基于空间聚类挖掘的城市应急救援机构选址研究①

樊 博
(上海交通大学国际与公共事务学院，上海 200030)

摘要：以城市突发事件的应急管理为应用背景，研究使用空间聚类技术解决应急服务机构的选址问题. 首先提出实施聚类分析的数据模型，然后改进既有空间聚类算法的基础上，提出了以 k-means 聚类算法缩减解空间的搜索范围、以模拟退火算法为解搜索策略. 基于 GIS 系统，充分考虑空间障碍物因素和空间环境因素等限制条件，提出(COD-MEANS- CLASA)的空间聚类算法以实现应急救援机构的科学选址. 实验结果表明，提出的算法在执行效率和解质量两方面具有更好的表现.

关键词：空间聚类；k-means 算法；模拟退火算法；应急救援机构选址
中图分类号：C931.6；L986 **文献标识码**：A **文章编号**：1007－9807(2008)03－0016－13

0 引 言

0.1 城市应急机构选址的研究现状

随着我国城市化速度的加快，自然灾害、事故灾害、突发事件不断的增加，城市聚集财富，同时也聚集风险. 统计资料表明，我国每年因突发公共事件造成的损失高达 GDP 的 6%. 这个数据突显出城市应急管理的重要性和紧迫性. 救援机构选址问题是影响应急响应效率的关键. 目前针对应急救援机构选址问题的研究[1~3]，将出救时间最短和成本最低作为系统目标线性优化模型分析. 但是现有的选址模型研究有两个缺点：(1)忽略了一些无法量化的空间制约因素，如选址的障碍物因素、环境制约因素、交通网络因素、地形制约因素等，会使选址结果缺乏科学性. (2)应急救援中心的选址是 NP－hard 问题，如果基于现有选址模型来分析这些空间因素，就会导致变量与约束条件数量的维数灾难，难以实现科学的应急救援中心选址计算. 本文的研究视角是将应急机构选址问题由模型驱动转变为数据驱动. 以潜在的海量救援对象为分析主线，克服线性模型的求解缺点，探索数据驱动的城市应急机构选址方法.

据 Open GIS[4] 统计，全世界数据库中信息的 75%到 80%带有与空间地理位置有关的特性，可见数据间包含着丰富的地理空间关系. 在城市应急管理中，居民小区、企业、楼宇等应急救援对象都是地理空间实体，应急救援行为是在一定的空间范围调度实现的，这些信息完全可以在 GIS 中可以实现存储并管理. 由于地理信息在公共突发事件和城市应急管理中的突出重要性，国内外很多学者提出基于地理信息系统[5](Geographic Information System, GIS)的城市应急事件管理方案，包括地震救援[6]、溢油应急[7]、消防应急[8,9]、9.11 应急响应[10,11]、洪水应急[12]、交通事故[13,14]、化学应急处理[15]等领域. 能够支持突发事件的可视化查询、应急调度、路线指挥等应用需求. 到目前为止，国内外 GIS 应急管理系统的研究还局限于使用 GIS 的基本视图功能[6]、初步分析功能[7~9]和特定领域的数学建模功能[7~11]，GIS 系统在城市应急管理中的作用远没有得到充分发掘.

① 收稿日期：2007－04－18，修订日期：2008－03－10.
基金项目：国家社会科学基金资助项目(07CTQ009)；
作者简介：樊 博(1975—)，男，黑龙江省呼兰人，博士后，讲师，Email：fanbo411@163.com

0.2　空间聚类挖掘的研究现状

空间聚类分析[16]（spatial clustering）是GIS功能的延伸. 它是一种数据挖掘方法，可以深入地发掘潜藏在地理空间信息的知识，找出某个或某几个空间数据集合的代表结点，在城市管理、商务领域中具有广泛的应用. 例如，某大型企业要在指定某市区范围内的 n 个客户建立 k 个客户服务中心，这种选址问题的基本原则是使城市中所有客户的行驶距离总和最短. 解决此类问题的空间聚类方法称为划分方法[13~16]. 典型的划分方法有两类；k-平均方法（k-mean）[13,16]和 k-中心点方法（k-medoid）[14,15]. k-平均方法是基于重心的聚类技术，其工作流程如下，首先，随机地选择 k 个对象，每个对象作为初始的聚类中心. 对于剩余的每个对象，使用它与初始聚类中心的距离值为相似度量，将每个对象付给最近的聚类中心. 然后，重新计算每个聚类对象集合的平均值，作为下一步的聚类中心. 这个过程不断循环，直到这个平方误差准则函数逐渐收敛.

