ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.9, September 2010, pp.2250−2261
doi: 10.3724/SP.J.1001.2010.03889

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种多变元网络可视化方法*

孙 扬[1+], 赵 翔[2], 唐九阳[1], 汤大权[1], 肖卫东[1]

[1](国防科学技术大学 C[4]ISR 技术国防科技重点实验室,湖南 长沙 410073)

[2](School of Computer Science and Engineering, University of New South Wales, Sydney NSW 2052, Australia)

Multivariate Network Visualization Paradigm

SUN Yang[1+], ZHAO Xiang[2], TANG Jiu-Yang[1], TANG Da-Quan[1], XIAO Wei-Dong[1]

[1](C[4]ISR Technology National Defense Key LAB, University of Defense Technology, Changsha 410073, China)

[2](School of Computer Science and Engineering, University of New South Wales, Sydney NSW 2052, Australia)

+ Corresponding author: E-mail: victor_830514@yahoo.com.cn

Sun Y, Zhao X, Tang JY, Tang DQ, Xiao WD. Multivariate network visualization paradigm. *Journal of Software*, 2010,21(9):2250−2261. http://www.jos.org.cn/1000-9825/3889.htm

Abstract: This article proposes a multivariate network visualization paradigm, MulNetVisBasc. Advanced Start Coordinates (ASC) are employed to place nodes on the basis of multivariate attributes and to devise an algorithm that that incorporates edge-merging and routing techniques to automatically lay-out edges; furthermore, a user-friendly human-computer interface is developed to assist users in further data analysis and mining. The experimental results suggest that the visualization of MulNetVisBasc not only uncovers the multivariate distributional characteristics of datasets intuitively, but also displays the associations of networks clearly and is helpful in discovering the implicit knowledge hidden behind datasets. The edge layout algorithm reduces the visual clutters caused by edge crossing and is suitable for relatively huge multivariate network datasets in virtue of its low complexity. Finally, the human-computer interface is flexible and convenient.

Key words: multivariate network visualization; Elamer; advanced star coordinates; network visualization; multivariate data visualization; information visualization

摘 要: 提出一种多变元网络可视化方法 MulNetVisBasc,根据节点的多变元属性,使用高级星形坐标法布局网络节点,以边融合及路由技术为基础设计算法,自动有效布局网络边,实现友好的人机交互界面辅助用户进一步对数据进行分析挖掘.实验结果表明,MulNetVisBasc 的可视化结果能够在直观揭示数据集多变元分布特性的同时清晰展现其网络关联特性,有助于用户发掘多变元网络数据集中潜在的隐性知识.边布局算法能够有效减少视图中的边交叉数量,且复杂度较低,适用于较大规模数据集,人机交互界面灵活方便.

关键词: 多变元网络可视化;基于边融合及路由技术的边布局算法 Elamer;高级星形坐标法;网络可视化;多变元可视化;信息可视化

* Supported by the National Natural Science Foundation of China under Grant No.60903225 (国家自然科学基金); the Excellent Graduates Innovation Foundation of University of Defense Technology of China under Grant No.B080503 (国防科学技术大学优秀研究生创新基金)

Received 2009-10-14; Revised 2010-01-20; Accepted 2010-06-10

中图法分类号: TP391 文献标识码: A

多变元网络是节点(或边)具有多变元属性的结构网络,这种数据组织结构普遍存在于各应用领域,如社会网络中以网络节点表示社会个体成员,网络边代表成员间的关系,而每个社会成员都同时拥有自己的属性信息(姓名、年龄、性别等);生物学中的蛋白质分子交互网络以节点代表蛋白质分子,边表示分子间的相互作用,而每个分子又都包含了名称、功能等多变元信息.对于各类多变元网络数据,研究人员的需求不只局限于单纯分析其多变元特性或网络特性的信息,他们通常还会提出同时分析二类特性以揭示数据集的多变元统计特征与网络结构的内在联系的要求.如社会学家经常会提出拥有相似特点的人是否会联系紧密的疑问,分子生物学家对相互作用的蛋白质分子是否具有相似功能十分感兴趣.但是,由于多变元数据分析与网络数据分析的目的、方法、手段、过程差别非常大,因此,目前尚不存在相应的分析技术辅助研究人员有效地研究、探索、发掘及理解大量抽象复杂的多变元网络数据.另一方面,信息可视化技术作为抽象信息有效地展现分析工具日益得到了研究人员的认可,并且已分别在多变元数据和网络数据方面进行了较深入的研究.但二者的关注点不同:多变元可视化方法侧重于构建能够保持原多维数据拓扑结构的低维展现,以辅助用户在可视空间中分析各多变元对象间的相互关系,主要方法包括枝形图[1]、平行坐标系[2]、星形坐标系[3]、多维标度法(MDS)[4]、等距映射法(ISOMAP)[5]等;而网络可视化方法则强调通过优化网络图布局增强节点间关联信息的可理解性及易读性,重要的评价指标有网络图的对称性、网络边的交叉数量等[6].Eades 提出的力导引布局方法是一种经典的网络可视化布局方法[7].文献[8,9]分别对两类信息可视化方法进行了详细的综述.

然而,信息可视化技术领域对于能够同时展现多变元网络双重特性的可视化方法的研究仍然停留在较为简单的层次:首先使用网络可视化方法布局节点,然后在其上叠加多变元展现形式,如使用节点(边)的可视化特征(如形状、大小、颜色等)、将节点直接替换为枝形图、在节点上附加类似平行坐标系的方式来展现网络节点(边)的多变元属性.Becker 等人使用长方形节点的宽和高表示美国电话网络中部分主要城市的电话主叫及被叫数量[10],Stephen 等人使用节点大小和颜色分别代表 e-mail 网络中邮箱节点的邮件数量及所有者的职位信息[11],Xu 等人在已布局的网络节点上叠加多个 landscape 曲面展示其多变元属性[12].此类方式虽然能够较好地体现多变元网络中的网络结构特征,但也存在很多局限性,如变元可视数量较少、对多变元特性揭示不直观、容易引起交叠问题等.因此,仅在网络布局的基础上叠加多变元可视技术是不够的,需要在网络布局的同时即考虑其多变元特性.Wu 等人将涉及多变元属性的测地自组织映射法(GeoSOM)与节点布局过程结合以可视化多变元网络,使其网络布局能够体现出数据集的聚类及奇异点等特征[13,14].但是,由于 SOM 方法本身固有的缺陷,该方法计算复杂度较高,无法直观体现数据对象的维分布情况;并且在节点位置固定的情况下,该方法也未考虑通过优化边布局解决可视混乱问题.

