Journal of Computer Applications
计算机应用,2013,33(10):2981-2983,2988
ISSN 1001-9081
CODEN JYIIDU
2013-10-01
http://www.joca.cn

文章编号:1001-9081(2013)10-2981-03
doi:10.11772/j.issn.1001-9081.2013.10.2981

基于聚类的空间数据可视化方法

张 洋*,王 辰
(国防科学技术大学 信息系统与管理学院,长沙 410073)
(*通信作者电子邮箱 robin.zhangyy@gmail.com)

摘 要:首先介绍了目前空间数据可视化技术的研究内容和基本方法,对基于实体和基于区域两类常用方法进行了分析和总结。在此基础上提出了一种基于聚类的空间数据可视化方法,其基本思想是利用以Delaunay三角网的自适应空间聚类算法(ASCDT)为代表的空间聚类算法进行聚类分析,并获得结果描述参数,结合基本方法和参数特征设计专门用于聚类结果表达的可视化对象,进而实现空间数据的图上投影。最后对该类方法有待进一步探讨和改进的内容进行了展望。

关键词:空间数据;空间聚类;Delaunay三角网的自适应空间聚类算法;空间数据可视化
中图分类号:TP311.131 **文献标志码**:A

Spatial data visualization based on cluster analysis

ZHANG Yang*, WANG Chen
(*College of Information Systems and Management, National University of Defense Technology, Changsha Hunan* 410073, *China*)

Abstract: Firstly, the paper introduced the researches and basic methods of spatial data visualization technology, and analyzed two common kinds of methods, namely entity-based and region-based. A clustering-based spatial data visualization method was proposed, which firstly made a cluster analysis of spatial data and got the description parameters of the result through the use of spatial clustering algorithms represented by algorithm ASCDT (Adaptive Spatial Clustering algorithm based on Delaunay Triangulation). Secondly, it designed visual objects aimed at the cluster result by combining the basic visualization methods and the characteristics of the parameters. As a result, the mapping relationship was established. Finally, some issues that needed to be further studied and improved were discussed.

Key words: spatial data; spatial cluster; Adaptive Spatial Clustering algorithm based on Delaunay Triangulation (ASCDT); spatial data visualization

0 引言

随着信息技术的高速发展,空间数据呈现出海量式的增长,而人们对数据的分析与处理要求不断提高,“数据爆炸,知识匮乏”的问题逐渐凸显。如何建立一种有效的空间数据表达方法,为相关规划与管理部门提供决策支持工具,解决数据与决策者之间的鸿沟问题,即用户如何从海量的空间数据中获取有用信息,成为这一领域研究的热点和难点。可视化技术作为一种重要的数据挖掘手段,通过对空间信息的有效组织和直观展现,在解决海量数据的知识发现(Knowledge Discovery in Database,KDD)问题中发挥了重要的作用。

空间数据最主要的特征体现在其依赖于地理位置(如经度、维度等)分布。目前,空间数据可视化最常用的解决方法是采用基于地理信息系统(Geographic Information System,GIS)的方法。但是,随着相关领域空间数据规模和复杂程度的不断提高,单纯依靠地理信息系统的方法存在明显的缺陷——重“数据对象显示”,轻“信息结构刻画”[1]。也就是说,基于地理信息系统的方法对数据中潜在的关系、规律、趋势和模式缺乏有效的表达,其结果通常很难直接作为决策者辅助决策的依据。

对于相对庞大的数据集而言,我们首先关注的往往是空间数据的宏观分布特征,即不直接以单个空间实体作为研究对象,而是通过一些统计性的手段对数据进行预处理,获取数据分布的基本模式和规律,然后通过合适的可视化手段进行表达,进而提高数据挖掘的效率。针对目前空间数据可视化技术在实际应用中的要求,结合现有方法存在的优缺点,本文提出了一种基于聚类的空间数据可视化方法。其基本思想是:通过聚类分析按照一定规则对空间数据进行有效的组织和规范;针对空间聚类结果,设计一种表达直观、易于学习、交互良好的可视化方法对其进行描述和展现,并且为用户提供必要的交互接口,进而提高空间数据挖掘进程的效率和结果的可信度。

