文章编号：2096-1472(2018)-10-07-03
DOI:10.19644/j.cnki.issn2096-1472.2018.10.002

基于差分进化的社交网络可视化研究

毕璐琪，杨连贺
(天津工业大学计算机科学与软件学院，天津 300387)

摘　要：社交网络对于个人及社会的重要性日益凸显。随着社交网络数据规模的不断扩大，如何清晰美观地展现社交网络关系结构成为信息可视化领域研究的一大难点。针对此研究难点，本文应用网络理论和实验领域的专家之间的合作关系数据集，通过度中心性、介数中心性指标发现数据中的关键节点，改进差分进化算法的变异、交叉和选择过程，提出了基于差分进化的社交网络可视化布局算法，有效减少初始位置对可视化结果的影响，并且最终呈现的可视化结果可以清楚美观地展现社交网络结构。

关键词：社交网络；可视化；差分进化；关键节点
中图分类号：TP391.9　**文献标识码**：A

Study on Visualization of Social Network Based on Differential Evolution

BI Luqi,YANG Lianhe
(*School of Computer Science & Software Engineering*,*Tianjin Polytechnic University*,*Tianjin* 300387,*China*)

Abstract: Social networks have become increasingly prominent for both individuals and the society.As social network data continues to grow in size, how to clearly and attractively display the social network relationship structure has become a major difficulty in the field of information visualization. In view of the difficulty of this research,this paper applies the cooperation relationship data between experts in network theory and experimentation to find key nodes in the data through degree-centrality and betweenness-centrality indicators to improve the variation,crossover and selection of differential evolution algorithms.Therefore,a social network visual layout algorithm based on differential evolution is proposed,which effectively reduces the impact of the initial position on the visualization results.The visual results presented finally can clearly and beautifully reflect the social network structure.

Keywords:social network;visualization;differential evolution;key nodes

1 引言(Introduction)

当今时代，社交无处不在。随着通讯技术的不断进步，社交形式更加趋于多样化，其中包括面对面的人际交往型社交、网络平台如微博、微信、电子邮件等线上互动型社交。在大数据和人工智能的时代背景下，对海量社交网络数据的分析理解至关重要，因为它有利于理清个人及群体之间的联系，在好友推荐、个性化服务、舆情控制和信息传播等方面发挥重大作用。

随着数据规模的不断扩大，人们对实用性和美观性的要求越来越高。在实用性上，必须提高布局算法的效率，尽可能在保持结构的前提下达到全局优化；在美观性上，节点和边应均匀分布，尽量减少边的交叉，整体效果应对称，等等。

本文针对无向图，结合关键节点检测指标识别关键节点，结合差分进化算法较强的全局收敛和鲁棒性的优点，以及力导引算法布局美观、充分展现网络数据自身结构的优点提出差分进化布局算法，可有效降低初始位置对可视化结果的影响，使系统稳定的同时，减少视觉混乱，得到美观性和实用性兼具的可视化结果。

2 相关研究(Related research)

社交网络可视化是信息可视化的一个重要领域，社交网络可视化的核心是节点布局问题，节点布局既要求符合社交网络的自身结构，也要求清晰美观的效果。因社交网络具有小世界和无尺度的特点，为使社交网络的节点在有限空间内合理分布，布局算法的选择至关重要[1]。最常用的布局方法为节点-链接法。其中节点-链接法最常用的布局算法是力导引布局算法，最早由Eades提出，他将社交网络假设成一个物理系统，节点为钢环，链接为弹簧，用弹簧模拟两个点之间的关系，在弹力的作用下节点的位置不断移动，经过多次

迭代，布局达到动态平衡状态[2]。此后，Kamada等人基于力导引算法，以整个系统能量最小为准则确定节点的位置，从而提出KK算法[3]。Fruchterman等人在粒子物理学原理的基础上，通过计算所有节点之间的作用力来确定节点的具体位置，提出FR布局算法[4]。刘芳等提出基于粒子群优化的布局算法，设计了适应社交网络布局的目标函数，减少边交叉，用曲线替代直线，使布局效果更清晰[5]。

差分进化算法(Differential Evolution,DE)是一种高效的启发式搜索算法[6]，具有控制参数少、收敛快、优化结果稳健等优点，并在神经网络优化、机器智能、医学等工程领域获得了广泛应用[7]。同时，差分进化算法在可视化领域也有应用，如YUE等人研究了基于差分进化算法构建地理信息可视化建模的环境[8]。关于差分进化的优化研究，Skanderova等探索了基于复杂网络对差分进化动力学进行建模[9]。研究表明，差分进化算法对于网络数据的可视化是可行且有效的。

3 差分进化布局算法(Differential evolution layout algorithm)