因此,本文提出一种新的多变元网络可视化方法 MulNetVisBasc(multivariate network visualization based on advanced star coordinates).首先,应用高级星形坐标法(ASC)将多变元网络节点降维投影到多变元空间对应的低维参考坐标系中;然后,基于边融合及路由思想提出 Elamer 算法自动布局网络边,以尽量减少边交叉数量;最后,设计实现特色而友好的交互方式以辅助用户对多变元网络数据集进行可视分析.实验分析表明,MulNetVisBasc 产生的可视化效果易于理解,能够在直观揭示数据集的多变元特性(尤其是维分布信息)的同时清晰展现其网络关联特性,有助于发掘多变元网络数据集中潜在的隐性知识.Elamer 算法可以大幅度减轻因边交叉引起的可视混乱,且时间复杂度较低,适用于规模较大的数据集,人机交互界面灵活方便,用户满意度较高.

1 预备工作

1.1 假设与符号

虽然多变元网络的节点和边都可能具有多变元属性,但本文只讨论网络节点具有多变元性的情况.为便于全文论述,首先给出如下假设说明及定义:

假设说明.严格地定义,多属性数据集中相关属性才能称为“变元(variate)”,而相互完全独立(正交)的属性应该称作“维(dimension)”.但是,我们也可认为维是一类特殊的变元(相互间存在正交性).因此,本文为使论述更加符合读者的阅读习惯,对变元和维不进行详细区分.

定义 1. 多变元网络 G 由网络节点、边及节点具有的变元集组成,可形式化描述为 $G=\langle V,E,A\rangle$.其中,V 表示节点集合,E 代表边集合,A 是 V 中元素属性值项 $F(V)=(f_1,f_2,\ldots,f_m)$的集合.$V$,$A$ 的势相等,即$|A|=|V|$.

定义 2. 假定 G 中节点 $V=\{V_1,V_2,\ldots,V_n\}$来自相同的应用领域,由一组相同的数值型变元(属性)$A_1,A_2,\ldots,A_m$进行描述,且每个节点 V_i 都包含完整的 m 个属性值,不存在变元值缺失情况.

定义 3. G 中无向(或有向)边 $E=\{e_1,e_2,\ldots,e_l\}$是 V 中元素的无序偶或有序偶$\langle V_i,V_j\rangle$的集合,并且称与节点 V_i 相连边的条数为 V_i 的度,记为 $d(V_i)$.

1.2 高级星形坐标法(advanced star coordinates,简称ASC)

ASC 是我们前期提出的一种多变元可视化方法[15],它改进了星形坐标法(SC)降维过程信息损失较为严重、可视化结果未体现多变元数据集维分布信息、手动配置维度轴繁杂、耗时的不足.如图 1 所示,ASC 使用圆形中沿直径方向的向量 $\overrightarrow{D_{sj}D_{ej}}$ 作为维度轴表示多变元数据集的一维,将多变元数据集中所有数据任意两维对应数值近似相等的数量定义为相应两维的相关系数,并根据各相关系数值排列维度轴,数值越大,二维度轴夹角越小.而后,依据变元间的语义相关性标定维度轴正向,完成高级星形坐标系的构建;ASC 使用平面直角坐标系中的一点 $F^i(x_i,y_i)$代表多变元对象 $F(V_i)$,其属性值 $(f_1^i,f_2^i,\ldots,f_m^i)$ 定义为 $F(V_i)$在原多维空间中的维坐标.将 $F^i(x_i,y_i)$在高级星形坐标系各维度轴的投影点在相应维度轴上的坐标定义为 $F(V_i)$的可视坐标,并以减小相应多变元对象的可视坐标与维坐标的差别为准则,使用最优化方法求解全部对象的可视坐标,实现对用户有意义的降维运算,将抽象的多变元对象映射到低维可视空间-高级星形坐标系中.ASC 的降维算法效率较高,适用于数据量较大、维数较高的数据集合,可视化结果易于理解.与 SC 方法相比,提高了手动配置维度轴的效率,缩短了发掘有效结论的时间,减少了简单降维过程引入的信息损失,体现了多变元数据集合丰富的维分布信息.

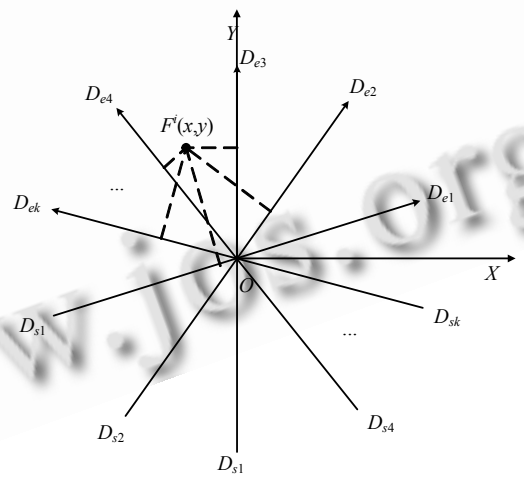


Fig.1 ASC multivariate data visualization model

图 1 ASC 多变元可视化模型

2 基于 ASC 的多变元网络可视化方法(MulNetVisBasc)

MulNetVisBasc 方法是根据 Card 提出的信息可视化参考模型[16]设计的多变元网络可视化方法(图 2).首先组织原始数据集形成规范的多变元网络节点及边的数据表;然后计算多变元节点数据表中各维度的相关系数,并以此为基础绘制高级星形坐标系,使用 ASC 降维算法将多变元节点投影到高级星形坐标系中;通过高效的边

融合及边路由算法合理布局网络边,以尽量减少边交叉数量,清晰展现节点间关联信息;而后,对形成的多变元网络可视化结构进行渲染,使用样条曲线绘制网络边,使用不同颜色按照不同分类标准标注节点,确保可视化视图美观,易于理解;最后,设计友好的人机交互操作对可视化过程进行控制,以辅助用户更加直观、准确地分析、发掘相关隐性信息.