1 空间数据可视化基本方法分析

地理信息系统作为相关信息系统设计与开发的技术支撑,可以将空间实体按照地理位置进行图上投影,并结合常用的可视化方法,实现专题属性的显示、查询和统计。空间数据可视化的基本方法包括基于圆点图标的方法、基于统计图的方法、基于密度图的方法和基于热区图的方法等。

基于圆点图标的方法用圆点表示空间对象,并利用空间对象的空间维信息将其投影到地图上,常用的投影方法有最近邻接点算法和基于曲线的算法。同时,可以利用圆点的尺寸、颜色和透明度等特征来表示空间对象的专题属性。但是,若数据量较大、地图范围有限,圆点图标会出现拥挤、重叠的

收稿日期:2013-04-22;**修回日期**:2013-06-04。

作者简介:张洋(1989-),男,重庆人,硕士研究生,主要研究方向:多媒体、虚拟现实; 王辰(1973-),男,天津人,副教授,主要研究方向:多媒体信息系统、人机交互、指挥决策。

现象,不能满足应用需要。利用栅格划分算法[2]能够在一定程度上缓解重叠问题带来的不良影响。

基于统计图的方法,首先对关注的地图范围按照一定规则进行区域划分,针对每个区域进行统计分析,获得描述该区域专题属性的主要统计特征参数,最后采用适合的统计图显示。该方法具备一般统计图的优势,例如,方便对数据进行精确度量和对比,专题维维度数量限制小等。常用的统计图有饼状图、柱状图和雷达图。但是,由于统计图基本显示尺寸的要求,其出现的重叠问题可能更加突出。

基于密度图的方法是应用最为广泛的数据可视化方法之一,尤其是在人口密度方面。该方法的主要思想是用颜色对空间对象在地图上分布的密度值进行编码,不同颜色对应不同密度值或者密度值范围,通常先定义密度最大值和最小值分别对应的颜色,两者之间在视觉上要有明显差异且符合人们的认识习惯,如用红色表示最大值,用蓝色表示最小值。

基于热区图的方法用事先定义好的标准圆形分别描绘各空间实体对象,圆心位置与实体坐标位置重合,所有圆形半径相等,并根据用户关注的具体数量值对圆形进行径向渐变色彩的填充,进而表达空间实体的聚集性特征。基于热区图的方法利用形状、色彩以及透明度传递空间数据信息,对非精确信息表达合理,容易接受和理解;采用渐变平滑过渡的显示方法,视觉上更加柔和。

综上所述,常用的空间数据可视化方法各有优缺点。其中,基于圆点图标的方法和基于热区图的方法是从空间实体个体本身出发,对于数据量较大的情况,一方面,可能出现显示的"拥挤"和"重叠";另一方面,也增加了渲染的成本。基于统计图的方法和基于密度图的方法,则是在对原始空间数据进行统计预处理的基础上实现的,其研究对象是"某一区域",这样能够在一定程度上满足大数据量的可视化需求,但它们都侧重于空间实体在专题维度的统计信息,对实体本身在区域内的分布信息缺乏表达。

2 基于ASCDT的空间聚类分析

根据地理学第一定律可以得到结论:空间上距离近的实体间的相似性比距离远的实体的相似性大,即空间实体间的依赖关系。因此,通过空间聚类分析可以将空间实体依据一定的相似性度量标准划分成若干具有一定意义的空间簇,簇内实体尽可能相似,簇间实体尽可能相异。与传统聚类相比,其特殊性主要表现在实体的定义、相似性的定义以及类的定义三个方面。解决空间聚类相关问题通常要与地理信息相结合,重点考虑空间属性的关联性,即实体在空间位置上的直接或间接邻近关系。常用的空间聚类方法主要有基于划分的算法、基于层次的算法、基于密度的算法和基于图论的算法等。