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} | x - m_i | \qquad (1)$$

在式(1)中，x 是空间对象，m_i 是聚类集合 c_i 的均值. k-means 方法经常以局部最优结束，该算法力图找出平均误差值最小的 k 个划分. 它的复杂度是 $O(nkt)$，其中 n 是空间对象的总数目，k 是聚类的数目，t 是算法迭代的次数. 在通常情况下，$k \ll n$，$t \gg n$. 由于算法的复杂度较小，它的主要优点是能高效的处理大数据集；但是，k - 平均方法必须以平均值计算为前提，如果存在少量的“噪声”和孤立点数据，则会对平均值的计算产生极大的影响.

由于 k - 平均方法对噪声数据存在敏感性，因此产生 k - 中心点方法. 该方法不以每个聚类集合中空间对象的平均值为参照点，而以最接近聚类集合中心的空间对象作为中心点. 中心点方法不像平均值那么容易被极端的噪声数据所影响. 但是 k - 中心点方法要比 k - 平均值方法的执行代价高——计算复杂度大约是 $O(n*n)$. 这种划分方法的相似度函数仍然是基于最小化所有空间对象与参照点的距离总和. k - 中心点的工作流程是：首先为每个聚类集合选择一个代表对象；剩余的对象根据其与代表对象的距离分配给最近的一个聚类集合. 然后反复地利用其他对象来代替代表对象，从而不断的改进聚类的质量. 典型的 k-medoid[14,15]是CLARANS方法（Clustering LARge Application based on RANdomized Search，基于随机采样的大数据集聚类）. 现有的 k- medoid 方法在运行效率和解质量方面都不理想，需要进行针对性的提升. 本文拟改进现有空间聚类技术，将其应用到应急救援机构的选址上，实现更加科学合理的选址.

1　应急机构选址的数据模型

1.1　应急救援对象的空间数据建模

救援对象的地理空间位置是应急机构选址的根本依据，选址的原则是所有潜在救援对象与该应急中心的总距离最短. 本文以空间数据为分析主线，建立面向应急救援对象分析的数据模型，将应急管理对象的大量属性管理起来，其状如星型，称为星型模型，如图1右. 空间数据立方体是通过空间数据引擎[5,7]（Spatial Data Engineer，SDE）将应急救援对象的多个属性信息（星型模型）和GIS（见图1左）集成起来，实现城市信息和救援对象位置的可视化，在此基础上可以设计更为深入地空间分析功能，实现数据驱动的应急机构选址.

1.1.1　空间数据立方体的模型

该模型可以表示为 $MD = (D, H, M, \rho, \Gamma)$. 其中，

定义1　空间数据立方体的维

D——应急对象 O 的维属性集合. 它由一组 M 个非空间属性 ND 和 N 个空间属性的集合 SD 组成，$ND \subseteq D$，$SD \subseteq D$. 其中 $ND = \{ND_1, ND_2, \cdots, ND_m\}$，$SD = \{SD_1, SD_2, \cdots, SD_n\}$. 对于应急对象 O_i 的维值 D_i 可以表示为 $O(D_i)$.

$D_i(H) = \{H_1, H_2, \cdots, H_n\}$——维的层次关系集合. 其中 H_n 是 D_i 的对应的概念层次，如 $H_1 \leftarrow H_2 \leftarrow H_3$，：$year \leftarrow month \leftarrow day$

$D_i(H_j) = (\beta_{ij,1}, \beta_{ij,x}, \cdots, \beta_{ij,t})$——维的具体值集合. 其中，$\beta_{ij,t}$，是维 $D_i(H_j)$ 的一个具体值，t 是 $D_i(H_j)$ 的属性个数.

