 2012, 25(6):64-66, 78.
- WANG Qi. Research on an Indoor Positioning Technology Based on RSSI Ranging [J]. Electronic Science and Technology, 2012, 25(6):64-66, 78.
- [16] 刘信新,陈鲲. 基于RSSI与TDOA的混合测距加权定位算法[J]. 计算机与数字工程, 2011, 39(4):25-28.
- LIU Xinxin, CHEN Kun. A Weighted Localization Algorithm Using a Hybrid of RSSI and TDOA Range Technologies [J]. Computer and Digital Engineering, 2011, 39(4):25-28.
- [17] 任维政,徐连明,邓中亮,等. 基于RSSI的测距差分修正定位算法[J]. 传感技术学报, 2008, 21(7): 1247-1250.
- REN Weizheng, XU Lianming, DENG Zhongliang, et al. Distance Difference Localization Algorithm Based on RSSI for Wireless Sensor Networks [J]. Chinese Journal of Sensors and Actuators, 2008, 21(7): 1247-1250.
- [18] 刘运杰,金明录,崔承毅. 基于RSSI的无线传感器网络修正加权质心定位算法[J]. 传感技术学报, 2010, 23(5):717-721.
- LIU Yunjie, JIN Minglu, CUI Chengyi. Modified Weighted Centroid Localization Algorithm Based on RSSI for WSN [J]. Chinese Journal of Sensors and Actuators, 2010, 23(5):717-721.
- [19] 陈昌祥,达维,周洁,等. 基于RSSI的无线传感器网络距离修正定位算法[J]. 通信技术, 2011, 44(2): 65-66, 69.
- CHEN Changxiang, DA Wei, ZHOU Jie, et al. RSSI-based Range Collation Localization Algorithm in WSN[J]. Communications Technology, 2011, 44(2):65-66, 69.
- [20] 朱忠记,何熊熊,章晓,等. 基于RSSI的四边测距改进加权质心定位算法[J]. 杭州电子科技大学学报, 2014, 34(1):17-20.
- ZHU Zhongji, HE Xiongxiang, ZHANG Xiao, et al. The Quadrilateral and Improved Weighted Centroid Localization Algorithm Based on RSSI [J]. Journal of Hangzhou Dianzi University, 2014, 34(1): 17-20.
- [21] 王振朝,张琦,张峰,等. 基于RSSI测距的改进加权质心定位算法[J]. 电测与仪表, 2014, 51(21):63-66.
- WANG Zhenchao, ZHANG Qi, ZHANG Feng, et al. Improved Weighted Centroid Positioning Algorithm Based on RSSI Ranging [J]. Electrical Measurement & Instrumentation, 2014, 51(21):63-66.

~~~~~

(上接第 217 页)

- [16] Jaccard P. Etude comparative de la distribution florale dans une portion des Alpes et du Jura[M]. Impr. Corbaz, 1901.
- [17] Ravasz E, Somera A L, Mongru D A, et al. Hierarchical organization of modularity in metabolic networks[J]. science, 2002, 297(5586): 1551-1555.
- [18] Leicht E A, Holme P, Newman M E J. Vertex similarity in networks [J]. Physical Review E, 2006, 73(2): 026120.
- [19] Zhou T, Lü L, Zhang Y C. Predicting missing links via local information [J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2009, 71(4): 623-630.
- [20] Liu J G, Zhou T, Che H A, et al. Effects of high-order correlations on personalized recommendations for bipartite networks [J]. Physica A: Statistical Mechanics and its Applications, 2010, 389(4):881-886.
- [21] Symeonidis P, Tiakas E, Manolopoulos Y. Transitive node similarity for link prediction in social networks with positive and negative links[C]//Proceedings of the fourth ACM conference on Recommender systems. ACM, 2010: 183-190.
- [22] Barabási A L, Albert R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439):509-512.
- [23] Zachary W W. An information flow model for conflict and fission in small groups[J]. Journal of anthropological research, 1977, 33(4):452-473.
- [24] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proceedings of the national academy of sciences, 2002, 99(12):7821-7826.
- [25] Barabási A L, Albert R. Emergence of scaling in random networks[J]. science, 1999, 286(5439):509-512.