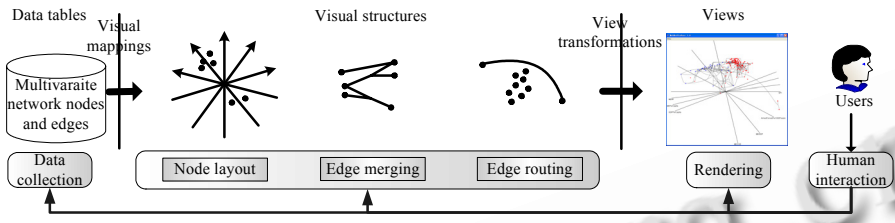


Fig.2 MulNetVisBasc information visualization framework

图 2 MulNetVisBasc 可视化框架

2.1 基于ASC的节点布局

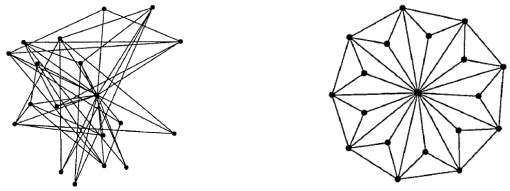
首先,不考虑多变元网络数据中的关联信息,只抽取其中由所有节点属性值形成的多变元数据集,并以此为数据基础,利用 ASC 方法绘制高级星形坐标系.然后,将多变元网络节点降维映射到高级星形坐标系中,并在二维平面中计算全部节点相互之间的欧式距离组成距离矩阵,而后以距离矩阵作为输入使用 k-means 方法对多变元网络节点进行聚类(ASC 方法最主要的优点在于能够清晰、直观地揭示多变元数据集的聚类信息及其维分布信息,如果在可视化视图中未形成簇状分布,则对相应数据集 ASC 方法就丢失了它的使用价值.因此,本文假定降维后的多变元网络节点也会在高级星形坐标系中形成多簇聚类,而此处的 k 值可由用户根据可视化效果图交互设定).最后,计算每个聚类的长方形最小包围盒,为下一步进行边融合及路由布局做准备.

2.2 基于边融合及路由技术的边布局算法——Elamer

传统网络可视化方法主要通过各种算法布局节点使产生的网络图能够尽量符合诸多绘图美学标准[17],从而提高网络结构信息、关联信息的可读性,如图 3 所示.而 Purchase 在研究中发现,最重要的绘图美学标准是边交叉数量最小原则[18].但 MulNetVisBasc 为能展现数据集的多变元分布特性,在未考虑节点关联信息的情况下使用 ASC 方法布局网络节点,使得网络节点的位置相对固定,无法进行大幅度调整,如果直接将相关节点以直线相连势必会产生相当数量的交叉,引起多变元网络图的可视混乱.因此,为能够清晰表现节点间的关联信息就需要重点研究网络边的布局,以减少多变元网络图中边交叉的数量.但是,使用最优化方法求解最小化网络边交叉的布局已被证明是一个 NP 难问题[19],因此,我们只能设计尽量减少边交叉数量的方法.恰当地使用边融合及路由技术能够极大地减少网络图中的交叉,如图 4 所示,通过将拥有共同端点的边部分进行融合,可以减少边交叉的几率;在 ASC 产生的多簇布局中,若非相邻簇节点连线穿越其他簇,则会产生很多交叉,若将其绕行(路由)即可达到较好的减少边交叉数量的效果,如图 5 所示.

边融合技术是计算机视觉研究中的重要内容[20],而边路由方法是 VLSI、计算几何、机器人路径设计领域中的热点问题[21].Doantam 等人首次将二者结合应用于流图(flow map)边的自动绘制[22],他们以流图节点的距离矩阵为基础,使用聚合式层次聚类法将所有节点组织为二叉树形式;然后“根化”流图源点对该二叉树进行重新组织,其边融合效果由父亲节点连接其子节点时进行分叉实现,边路由算法则应用于父亲节点连接子节点时穿越子节点的同层兄弟节点的情况,该方法对单源流图的可视化效果较好.虽然我们可将多变元网络近似“拆分”为多层单源流图,并且 Doantam 也使用压条法(layering)实现了多源流图(可认为多层单源流图的重叠)边的自动绘制,但是由于该方法对多源流图绘制的效果与单源流图相比存在较大差距,而且在使用压条法时需要对每一源点进行重新“根化”,“根化”算法复杂度又相对较高 $O(n^2)$,不适于处理多变元网络中“源点”较多的情况,因此,本文不能直接套用该方法进行网络边的布局,但是该方法的部分思想可借鉴用于本文边布局算法中的边融

合及路由过程.



(a) A graph with a bad layout (b) Improved layout of the graph in Fig.3(a)
(a) 布局较差的网络图 (b) 布局改进后的网络图

Fig.3 Importance of drawing aesthetics in drawing graph
图 3 绘图美学标准在网络图绘制中的重要性

Fig.4 Edge merging reduces edge crossing
图 4 边融合技术能够减少边交叉

Fig.5 Edge routing reduces edge crossing
图 5 边路由技术能够减少边交叉

为适应多变元网络的结构特点及 MulNetVisBasc 的节点布局特点,实现多变元网络边的有效自动布局,本文提出 Elamer 算法(edge layout algorithm based on edge merging and routing).算法分为 3 部分:簇间网络边融合(edge merging between clusters,简称 EMBC)、簇间网络边路由(edge routing between clusters,简称 ERBC)及簇内网络边融合(edge merging inner cluster,简称 EMIC).簇间网络边融合解决拥有相同端点的不同簇间节点连线的融合分叉问题,算法思想如图 6 所示;簇间网络边路由解决非相邻簇节点连线可能穿越兄弟簇的问题,算法思想如图 7 所示;簇内网络边融合解决相同簇内节点连线的融合分叉问题,算法思想如图 8 所示.