定义2　空间数据立方体的度量

$M = \{NM, SM\}$——空间数据立方体的度

量. 它包括数值度量 NM(如图 1 中的“事发次数”)和空间度量 SM(如图 1 中的“对象指针”),SM 是指向“空间对象地理位置”的空间指针集合.

$NM = (NM_1, NM_2, \cdots, NM_n)$—— 空间数据立方体的数值度量. 它是一组数值函数,如火灾次数,救火车数量等.

$SM = (SM_1, SM_2, \cdots, SM_n)$—— 空间数据立方体的空间度量. 它是大量的空间对象指针,例如救援对象位置点的指针等.

定义 3　空间数据立方体的度量函数

$\rho = (f_1(NM), \cdots, f_r(NM))$—— 数值度量的函数运算. 例如,某月火灾总次数、某区域急救机构总数,分别用 $Sum(M)$,$Count(M)$ 等函数表示.

$\Gamma = (F_1(SM), \cdots, F_s(SM))$—— 针对空间度量的函数运算,例如取集合函数 $Set()$、平均值函数 $Mean()$、空间拓扑关系函数 $Cover()$,$Touch()$,$Overlap()$ 等、空间距离关系函数 $Distance()$ 等.

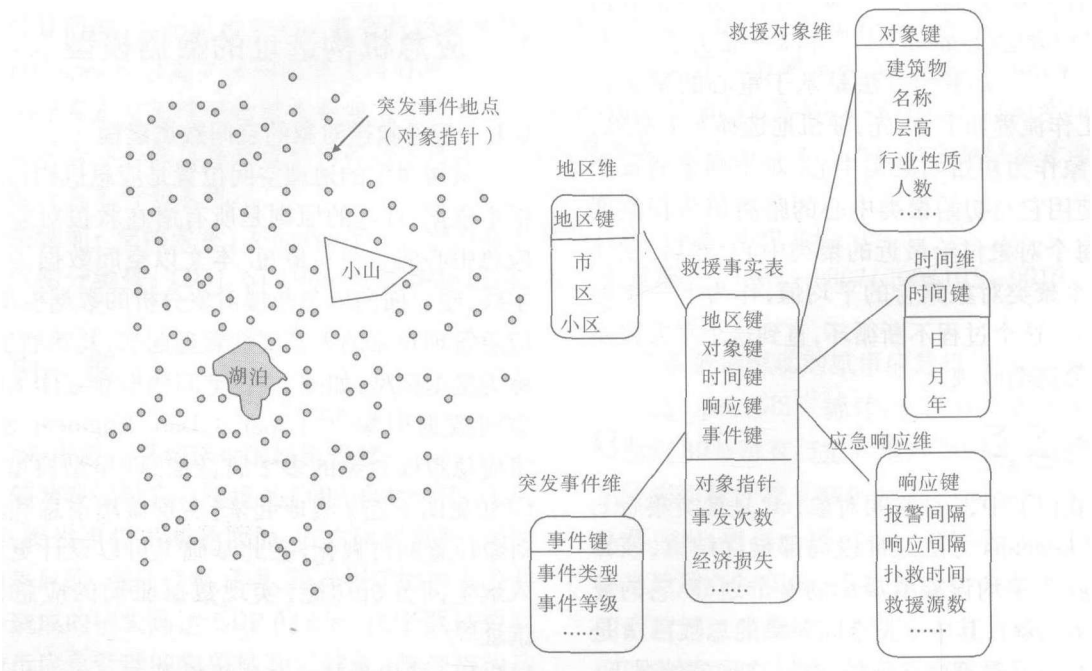

图 1　面向应急救援分析的空间数据立方体星型模型及救援对象的位置图

Fig. 1 Spatial data cube of urban emergency management and digital map of emergency spots