2.1 ASCDT空间聚类算法

对于任意的平面点集,经过剖分得到的Delaunay三角网是唯一的,并且具有空外接圆和最大最小角等整体最优性质,因此得到了广泛的应用。Deng等[3]提出了一种基于Delaunay三角网的自适应空间聚类算法(Adaptive Spatial Clustering algorithm based on Delaunay Triangulation, ASCDT),其基本思想是首先通过建立Delaunay三角网来表达空间实体间的空间邻近关系,利用网边的基本统计量来定义整体和局部的约束准则,删除其中的不一致边(包括整体长边和局部长边以及局部"颈"或"链")。针对障碍和区域限定的需求,有顾及空间障碍的自适应空间聚类算法——ASCDT+,采用障碍图层与Delaunay三角网的边进行叠置分析,并打断与障碍物相交的边,以此来考虑空间障碍对实体间可达性的阻隔,同时可将区域划分等价为闭合的空间障碍。与目前常用的空间聚类算法相比,其优势主要体现在对不同形状空间簇的识别、对实体不同密度分布的适应和对颈链问题的处理上。具体到本文需要解决的可视化问题,尤其表现为其采用物理学中凝聚力和凝聚场的思想,为聚类结果的可视化表达提供了必不可少的基础。利用ASCDT进行空间聚类分析的基本流程见图1。

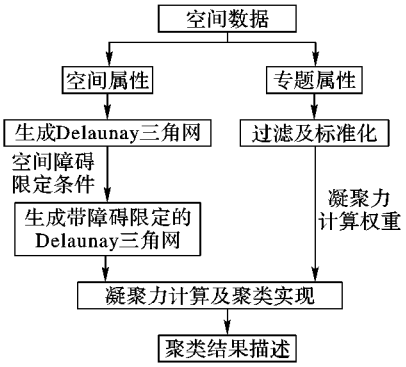


图1 利用ASCDT进行空间聚类分析的基本流程

2.2 ASCDT在空间数据可视化应用中的可行性分析

针对空间数据特点及其可视化实现需求,对ASCDT在空间数据可视化应用中的可行性作如下分析:

1)算法复杂度。ASCDT的整体复杂度约为$O(N\log N)$,主要体现在构建Delaunay三角网的过程,其他步骤的时间复杂度近似线性为$O(N)$。在实际操作中,ASCDT+算法只是引入了多图层操作的步骤,其复杂度也是线性的,所以整体复杂度并没有改变,仍然能够适应海量数据的应用需求。

2)聚类结果的易描述性。空间聚类结果的描述问题主要体现在空间簇的聚集程度上,常用的方法都是通过空间实体对应的Voronoi邻近实体集合计算得到的,而Delaunay三角网是Voronoi图的对偶图,所以实际操作中只需要根据Delaunay三角网定义的空间邻近关系来计算凝聚力,避免了计算的复杂性[7]。Delaunay三角网中任意两个空间实体P和Q间的凝聚力可定义为

$$F_{agg}(P,Q)=\begin{cases}k\dfrac{m_P m_Q}{d(P,Q)^2}, & Q\in NV(P)\\ 0, & Q\notin NV(P)\end{cases}$$

其中:k为引力系数,可设为1;m_P和m_Q分为P和Q的质量,可视为空间实体专题维度的权重;$d(P,Q)$为实体P和Q间的欧氏距离,$NV(P)$表示实体P的直接邻近实体集合。

因此,一方面,可以通过设定凝聚力计算权值来反映专题属性(即用户关注的空间实体的非空间属性)对聚类结果的影响;另一方面,增强了聚类结果的可解释性和可用性,获得的空间簇可根据其特征参数进行描述和表示。

综上所述,ASCDT能够有效地结合空间实体对象的空间维属性和专题维属性进行聚类分析,通过三角网的建立约束空间位置,通过权重设定考虑用户感兴趣的专题维内容,并且获得的聚类结果易于可视化表达。

3 聚类结果可视化设计

3.1 可视化对象设计

可视化对象参考常用的圆点图标来表达,圆点图标的圆心、半径分别与空间簇的质心、范围半径相对应。用径向渐变色彩来表达每个空间簇的整体和局部聚集性特征,簇内实体数量通过渐变色彩范围来反映,实体数量越大,渐变色彩最大值越大,即圆点图标圆心处颜色值越大。具体实现中,根据不