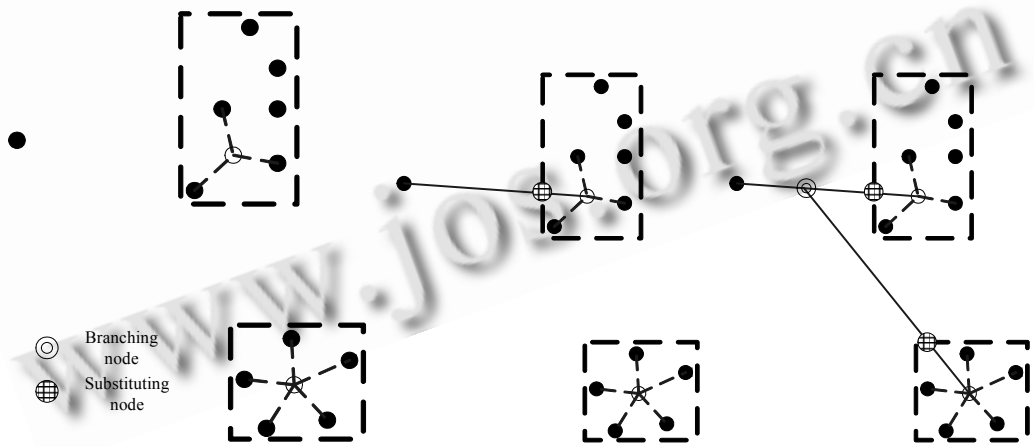


Fig.6 Edge merging between clusters
图 6 簇间网络边融合

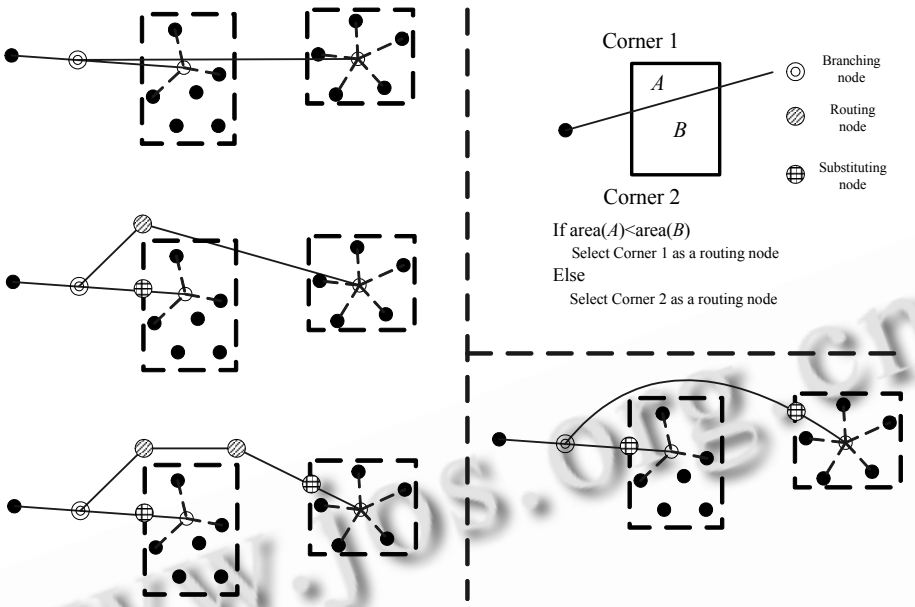


Fig.7 Edge routing between clusters

图 7 簇间网络边路由

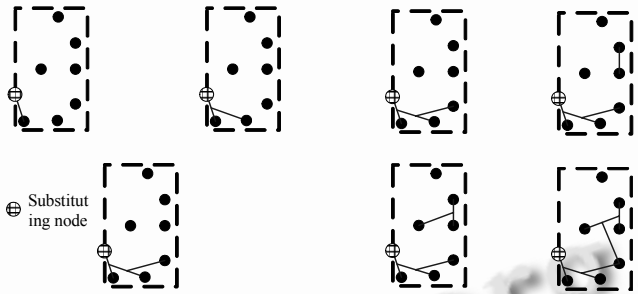


Fig.8 Edge merging inner cluster

图 8 簇内网络边融合

为了能够形式化描述 EMBC,ERBC 及 EMIC 这 3 部分算法,首先给出如下定义:

定义 4. 任意节点 V_i 都归属于一聚类簇 $cluster_k$,称与节点 V_i 相连的非 $cluster_k$ 内节点为 V_i 的簇间相关连点,V_i 的簇间相关连点与 V_i 相连边的条数定义为 V_i 的簇间度,记为 $d_{between}(V_i)$,而 $cluster_k$ 内的节点与 V_i 相连边的条数定义为 V_i 的簇内度,记为 $d_{inner}(V_i)$,则 $d_{between}(V_i)+d_{inner}(V_i)=d(V_i)$.

定义 5. 如图 6~图 8 所示,将两条拥有共同端点的边融合,其分叉点定义为该组边的分支点,对 $d_{between}(V_i)>0$ 的节点 V_i,在其簇间相关连点所在包围盒边缘定义 V_i 的同位点以代替 V_i 与各簇间相关连点连接.

首先给出 Elamer 算法的簇间网络边融合算法:

簇间网络边融合算法(EMBC).

(1) 计算各节点的簇间度、簇内度;

(2) 选取 $d_{between}(V_i)$最大的节点 V_i,将其簇间相关连点按所属聚类进行划分并计算其重心点;

(3) 以 V_i 为原点划分四象限,将全部重心点按所属象限分组,在每一象限内分别计算 V_i 与各重心点的距离,并分别记入数组 CG_{quad};

(4) 在每一象限内,以 V_i 为起点,以直线连接 CG_{quad} 中最小的点,交该点所属聚类 $cluster_j$ 的包围盒于点 U_{ij},

则 U_{ij} 为 V_i 在 $cluster_j$ 中的同位点,V_iU_{ij} 的中点为分支点;

(5) 删除每一象限 CG_{quad} 中的最小点,将 $cluster_j$ 中与 V_i 相关连点的簇间度减 1,同时更新 $d_{between}(V_i)$;

(6) 以分支点为起点,重复执行步骤(4)、步骤(5),当 $d_{between}(V_i)$=0 时终止;

(7) 以 V_i、分支点及同位点为控制点,使用 catmull-rom 样条曲线绘制连线;

(8) 重复执行步骤(2)~步骤(7),直到所有节点的簇间度都为 0 时终止.