1.1.2　空间数据立方体的操作

定义 4　空间数据立方体的数值度量映射

空间数据立方体全面集成了面向应急救援对象分析的空间属性数据和非空间属性数据. 空间数据立方体的操作能够实现对救援对象位置数据的灵活分析. 它可以灵活地提供多角度的维属性分组操作,实现相关空间对象的数值度量查询,表示为 $\wedge_{\forall i,j,x}\{O[D_i(H_j)] \mid [D_i(H_j)]\theta\beta_{ij,x}\} \rightarrow f_r(NM)$. 例如取 D_i 分别为时间维、地区维和突发事件维;H_j 分别为年、城市和火灾的维层次,θ 为“等于”,$\beta_{ij,x}$ 的值分别为“1999 年,北京,重大火灾事件”,$f_r(NM)$ 为 $SUM()$ 函数,即累加事件数量. 则 $\wedge_{\forall i,j,x}\{O[D_i(H_j)] \mid [D_i(H_j)]\theta\beta_{ij,x}\} \rightarrow f_r(NM)$. 即为(1999 年,北京,重大火灾事件)→(事发次数).

定义 5　空间数据立方体的空间度量映射

空间数据立方体操作也可以灵活地提供多角度的维属性分组操作,实现相关空间对象的数值度量查询,表示为 $\wedge_{\forall i,j,x}\{O[D_i(H_j)] \mid [D_i(H_j)]\theta\beta_{ij,x}\} \rightarrow F_s(SM)$. 例如取 D_i 分别为时间维、地区维和突发事件维;H_j 分别为年、城市和火灾的维层次,θ 为“等于”,$\beta_{ij,x}$,的值分别为“1999 年,北京,重大火灾事件”,$F_s(SM)$ 为 $Set()$ 函数,即取位置点的集合. 则 $\wedge_{\forall i,j,x}\{O[D_i(H_j)] \mid [D_i(H_j)]\theta\beta_{ij,x}\} \rightarrow F_s(SM)$. 即为(1999 年,北京,重大火灾事件)→(事发地点集合).

定义 1 ~ 定义 5 描述了面向应急管理的空间数据立方体模型. 该模型的使用能够将应急管理

中的大量数据有机的组织到立方体模型中，通过空间数据立方体的多维、多层次组合操作，可以灵活地查询到有待深入分析的数据集合，为后续的选址分析提供数据源.

1.2　应急救援对象的空间拓扑关系谓词

定义6　空间数据立方体的空间度量函数

在定义3中，空间数据立方体的空间度量函数主要用于计算某实体对象与其它空间对象的关系，主要是指空间拓扑关系谓词的计算. 它可以用来描述救援对象周边的环境信息，这些信息是影响应急机构选址的重要因素. 空间拓扑关系是语义最丰富的空间谓词，主要是描述与位置临近的空间对象之间的空间关系，例如谓词 *Touch*(x, *hill*) 表示空间对象 x 与山界相交，空间拓扑关系谓词包括相交 *Touch*( )、邻接 *Adjacency* ( )、包含 *Contain*( )、相离 *Disjoin*( )、重合 *Overlap*( ) 等.

临近对象的属性信息是影响应急机构选址的重要因素. 例如 *Overlap*(x,山坡) 只能表述对象 x 与山坡位置重合，*Overlap*(x,高斜度山坡) 是更为有指导意义的谓词，因为山坡的斜度才是选址分析的关键. 所以，空间拓扑关系谓词应具有高信息精度，这需要将临近空间对象的属性信息进行科学的离散化和概念化，采用概念树(Concept tree)来组织和表示数值数据，经典的概化方法有AOI[13] 等.

算法1　空间度量函数中的空间拓扑关系谓词的计算(定义6的实现算法)

Input

1) 问题相关的空间数据(通过空间数据立方体模型的维度和层次的组合，实现空间度量映射，即模型中定义5的 *Set*( ) 操作).

2) 临近对象的非空间属性概念树——Concept tree(通过属性离散[16] 和概念攀升[13] 实现).

3) 问题相关的空间拓扑关系谓词(即选址问题中的空间制约因素，包括空间环境因素等. 例如，空间环境因素“选址出口道路的宽度大于60M——*Touch*(x,*wideroad*)”，“不能临近铁路——*Not-close-to*(x, *railway*)” 等.