社交网络图通常用G(Graph)表示，G由节点集N和边集E组成，即$G=(N,E)$。本文用到的相关符号及说明详见表1。

表1 相关符号及说明

Tab.1 Related symbols and description

符号	说明
Nodes(N)	节点集
Edges(E)	链接集
n	节点个数
pos	节点坐标集

3.1 关键节点检测

(1)度中心性

对无向图而言，节点的度(Degree)表示与此节点相连的节点个数。节点的度越大，与之相连的节点越多，表示其在社交网络中越重要，传播信息的能力越强。度中心性(Degree Centrality)描述节点度的大小，用C_D表示，其表达式为：

$$C_D=\frac{Degrees(i)}{n-1} \tag{1}$$

其中，$Degrees(i)$表示节点i的度，即与之相连的节点个数。分母$n-1$表示该节点最大的度值，即除自身外与其他$n-1$个节点均相连。

(2)介数中心性

在网络拓扑结构中，最短路径是两个节点之间可选路径中花费时间精力最小的路径。而节点i在其他连通节点间最短路径的比重大小表示为介数中心性(Betweenness Centrality)，用C_B表示，节点i的介数中心性表示为：

$$C_{Bi}=\sum_{s\neq i\neq t\in Nodes}\frac{\alpha_{st}(i)}{\alpha_{st}} \tag{2}$$

其中，α_{st}表示节点s和t之间最短路径的个数，$\alpha_{st}(i)$表示节点i在节点s到t的最短路径中的次数。节点的介数中心性值越大，其在网络中媒介的作用越大。

3.2 差分进化布局算法

差分进化算法是一种启发式优化算法，广泛应用于复杂问题的优化。差分进化算法的流程为：初始化—变异—交叉—选择。可将社交网络可视化布局视为一个无约束优化问题，结合力导引布局算法的思想，将节点视为一个个原子，原子间的作用力有斥力F_r和引力F_a，任意两个节点间存在斥力，两个相连接的节点间存在引力，最终的布局使得整个系统处于动态平衡状态。基于布局算法中节点和边应均匀分布的美学标准，本文设置适应度函数为节点和边分布的偏差函数：

$$deviation\_edge=\frac{max(len)-min(len)}{avg(len)\times num} \tag{3}$$

$$deviation\_node=\frac{\left|k-min(distance)\right|}{area(G)} \tag{4}$$

$deviation_edge$表示相连节点之间边长度的偏差值，其中len表示边的长度，分子即最大边长与最小边长的偏差值，num表示边的总数。$deviation_node$表示节点间距离的偏差值，其中$distance$表示节点间距离，$area(G)$表示布局大小，k表示理想节点距离，计算公式为：

$$k=\frac{area(G)}{n} \tag{5}$$

适应度函数F的计算公式为：

$$F=deviation\_edge+deviation\_node \tag{6}$$

差分进化布局的算法流程为：

(1)初始化

首先初始化图G，随机初始化节点坐标，获得初始适应度函数值F_0，以及初始节点坐标集pos^0。

(2)变异

变异过程结合力导引算法基本思想，将节点看作原子，将两点之间的链接视为弹簧，通过计算两个相连节点之间的引力和任意两个节点间的斥力来确定节点的位置。节点i和节点j之间的位置差表示为$d(i,j)$：

$$d(i,j)=pos(i)-pos(j) \tag{7}$$

经过变异过程，节点i的位移$d(i)$为：

$$d(i)=\sum_{i\neq j\in N}\frac{d(i,j)f_r(|d(i,j)|)}{|d(i,j)|}-\sum_{i\neq k,(i,k)\in E}\frac{d(i,k)f_a(|d(i,k)|)}{|d(i,k)|} \tag{8}$$

为防止置换超出边界，应用变异算子H对位移进行控制，节点i经变异后坐标更新为：

$$pos(i)^1=pos(i)^0+\frac{d(i)}{|d(i)|}\times H \tag{9}$$

变异算子控制节点坐标不超出边界，这里取$H=0.8$。

经变异过程后，节点的坐标更新为pos^1。

(3)交叉

将变异后的节点坐标与初始节点坐标进行参数混合，得到坐标集pos^2。参数为交叉算子$CR(CR\in[0,1])$，交叉算子用于控制节点各坐标值对交叉的参与度。

$$pos^2(i)=\begin{cases} pos^1(i), if\ \ rand(0,1)\le CR \\ pos^0(i), \text{otherwise} \end{cases} \tag{10}$$

由式(10)可以看出，CR越大，交叉后节点坐标为pos^1的几率越大，使得布局算法收敛较快；反之，节点坐标为初始坐标pos^0的几率越大，坐标更新速度整体减慢，更有利于全局优化。所以权衡利弊后，取$CR=0.68$。