当执行簇间网络边融合第(4)步时,若产生的连线会穿越其他聚簇,则执行簇间网络边路由过程,算法如下:

簇间网络边路由算法(ERBC).

(1) 依据连线穿越所划分面积选取包围盒的顶点作为路由点;

(2) 连接路由点和重心点,若仍然会穿越其他聚簇,则继续按照步骤(1)选取路由点,直到路由点可无障碍的连接重心点;

(3) 通过全部路由点,将分支点与重心点以 catmull-rom 样条曲线连接.

经过簇间网络边融合及路由过程,所有簇间网络边布局完成,簇间相关连点在相应的包围盒边缘都产生了对应的同位点.最后,Elamer 算法进行簇内网络边融合完成整个边布局过程.

簇内网络边融合算法(EMIC).

(1) 在每一簇内,首先给出所有同位点的簇内度;

(2) 选取 $d_{inner}(V_i)$最大的同位点 V_i,以 V_i 为原点划分四象限,将与其相关的点按所属象限分组,在每一象限内分别计算 V_i 与各相关点的距离,并分别记入数组 CG_{quad};

(3) 在每一象限内,以 V_i 为起点连接 CG_{quad} 中最小的点,计算连线中点作为分支点;

(4) 删除每一象限 CG_{quad} 中的最小点,将其簇内度减 1,同时,$d_{inner}(V_i)=d_{inner}(V_i)-1$;

(5) 以分支点为起点,重复执行步骤(3)、步骤(4),当 $d_{inner}(V_i)$=0 时终止;

(6) 重复执行步骤(2)~步骤(5),直到簇内所有同位点的簇内度都为 0 时终止;

(7) 使用相同方法计算簇内其余节点的全部分支点;

(8) 以同位点、节点及分支点为控制点,使用 catmull-rom 样条曲线将其进行连接.

为避免折线拐点给用户带来的视觉刺激,Elamer 算法参照文献[22]使用 catmull-rom 样条曲线代替直线对各端点进行连接;同时,若绘制的样条曲线会覆盖除端点外的其他节点,则 Elamer 算法会对该节点产生一个极小的随机扰动微调其位置,使其避开连线;对于簇内网络边融合过程中可能产生的共线问题,如图 9 所示,Elamer 算法则以分支点、分支点的相关点及其之间的节点为控制点绘制 Bézier 曲线绕开该节点(簇间网络边融合不会产生共线问题,若存在共线情况,则 Elamer 算法必定会调用网络边路由算法直接绕开聚簇).

Fig.9 Solution to co-linearity problem

图 9 共线问题的解决

2.3 人机交互设计

MulNetVisBasc 结合其节点、边的布局以及针对多变元网络数据可视分析任务的特点,设计了友好便捷的人机交互界面,以辅助用户进一步发掘分析该类数据集的结构特征及其中蕴含的隐性知识.

为能更加清晰地展现不同类别节点的分布及关联信息,MulNetVisBasc 按照节点所属分类,使用不同颜色对其进行标注;由于 ASC 方法将节点聚簇,簇内节点较为密集,因此设计了信息可视化工具通用的缩放、平移功能,以放大查看簇内节点的关联情况;为减少可视混乱,发掘潜在关键节点,集成连接度滑块根据节点度的大小对其进行动态过滤;高亮和导航功能可以辅助用户研究网络连接的拓扑结构,当用户将鼠标移动到某一节点时,与该节点相连的节点及边高亮显示,并且文字标注节点的特征信息(如名称等),用户鼠标沿连线运动至一相关

点,则与该点相关的节点及边加入上一操作形成的高亮视图(初始视图节点布局集中,文字标注会产生可视混乱,因此在交互筛选之后,才进行节点的文字标注).

3 实验结果

我们首先通过实验验证 MulNetVisBasc 方法的有效性,同时观察整个多变元网络数据集在改进星形坐标系中的多维分布及其之间的相互关联情况,获取较为丰富的隐性知识,并发掘不同种类相互关联节点的维分布特性;然后,通过对比 Elamer 算法作用前后的可视化结果,检验边布局的重要性及 Elamer 算法的正确性;最后,通过对不同规模数据集应用 Elamer 算法及 Doantam 提出的边布局算法比较二者性能.

3.1 实验设置

实验原型采用 Eclipse 基于 prefuse 开发包[23]编写实现,操作系统平台 Microsoft Windows XP,机器配置为:Intel P4 2.6,1GB,120GB 硬盘.实验数据采用真实的国际军火贸易数据集,每一个网络节点代表一个国家或地区,每个节点的衡量指标包括该国或地区的全球和平指数(GPI)、人均 GDP(GDP per capita)、人均军费(ME per capita)、部队人均军费(ME/AF)、每千人的军队数量(armed forces per 1000 people)、军费占 GDP 的比重(ME/GDP)、军费占中央财政的比重(ME/CGE),共 7 个变元属性,网络边表示相连两个国家在 2005 年间进行过军火贸易(此处暂未考虑二者在交易过程中的买卖角色,即该网络为无向网络).数据集来源于 http://www.prio.no/NISAT/Small-Arms-Trade-Database/与 http://first.sipri.org/.