Output

空间拓扑关系谓词集合

Method

Step 1 Relevant_Themes = Extract_task_relevant_theme(*Spatial Date Cube*);/ 基于空间数据立方体提取GIS中相关的空间图层.

Step 2 Relevant_SDB = Extract_task_relevant_objects(*Relevant_Themes*)/ 从相关空间图层中选出任务相关的空间对象集合.

Step 3 MBR_Predicate_DB = Find_MBR_predicates(*Relevant_SDB*, *Taskrelated predicate*);/ 用MBR方法粗略估计出与分析问题有关的空间拓扑关系谓词，MBR是GIS中近似空间实体的最小边界矩形，即空间对象的索引. 其主要思路是利用空间对象的索引MBR来近似这个实体，使用空间对象的MBR来判断两个对象的关系(详见文献16).

Step 4 Spatial_Predicate = Compute(*MBR_Predicate_DB*, *Taskrelated objects*, *Concept tree*);/ 依据属性值的概念树，计算问题需要的高精度空间拓扑关系谓词.

2　空间距离关系的函数计算

2.1　空间距离关系的函数

目前空间距离的计算多采用空间对象坐标点之间的直接几何距离(Euclidean distance)进行计算. 现实世界中，空间实体之间的实际距离往往不是直接的空间几何距离，两者间可能存在着不可穿越的障碍物，如建筑物，河流，高速公路等. 因此，空间对象之间距离的计算就要充分考虑到障碍物的影响，如图1中的左侧地图[8].

定义7　直接空间距离[9]

已知(1) 一个 n 个点的 P 集合：$\{p_1, p_2, \cdots, p_n\}$，其中 $\forall p_i$ 为某个救援位置的点坐标，$1 \leqslant i \leqslant n$；(2) 一组互不相交的障碍物集合 O，$O = \{o_1, o_2, \cdots, o_m\}$，$\forall o_i \in R, 1 \leqslant i \leqslant m$，$R$ 为二维空间区域. 任意两点 p_i 和 p_j，两者的直接空间距离(Euclidean distance) 如公式2

$$d(p_i, p_j) = \sqrt{(p_i x - p_j x)^2 + (p_i y - p_j y)^2} \tag{2}$$

定义8　如果考虑障碍物的存在，则两点的距离可表示为 $d'(p_i, p_j)$，它是指绕过障碍物后两点间的最短距离. 障碍物 o_i 是一个拓扑多边形，用

$P(V,E)$ 来表示，V 是形成障碍物多边形的系列点集合：$V = \{v_1, v_2, v_3, \cdots, v_k\}$，$E$ 是形成该多边形的线集合：$E = \{e_1, e_2, \cdots, e_k\}$，其中 e_i 是点 v_i 和 v_{i+1} 连线，$1 \leqslant i \leqslant k$. 障碍物多边形有两类：凹状体和凸状体.

定义 9 “可见性”是两个数据点之间的一种关系，如果两个数据点之间的连线与 $P(V,E)$ 所代表的障碍物不相交，则称两个数据点具有“可见性”. 一个具有 n 个数据点的 D 集合，$D = \{d_1, d_2, \cdots, d_n\}$，$l$ 为连接 d_i 和 d_j 的线段，$d_i, d_j \in D, i \neq j$，$i, j \in [1, \cdots, n]$. $\forall e_i \in E$，如果 $\neg\ \exists p \in l \cap e_i$，则 d_i 和 d_j 是“可见的”[10].

在文献[10]中研究考虑障碍物的空间对象距离计算方法. 该方法提出 BSP_tree 的空间数据结构来确定两个空间对象 p_i 和 p_j 是否具有“可见性”(Step 2). 如果具有“可见性”，即空间对象之间的距离可以采用直接的空间距离来计算(Step 6). 否则当 p_i 和 p_j 之间存在障碍物时，首先判断该障碍物的形状，根据该障碍物体的形状为凹状体(Step 4) 或凸状体(Step 5) 来确定救援位置与障碍对象的连接策略. 凹状体障碍物与凸状体连接策略并不相同. 对于凹状体来说，凹点是不需要连接的，只有凸出点才是跨越该障碍物的连接结点；对于凸状体来说，通过选择空间对象与障碍物的“可见点”将空间对象 p_i 和 p_j 连接成若干通路，每个凸点都需要连接，见图 2. 这种情况下，空间对象间的距离计算就需要在图中选择出空间对象 x 和 y 之间的最短路径 $d'(p_i, p_j)$，通用的计算方法是运筹学中的 Dijkstra 算法.