交叉过程后计算得到适应度值F_1。

(4)选择

设置阈值ε，如果$F_1\le F_0$，即变异交叉后的总偏差值小于初始时的总偏差值，则在下一次迭代中F_1更新为初始适应度值F_0，初始坐标更新为pos^2；若$F_1\le\varepsilon$，则结束迭代，最终节点坐标为pos^2。

若$F_1>F_0$，则开始新一轮变异−交叉−选择；同时，还可能存在迭代停滞，即适应度值在迭代过程中没有改变，且不满足阈值条件。为避免此情况，需要设置一个标记$flag$，取$flag$为当前迭代后的F_0。若下一次迭代后$F_0=flag$，则再下一次迭代从初始化操作开始，重复迭代，直到满足阈值条件或迭代完成。

4 数据分析及仿真结果讨论(Data analysis and simulation results discussion)

4.1 数据来源

本文使用的数据来源于M.E.J.Newman的一篇论文[10]，数据集的内容是网络理论和实验领域的专家之间的合作关系，包括1589个节点、2742条边，其中包含边的权重值，边权重值表示相连的两节点即两位专家之间合作关系的强弱。

4.2 数据分析

度中心性衡量节点在社交网络中的整体中心性，与之相连接的节点越多，在社交网络中越占据中心位置。介数中心性表示节点的媒介程度，其处于各连通节点的最短路径次数越多，表明该节点的活跃程度越高。这两个中心性指标都表明节点在社交网络图中的重要地位。根据度中心性和介数中心性的计算公式，对数据节点进行整理，取度中心值+介数中心值前五位的节点数据，详见表2。

表2 度中心值+介数中心值前五位的节点

Tab.2 The top five nodes of degree−centrality & betweenness−centrality

id	姓名	度中心值+介数中心值
78	NEWMAN，M	0.03946189
34	JEONG，H	0.03117465
33	BARABASI，A	0.03000883
150	PASTORSATORRAS，R	0.02896270
216	BOCCALETTI，S	0.02640828

经查阅表2中关键节点得知，表中大部分科学家都是在网络理论和实验领域中相当活跃的。

分析处理上述数据时发现，有128个节点度中心值和介数中心值为0，即这些节点不与其他节点有连接，在数据集中的活跃度基本为0，所以最终结果中对这些节点不予以展示。

4.3 仿真结果分析

首先根据节点的度中心值和介数中心值得到节点的权重集，根据节点权重集调整节点大小和颜色，权重最大的节点尺寸最大、颜色最深，依次类推。

根据边的权重值，将权重值小于0.5(即弱合作关系)的边用虚线表示，并且设置虚线透明度为0.7；权重值不小于0.5(即较强合作关系)的边用实线表示，透明度为1。

设置不同的迭代次数，可以看到布局的变化。设置迭代次数为50、100、200，得到的可视化布局图如图1所示。

(a)迭代次数为50　(b)迭代次数为100　(c)迭代次数为200

图1 不同迭代次数的可视化结果

Fig.1 Visualization results for different iterations

随着迭代次数的增加，边的分布趋于规整；部分节点关系紧密，独立成团；节点的分布趋于中心化发散，并且中心位置的节点更加集中，易造成节点重叠。

差分进化布局算法的适应度函数根据节点和边分布偏差设置，随迭代次数增加产生的适应度函数值F的折线图如图2所示。

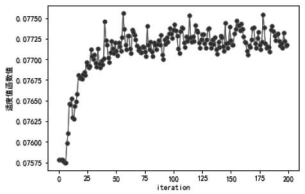

图2 适应度函数迭代结果

Fig.2 Fitness function iterative results

从图2中可以看出，F值随着迭代次数增加呈上升趋势，并且自迭代次数超过50后，F值在[0.7700,0.7750]内的动态变化。

随着迭代次数的增加，可以明显看到网络图中心位置的重叠节点逐渐增多。为避免中心位置过多节点重叠，迭代次数应控制在100—150次。当迭代次数为120时，节点重叠相对最少，适应度函数F=0.7704，边和节点的分布均匀且整体对称；所以最终将可视化结果的迭代次数设置为120，如图3所示。

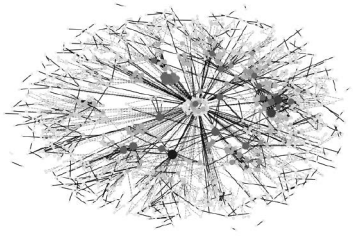

图3 迭代次数为120的可视化结果

Fig.3 Visualization result with 120 iterations

从可视化结果可以看出：

(1)权重系数大的节点(即尺寸大颜色深的节点)周围链接较多，且周围的实线较多，表示其在社区网络中的地位重要。节点权重系数小的节点(即尺寸小颜色浅的节点)多分布在最外围，符合社交网络结构。