3.2 可视化效果

图 10 是采用 MulNetVisBasc 方法对 2005 年国际军火贸易数据的可视化结果图,并根据各国家的发达程度进行了标注.其中,三角形表示发达国家,圆形代表发展中国家,维度轴标注端为数值增大方向.观察分析该图,我们可以直观获得如下结论:首先,GDP Per Capita,ME Per Capita,ME/AF,GPI 维相关度较高聚集在一起, Armed Forces Per 1000 People,ME/GDP,ME/CGE 维之间也存在一定相关度;其次,图中大体形成了 4 个聚类:左上方以发达国家为主的聚类,发达、发展中国家各占一部分的中间区域,右上方的发展中国家聚类和右下方的两个孤立点(通过交互操作可知是 Eritrea 和 North Korea);再次,与发展中国家相比较而言,发达国家人均 GDP、人均军费及部队人均军费都比较高,而全球和平指数较低(指数越低,说明该国或地区和平程度越高),发达国家每千人的军队数量、军费占 GDP 的比重、军费占中央财政的比重略低于发展中国家,但差别不大(除 Eritrea 和 North Korea 外);最后,有部分发展中国家未进入国际军火贸易网络,在图中表现为一些无边连接的节点;而发达国家基本都是国际军火贸易的主角,在图中表现为连接较多或拥有未分支网络边的节点.

经过简单地缩放、平移、高亮显示等一系列交互操作,我们分析了美国、俄罗斯和中国的军火贸易伙伴关系,如图 11~图 13 所示,对比 3 图可明显发现:美国的贸易伙伴遍及各类国家,但未与俄、中建立联系;俄罗斯的军火贸易主要针对中间区域的国家及 GPI 指数较高、GDP Per Capita,ME Per Capita,ME/AF 指数较低的国家;而中国的军火贸易却只局限于 GPI 指数较高、GDP Per Capita,ME Per Capita,ME/AF 指数较低的国家,并且与俄贸易伙伴的交集甚少.

为进一步证明 MulNetVisBasc 的优越性,我们使用 Wu 等人提出的 GeoSOM 方法[14]可视化同一国际军火贸易数据集,效果如图 14 所示.对比图 10 与图 14 易知:首先,MulNetVisBasc 可展现全部节点及其关联信息,而 GeoSOM 只能处理相互间有连接的节点,丢失了上文提到的未进入贸易网的部分发展中国家,因此存在信息缺失问题;第二,虽然 GeoSOM 的节点位置也是由多变元属性共同决定,但是由于没有 MulNetVisBasc 中类似高级星形坐标系的“参照物”,所以只能模糊描述节点的多变元分布情况.即使 GeoSOM 采用了节点背景颜色标注变元值,但这只限于一个变元的表达(如图 14 中颜色即表示军费占 GDP 的比重);第三,GeoSOM 未考虑边布局优化问题,图 14 中存在较多交叉及一定程度的可视混乱,给用户发掘数据集的隐性信息带来认知困难;而 MulNetVisBasc 中的相互关联较为清晰,有益于用户进一步的分析挖掘;第四,GeoSOM 的边带有贸易方向信息,同时,边颜色表示了贸易量,但是该信息在 MulNetVisBasc 却没有涉及,这是 MulNetVisBasc 的不足之处.总

之,MulNetVisBasc 与 GeoSOM 虽各有优劣,但是相比较而言,MulNetVisBasc 在同时直观揭示数据集多变元及网络特性(尤其是多变元分布信息)方面的优势较为明显,而且也不存在前文提到的多变元网络可视方法变元可视数量较少问题.

为验证边布局算法 Elamer 在 MulNetVisBasc 中的重要性,我们同时生成了未进行边优化的 MulNetVisBasc 效果图,如图 15 所示.由于节点排列较为密集,网络边较多,可视混乱程度较高.尤其是方框部分,根本无法分辨其中的节点连接关系,很多信息被覆盖在由大量交叉边形成的"密集网"之下.而图 10 在使用了 Elamer 算法之后,该问题得到了很大程度的改善,与图 15 相比,方框区域中节点间的连接关系变得较为清晰.但是,图 15 的可取之处在于对"密集网"边缘节点的连接度展现较为直观(图中圆圈中的节点度较高);而在图 10 中却只能根据边的起点分支走向信息判断该节点为高度节点.因此,我们正在考虑根据边权重信息以由深及浅颜色渐变方式渲染网络边,以弥补 Elamer 算法因边融合及路由导致节点度展现不直观的缺陷.