同的颜色空间模型(如 RGB、HSL、HSV 等)和使用的颜色分量通道,"值"的定义也有所区别,但都必须符合可视化的基本原则和人们的日常认识习惯。

簇内紧密度通过渐变色彩中间控制滑块位置来表达(可以参考常用的图形图像处理软件中的渐变工具,如图 2),根据实际需要,滑块的数量可以是单个或多个,只要不引起视觉上的分辨压力。以单个滑块为例,当簇内实体分布紧密时,滑块位置越趋向于圆点图标的圆心位置,色彩越集中,对比越强烈;当簇内实体分布分散时,滑块位置越趋向于圆点图标的边缘位置,色彩越均匀,过渡越平滑。如图 3。依据上述方法对各个空间簇进行可视化表达之后,有必要考虑簇间的位置关系,当空间簇对应的渐变圆点图标之间存在"拥挤"和"重叠"现象时,采用基于热区图的方法进行颜色和透明度的渐变叠加,实现平滑柔和过渡的效果,以符合一定的可视化美学标准。

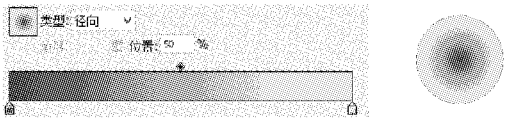
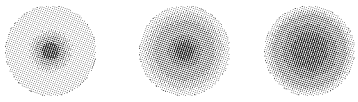


图 2　矢量图形处理软件 Adobe Illustrator 中的渐变工具



(a) 15%　(b) 50%　(c) 75%

图 3　渐变滑块位置分别为 15%、50%、75% 时的图标举例

3.2　聚类结果到可视化对象的映射关系

在设计可视化对象的基础上,建立聚类结果到可视化对象的映射关系,见图 4。

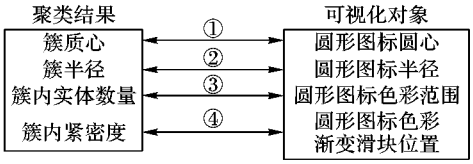


图 4　聚类结果到可视化对象的映射

其中,①~④映射通常都是包含一定变换的间接映射。映射①需要从空间坐标投影到用于实际显示的屏幕坐标,常用的投影方法类似于基于圆点图标的常用方法。映射②必须结合实际显示的地图范围来设定图标尺寸,可用面积或者半径来度量:尺寸过大,会出现"重叠"现象;尺寸过小,空间簇之间区别不明显。映射③要求在区别不同实体数量空间簇的同时,保证可视化的美学标准,尤其是对于同一类型的数据对象要保证规范统一。映射④可以以簇内紧密度的最大值为基准进行归一化,按不同的比例决定渐变色彩滑块位置。

虽然聚类结果的描述参数能够在一定程度上反映簇内实体的分布特征,但这种表示是模糊而不准确的。例如,由于采用圆形图标和径向渐变的可视化对象,簇内实体分布的方向性不明确,如果空间实体分布在相对于簇质心的不同方向,但其紧密程度相似,其可视化效果是没有差异的。为了满足用户对局部态势信息的需求,可以采用基于层次细节(Levels of Detail,LOD)的交互式可视化方法来实现宏微观态势的切换过渡。例如,设计实现从空间簇对应的圆形图标对象,到簇内空间对象对应的拓扑结构之间的平滑转换方法。

4　应用分析

以公路交通运输企业运力数据可视化为例,当获得关于车辆运输原始数据(即所有客/货运输记录)后,通过数据库操作过滤出相关字段,利用统计计算方法,得到表现其运力的主要指标,包括货物运送量(吨/年)、人员运送量(人/年)、计划完成率(%)等。然后,采用 ASCDT 对运输企业进行聚类分析,利用企业的经纬度原始定位,将用户感兴趣的具体运力指标作为凝聚力计算权重。在获得聚类结果之后,将包含多个运输企业的空间簇作为可视化对象,充分利用空间簇的描述参数,设计图形化表示方法,建立图上映射。最后,为用户提供必要的交互支持,包括运力指标选择、运输范围限定、宏微观态势切换等。其中,公路交通运输企业货物运送量数据可视化结果如图 5 所示,渐变色彩的取值反映了对应区域运输企业的年货运量,由于对重叠图标进行了平滑过渡处理,实际显示结果呈现出不规则形态。