(2)设置不同迭代次数得到的结果均发现有很多独立成团的节点集合，即这些节点之间合作关系紧密。

(3)整体布局呈现中心发散对称，节点和边分布较为均衡，不混乱。

5 结论(Conclusion)

本文针对无向图，提出了基于差分进化的社交网络布局算法，变异过程结合力导引算法思想更新节点坐标，将适应度函数设置为偏差函数。首先根据中心性指标对节点属性进行区分，根据边权重值区分链接的强弱，最后应用差分进化布局算法得到的最终可视化结果。

研究结果表明，本文提出的差分进化布局算法可以在保持网络结构的同时，通过一次迭代多次更新节点坐标有效降低了初始位置对可视化结果的影响，并且符合社交网络节点和边分布均衡的美学标准，有效降低了视觉混乱。在后续的研究中，还需提高对海量规模社交网络数据布局的效率，并将进一步学习研究三维社交网络可视化。

参考文献(References)

[1] 孙扬,蒋远翔,赵翔,等.网络可视化研究综述[J].计算机科学,2010,37(02):12-18;30.

[2] EADES P.A heuristic for graph drawing[J].Congressus Nutneran tiunt,1984,42(11):149-160.

[3] KAMADA T,KAWAI S.An algorithm for drawing general undirected graphs[J]. Information Processing Letters(Elsevier),1989,31(1):7-15.

[4] FRUCHTERMAN T M J,REINGLOD E M.Graph drawing by force-directed placement [J]. Software Practice and Experience(Wiley),1991,21(11):1129-1164.

[5] 刘芳,孙芸,杨庚,等.基于粒子群优化算法的社交网络可视化[J].浙江大学学报(工学版),2013,47(01):37-43.

[6] Storn R, Price K.Differential Evolution-A Simple and Efficient Heuristic for global Optimization over Continuous Spaces[J]. Journal of Global Optimization,1997,11(4):341-359.

[7] 肖婧.差分进化算法的改进及应用研究[D].哈尔滨:哈尔滨工程大学,2011.

[8] Liu G,Liu K,Wu C.Constraint solving approach based on Differential Evolution Algorithm for geo-visualization variant design method[C].International Conference on Natural Computation.IEEE,2010:2311-2315.

[9] Skanderova L,Fabian T.Differential evolution dynamics analysis by complex networks[J].Soft Computing,2017:1-15.

[10] Newman M E J.Finding community structure in networks using the eigenvectors of matrices[J].Physical review.E (Statistical,nonlinear,and soft matter physics),2006,74(3 Pt 2):036104.

作者简介：

毕璐琪(1993-)，女，硕士生.研究领域：信息可视化.

杨连贺(1965-)，男，博士，教授.研究领域：计算机仿真与辅助设计，数据挖掘.

(上接第13页)

5 结论(Conclusion)

通过对实验结果的分析，相对于经典的向量空间模型(VSM)，本文提出的基于公共词集对长篇小说相似度的计算方法在正确率和召回率有一定的提高，并且对于特点鲜明的小说类型，比如儿童类型小说、冒险类型小说等，效果尤其显著。

当然本文提出的方法也有不足之处，比如对构成元素复杂的小说，相似度的衡量不够准确。如果进一步添加对两篇小说在词集和词序上的不同之处的计算模型，并且影响最终的相似度计算，这样可以进一步提高相似度衡量的精确性。

参考文献(References)

[1] 黄承慧,印鉴,侯昉.一种结合词项语义信息和TF-IDF方法的文本相似度量方法[J].计算机学报,2011,34(5):856-864.

[2] 刁力力,胡可云,陆玉昌,等.用Boosting方法组合增强Stumps进行文本分类[J].软件学报,2002,13(8):1361-1367.

[3] 徐戈,王厚峰.自然语言处理中主题模型的发展[J].计算机学报,2011,34(8):1423-1436.

[4] Mariona Coll Ardanuy,Caroline Sporleder.Structure-based Clustering of Novels.

[5] 孟宪军.互联网文本聚类与检索技术研究[D].哈尔滨工业大学,2009.

[6] 黄贤英,刘英涛,饶勤菲.一种基于公共词块的英文短文本相似度算法[J].重庆理工大学学报(自然科学版),2015,29(8):88-93.

[7] 欧阳宁,罗艳.基于领域特征词加权的文本相似度计算[J].计算机工程与设计,2012,33(11):4338-4342.

[8] 张焕炯,王国胜,钟义信.基于汉明距离的文本相似度计算[J].计算机工程与应用,2001,37(19):21-22.

作者简介：

郭 涛(1996-)，男，本科生.研究领域：数据挖掘.

霸元婕(1997-)，女，本科生.研究领域：数据挖掘.

李绍昂(1997-)，男，本科生.研究领域：数据挖掘.