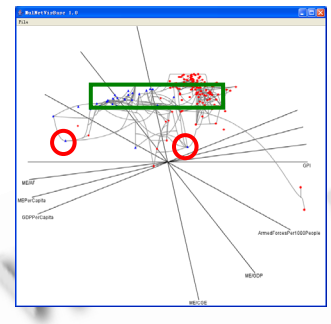


Fig.10 Visualization of international arms transfers

图 10 国际军火贸易数据可视化效果图

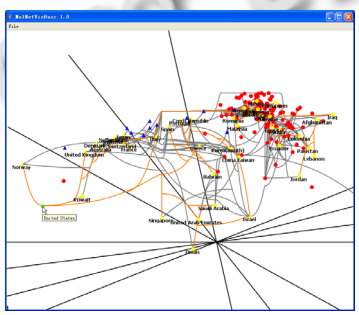


Fig.11 Highlight arms trade partnership of USA

图 11 高亮显示美国军火贸易伙伴

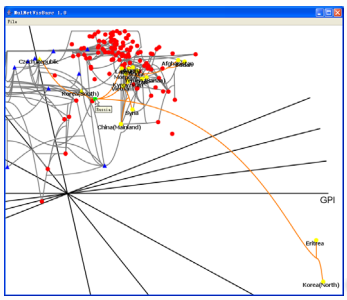


Fig.12 Highlight arms trade partnership of Russia

图 12 高亮显示俄罗斯军火贸易伙伴

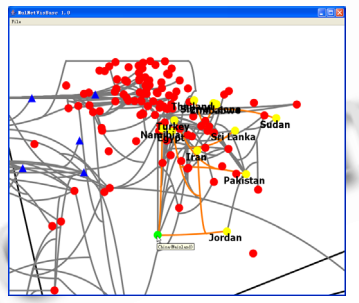


Fig.13 Highlight arms trade partnership of China

图 13 高亮显示中国军火贸易伙伴

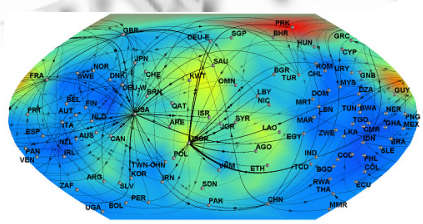


Fig.14 Visualization of international arms transfers using GeoSOM

图 14 国际军贸数据集的 GeoSOM 可视化效果图

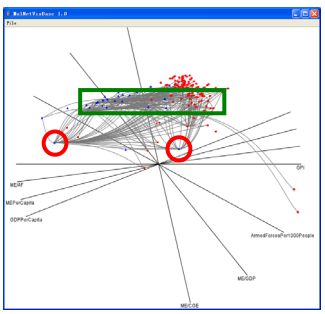


Fig.15 Visualization of international arms transfers without Elamer

图 15 未使用 Elamer 算法的国际军贸数据可视化效果

3.3 性能分析比较

下面分析 Elamer 算法与 Doantam 所使用的边布局算法的理论时间复杂度.Doantam's 算法包括网络节点布局调整、层次聚类、"根化"过程、边融合、边路由及渲染过程.首先,我们认为,其网络节点布局调整过程(时间复杂度 $O(n^2)$)与本文使用 ASC 方法布局网络节点步骤对应(时间复杂度接近 $O(n\lg n)$),都不应属于狭义的边布局过程;其次,二者又都需要聚类前计算任意两节点间距离(时间复杂度 $O(n^2)$)、聚类后计算最小包围盒及最后进行可视渲染,并且二者在上述过程的时间复杂度差异不大.因此,对比分析中我们只考虑了 Doantam's 算法中的层次聚类、"根化"、边融合及路由过程,Elamer 算法中的 k-means 聚类、簇间边融合及路由和簇内边融合过程.Doantam's 算法边融合及路由的时间复杂度为 $O(n)$,传统层次聚类方法的复杂度为 $O(n^2)$,"根化"过程为 $O(n^2)$(优化后能够达到 $O(\lg n)^2$)[22],因此,Doantam's 算法的总时间复杂度为 $O(n^2)$(只是针对单源情况,多源时复杂度会更高);而 k-means 算法的时间复杂度为 $O(n)$,Elamer 算法簇间边融合及路由过程中计算节点簇间(内)度步骤的最差时间复杂度为 $O(nl)$,平均为 $O(l\lg n)$(l 为边数量),根据节点度排序的最差时间复杂度为 $O(n^2)$,平均为 $O(n\lg n)$,计算分支、同位点的时间复杂度为 $O(n)$,簇内网络边融合过程中计算同位点与各相关节点距离的复杂度为 $O(kn)$,根据距离排序节点的最差时间复杂度为 $O(n^2)$,平均为 $O(n\lg n)$,计算分支节点的复杂度为 $O(n)$,因此,Elamer 算法总的时间复杂度最差为 $O(n^2)$或 $O(nl)$,平均为 $O(n\lg n)$或 $O(l\lg n)$.相比较而言,一般情况下,Elamer 算法的性能在理论上优于 Doantam's 算法.

为使用真实数据进一步证明 Elamer 算法性能的优越性,我们对比分析了 Elamer 算法与 Doantam's 算法对不同规模多变元网络数据集的运行时间效率.除国际军火贸易数据集(190 节点、312 条边)外,我们还收集了 Padgett's Florentine families network(16 节点、38 条边)[24]和 Krackhardt's high-tech managers advice network(21 节点、145 条边)[24]两个小型多变元网络,以及大型多变元网络 Jeong 等发布的 Biological network(1 870 节点、2 240 条边)[25],并使用 GEOMI[26]人工合成一个中等规模的多变元网络(516 节点、773 条边),结果如图 16 所示.

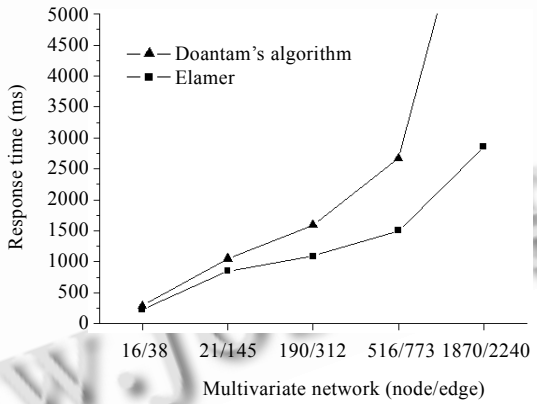


Fig.16 Performance comparison of Elamer and Doantam's algorithm

图 16 Elamer 算法与 Doantam's 算法的性能比较

从图中我们发现,两种算法对于小型网络运行效率较好,对中型网络的运行时间尚处于用户能够承受的范围.但随着网络规模的逐渐增大,Doantam's 算法所需时间大幅度上升,Elamer 算法的效率优势逐渐显现,其性能与网络整体规模之间大体呈现线性关系.我们还注意到,相比同等节点的一般多变元网络图,对于 high-tech managers advice network 这类近"完全图"(每两节点间必有边连接),两种算法的时间效率大幅下降,说明多变元网络边数量对他们性能的影响都比较大.这是因为网络边数量大幅增加,会促使 Doantam's 算法出现更多"源点",也会提高 Elamer 算法计算节点簇间(内)度的复杂度,这与本文对二者时间复杂度的理论分析是一致的.

4 结论与进一步工作

本文针对现有多变元网络可视化方法的局限性,提出了一种新的多变元网络可视化方法 MulNetVisBasc.根据节点的多变元属性,使用 ASC 方法布局网络节点,以边融合及路由技术为基础,设计了 Elamer 算法自动有效布局网络边,实现了友好的人机交互界面,以辅助用户进一步对数据进行分析挖掘.实验结果表明,MulNetVisBasc 的可视化结果能够在直观揭示数据集多变元分布特性的同时,清晰展现其网络关联特性,有助于发掘多变元网络数据集中潜在的隐性知识,人机交互界面灵活方便,用户满意度较高;Elamer 算法能够大幅度减少因边交叉引起的可视混乱,并且与 Doantam 所使用的边布局算法相比时间复杂度较低,适用于数据量较大的数据集合.

对于 MulNetVisBasc 方法,以后还需要在网络边多变元特性可视化、Elamer 算法的优化(尤其是聚簇包围盒的生成方式及边融合过程中区域划分问题)及进一步完善人机交互界面的等方面进行研究.