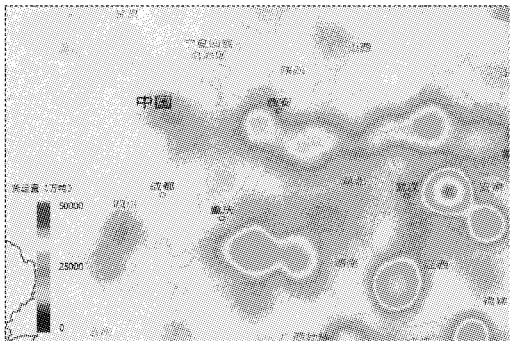


图 5　公路交通运输企业货物运送量数据可视化结果图例

5　结语

本文在分析现有空间数据可视化方法优缺点的基础上,结合实际需求,提出了一种基于空间聚类的方法,将数据聚类组织,对聚类结果进行描述和显示。本文方法在以下方面有待进一步深入研究:

1)聚类结果的不规则形状表达。对空间数据而言,其分布情况具有多样性,可以呈现圆形、带状,甚至是不规则的形状。因此,对于聚类结果的可视化问题,需要在反映空间实体实际分布范围的同时,保持可视化对象的规范性和可读性,即考虑是否对空间实体分布范围进行适当的抽象,以及如何建立此类情况下的映射关系。

2)可视化动态支持。当空间数据随时间变化(包括空间实体位置变化、专题维度属性值变化等)时,需要实时表达其主要特性。相比于静态可视化,时变可视化的主要目的是对空间数据动态演化过程的表达,通常分为离线可视化和在线可视化两类,主要区别在于时间序列中帧信息的来源。

参考文献:

[1]　樊明辉. 空间数据挖掘及其可视化系统若干关键技术研究[D]. 北京:中国科学院遥感应用研究所, 2006.

[2]　冯玉才, 刘嘉. 大量空间数据可视化的算法[J]. 计算机工程, 2003,29(13): 79-81.

[3]　DENG M, LIU Q, CHENG T, *et al*. An adaptive spatial clustering algorithm based on Delaunay triangulation[J]. Computers, Environment and Urban System, 2011, 35(4):320-332.

[4]　ANDRIENKO G, ANDRIENKO N, DYKES J, *et al*. Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research[J]. Information Visualization, 2008, 7(3): 173-180.

[5]　GONSCHOREK J, TYRALLOVÁ L. Geovisualization and geostatistics: a concept for the numerical and visual analysis of geographic mass data[M]. Berlin: Springer, 2012: 208-219.

(下转第 2988 页)

名称、零件编号,其中供应商名称下拉列表的显示内容依赖于片区名称。图 6 所示的是核心企业为供应商配置的代管库存查询表单,根据查看权限的配置,该企业用户只能查看自己的代管库存情况,因此供应商名称文本框的内容只读不可修改。

图 4 表单配置界面

图 5 核心企业代管库存查询表单

图 6 协作企业代管库存查询表单

5 结语

本文针对产业链协同 SaaS 平台的业务需求动态变化的现状进行分析,在传统的表单配置技术基础上,进一步对面向产业链协同平台的表单应用进行研究,通过将表单结构及表单元素与 XML 文档进行模式映射,支持表单配置模型的存储及动态加载,并实现平台表单在线动态更新。表单动态配置技术在产业链协同 SaaS 平台的应用使平台能基于 Internet 为企业提供灵活地服务,支持核心企业对协作企业群表单的动态配置,满足企业不断变化的需求,减轻维护人员的工作量,减少开发和维护成本,增强系统的复用性和生命力。

参考文献:

[1] 史玉良,栾帅,李庆忠,等. 基于 TLA 的 SaaS 业务流程定制及验证机制研究[J]. 计算机学报,2010,33(11):2054 - 2067.