致谢 非常感谢封孝生副教授在 Elamer 算法设计与实现过程中给予的指导和帮助,以及 Jeffrey Heer 副教授在 prefuse 开发包使用过程中给予的帮助.

References:

[1] Hoffman PE. Table Visualizations: A Formal Model and its Applications. Lowell: University of Massachusetts, 1999.

[2] Inselberg A. The plane with parallel coordinates. The Visual Computer, 1985,1(2):69−91. [doi: 10.1007/BF01898350]

[3] Kandogan E. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In: Proc. of the 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2001. 107−116.

[4] James XZL. Visualization of high-dimensional data with relational perspective map. Information Visualization, 2004,3:49−59. [doi: 10.1057/palgrave.ivs.9500051]

[5] Shao C, Huang HK, Zhao LW. A more topologically stable ISOMAP algorithm. Journal of Software, 2007,18(4):869−877 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/869.htm [doi: 10/1360/jos180869]

[6] Kozo S, Shojiro T, Mitsuhiko T. Methods for visual understanding of hierarchical system structures. IEEE Trans. on Systems, Man and Cybernetics, 1981,11(2):109−125. [doi: 10.1109/TSMC.1981.4308636]

[7] Peter E. A heuristic for graph drawing. Congressus Numerantium, 1984,42:149−160.

[8] Herman I, Melancon G, Marshall MS. Graph visualization and navigation in information visualization: A survey. IEEE Trans. on Visualization and Computer Graphics, 2000,6(1):24−43. [doi: 10.1109/2945.841119]

[9] Wong PC, Bergeron RD. 30 years of multidimensional multivariate visualization. In: Proc. of the Scientific Visualization, Overviews, Methodologies, and Techniques. Washington: IEEE Computer Society, 1997. 3−33.

[10] Becker RA, Eick SG, Wilks AR. Visualizing network data. IEEE Trans. on Visualization and Computer Graphics, 1995,1:16−28. [doi: 10.1109/2945.468391]

[11] Stephen GE, Graham JW. Navigating large networks with hierarchies. In: Proc. of the 4th Conf. on Visualization'93. Washington: IEEE Computer Society, 1993. 204−209.

[12] Xu K, Cunningham A, Hong SH, Thomas BH. GraphScape: Integrated multivariate network visualization. In: Proc. of the Asia-Pacific Symp. on Visualization. Los Alamitos: IEEE Computer Society, 2007. 33−40.

[13] Wu YX, Takatsuka M. Visualizing multivariate network on the surface of a sphere. In: Proc. of the Asia-Pacific Symp. on Information Visualization. Los Alamitos: IEEE Computer Society, 2006. 77−83.

[14] Wu YX, Takatsuka M. Visualizing multivariate networks: A hybrid approach. In: Proc. of the IEEE Pacific Visualization Symp. Los Alamitos: IEEE Computer Society, 2008. 223−230.

[15] Sun Y, Tang JY, Tang DQ, Xiao WD. An improved multivariate data visualization method. Journal of Software, 2010,21(6): 1462−1472 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/21/1462.htm [doi: 10.3724/SP.J.1001.2010.01462]

[16] Card S, Mackinlay J, Shneiderman B. Readings in Information Visualization: Using Vision to Think. San Fransisco: Morgan Kaufmann Publishers, 1999.

[17] Sindre G, Gulla B, Jokstad H. Onion graphs: Aesthetic and layout. In: Proc. of the IEEE Symp. on Visual Languages. Washington: IEEE Computer Society, 1993. 287−291.

[18] Purchase H. Which aesthetic has the greatest effect on human understanding? In: Proc. of the 5th Int'l Symp. on Graph Drawing. Springer-Verlag, 1997. 248−261.

[19] Garey MR, Johnson DS. Crossing number is NP-complete. SIAM Journal on Algebraic and Discrete Methods, 1983,4(3):312−316. [doi: 10.1137/0604033]

[20] Hussien B, Sridhar B. A robust line extraction and matching algorithm. SPIE Intelligent Robots and Computer Vision XII, 1993, 2055:369−380.

[21] David PD, Emden RG, Eleftherios K, North SC. Implementing a general-purpose edge router. In: Proc. of the 5th Int'l Symp. on Graph Drawing. Springer-Verlag, 1997. 262−271.

[22] Doantam P, Ling X, Ron Y, Pat H, Terry W. Flow map layout. In: Proc. of the IEEE Symp. on Information Visualization. Washington: IEEE Computer Society, 2005. 219−224.

[23] Heer J, Card SK, Landay JA. Prefuse: A toolkit for interactive information visualization. In: Proc. of the CHI 2005. New York: ACM Press, 2005. 421−430.

[24] Wasserman S, Faust K. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.

[25] Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. Nature, 2001,411:41. [doi: 10.1038/35075138]

[26] Ahmed A, Dwyer T, Forster M, Fu XY, Ho J, Hong SH, Koschutzki D, Murray C, Nikolov NS, Taib R, Tarassov A, Xu K. GEOMI: Geometry for maximum insight. In: Proc. of the 13th Int'l Symp. on Graph Drawing. Springer-Verlag, 2005. 468−479.

附中文参考文献:

[5] 邵超,黄厚宽,赵连伟.一种更具拓扑稳定性的 ISOMAP 算法.软件学报,2007,18(4):869−877. http://www.jos.org.cn/1000-9825/18/869.htm [doi: 10/1360/jos180869]

[15] 孙扬,唐九阳,汤大权,肖卫东.改进的多变元数据可视化方法.软件学报,2010,21(6):1462−1472. http://www.jos.org.cn/1000-9825/21/1462.htm [doi: 10.3724/SP.J.1001.2010.01462]

孙扬(1983－),男,山东济南人,博士生,主要研究领域为信息可视化,人机交互技术.

汤大权(1971－),男,博士,教授,主要研究领域为信息资源管理,信息可视化.

赵翔(1986－),男,博士生,主要研究领域为网络数据管理.

肖卫东(1968－),男,博士,教授,博士生导师,主要研究领域为信息管理,信息可视化.

唐九阳(1978－),男,博士,副教授,主要研究领域为对等网,知识管理.