[2] SHI Y L, LUAN S, WANG H Y. A flexible business process customization framework for SaaS[C]// Proceedings of the 2009 WASE International Conference on Information Engineering. Piscataway: IEEE, 2009:350 - 353.

[3] 李晓娜,李庆忠,孔兰菊,等. 基于共享模式的 SaaS 多租户数据划分机制研究[J]. 通信学报,2012,33(Z1):111 - 120.

[4] 唐文忠,莫伟栋. 面向领域的模型驱动智能表单系统的框架设计[J]. 北京航空航天大学学报,2007,33(9):1086 - 1089,1026.

[5] XU Y. The research of Workflow dynamic forms based on XML [C]// Proceedings of the 2011 IEEE International Conference on Computer Science and Automation Engineering. Piscataway: IEEE, 2011:331 - 334.

[6] 唐文忠,莫伟栋. 基于共享模型的工作流表单系统设计[J]. 北京航空航天大学学报,2008,34(4):391 - 395.

[7] 王瑞霞,隋宏伟,刘弘. 基于 XML 的表单设计器构件的设计与实现[J]. 计算机应用研究,2007,24(7):183 - 185.

[8] 李厚福,韩艳波,虎嵩林,等. 一种面向服务、事件驱动的企业应用动态联盟构造方法[J]. 计算机学报,2005,28(4):739 - 749.

[9] 王淑营. 面向产业链协同商务平台的动态数据交换技术研究[J]. 计算机集成制造系统: CIMS,2010,16(6):1336 - 1345.

[10] 王淑营. 面向制造业产业链协同商务平台集成框架[J]. 西南交通大学学报,2008,43(5):643 - 647.

[11] NAM C K, JANG G S, BAE J H. An XML-based active document for intelligent Web applications[J]. Expert Systems with Applications,2003,25(2):165 - 176.

[12] ZHANG K, ZHANG X, SUN W, *et al*. A policy-driven approach for software-as-services customization[C]// Proceedings of the 4th IEEE International Conference on Enterprise Computing, E-Commerce and E-Services. Piscataway: IEEE, 2007:123 - 130.

[13] LEE J H, KANG S J, HUR S J. Web-based development framework for customizing Java-based business logic of SaaS application [C]// Proceedings of the 14th International Conference on Advanced Communication Technology. Piscataway: IEEE, 2012:1310 - 1313.

[14] CHEN X J. Extending RMI to support dynamic reconfiguration of distributed systems[C]// Proceeding of the 22nd International Conference Distributed Computing Systems. Piscataway: IEEE, 2002:401 - 408.

[15] 王德俊,黄林鹏, 徐小辉. 事务控制的面向服务系统的动态更新协调[J]. 软件学报,2011,22(11):2652 - 2667.

(上接第 2983 页)

[6] GORRICHA J, LOBO V. Improvements on the visualization of clusters in geo-referenced data using self-organizing maps[J]. Computers & Geosciences, 2012, 43: 177 - 186.

[7] 刘启亮. 自适应空间聚类方法研究[D]. 长沙: 中南大学, 2011.

[8] 黄志敏. 带约束条件的交互式空间聚类算法研究[D]. 北京: 中国农业大学, 2007.

[9] 任永功, 于戈. 数据可视化技术的研究与发展[J]. 计算机科学, 2004, 31(12): 92 - 95.

[10] AOIDH E M, MARTINSOHN J. Geovisualization challenges of seascape genetics [EB/OL]. [2013- 01- 20]. http://www.researchgate. net/publication/228414861 _ Geovisualization _ Challenges_of_Seascape_Genetics.

[11] MACEACHREN A M, GAHEGAN M, PIKE W, *et al*. Geovisualization for knowledge construction and decision support [J]. IEEE Computer Graphics and Applications, 2004, 24(1): 13 - 17.

[12] WOLFF M, GONSCHOREK J. Geostatistical approaches for geovisual data exploration, analysis and 3D-visualisation in civil security[EB/OL]. [2013- 01- 20]. http://www. geomatik-hamburg. de/geoviz11/abstracts/12 _ GEOVIZ _ MWolff-JGonschorek_Potsdam. pdf.