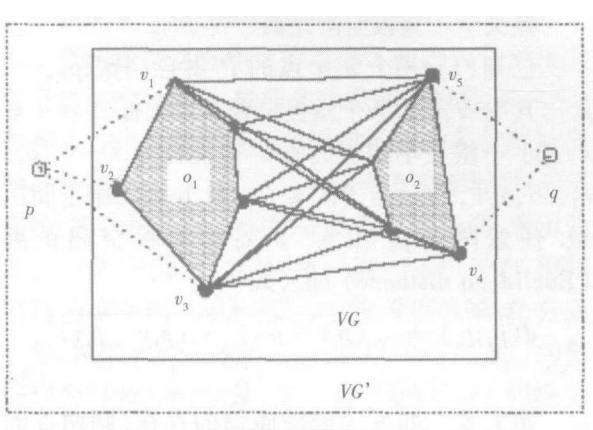

图 2 带有障碍物的可见图

Fig. 2 Visibility with obstacles

2.2 空间距离关系的实现算法

算法 2 存在障碍物的空间距离计算方法

Input

(1) 救援对象位置的电子地图 / 通过空间数据立方体模型操作和空间度量映射得出大量地理位置点

(2) 障碍物的电子地图 —— $P(V,E)$ / 通过 GIS 电子地图中的障碍物图层得出

Output

距离 ——*Distance* / 距离计算的值

method

Step 1 Element: = Fetch(location map) / 救援对象位置的电子地图

Step 2 Result: = CheckVisibility (p_i, p_j, $P(V,E)$) / 采用 BSP-tree[10] 检查两个空间对象之间是否存在障碍物

Step 3 If result: = true

Then shape: = Checkshape($P(V,E)$) / 检查障碍物的形状

Step 4 If shape: = Concex / 凸状体障碍物

Then Constructconcexgraph($p_i, P(V, E), p_j$)/ 采用凸状体障碍物的连接策略

Calculate_Dijkstra_distance(p_i, p_j)/ 当两个空间对象之间有障碍物时，采用 Dijkstra 算法计算两者间的间接距离 ——$d'(p_i, p_j)$

Step 5 Else / 即 Shape: = Concave—— 凹状体障碍物

Constructconcavegraph($p_i, P(V,E), p_j$)/ 采用凹状体障碍物的连接策略

Calculate_Dijkstra_distance(p_i, p_j)/ 当两个空间对象之间有障碍物时，采用 Dijkstra 算法计算两者间的间接距离 ——$d'(p_i, p_j)$

Endif.

Step 6 Calculate_Euclidean_distance(p_i, p_j)/ 当两个空间对象之间无障碍物时，采用公式 2 计算两者间的直接距离

Endif

Step 7 Return *Distance*

定义 10 空间对象的加权点

空间对象的加权可表示为 $w*(x,y) = \{(x,$

$y)_1,(x,y)_2,\cdots,(x,y)_w \mid w = 1,2,\cdots,n\}$. 因为每一个应急救援对象都是以点的形式表示在电子地图上的，应急救援对象的行业性质决定了其潜在突发事件的可能性. 例如，化工炼油厂和普通居民小区都是应急救援对象，但两者潜在的预警等级显然是不同的. 传统的研究往往将两者等量齐观，在选址问题中没有侧重. 因此，本文提出了加权点概念，根据预警等级对点坐标加权. 例如某建筑预警等级为 $w = 5$ 级，则作为 5 个的普通建筑看待.

3　面向选址问题的空间聚类算法

3.1　算法的数据流程框架

本文提出面向选址问题的空间聚类算法（算法 3）. 该算法包括五个子算法（算法 4,5,6,7,8），算法 3 中的这五个算法之间的输入输出关系，算法 3 与上文算法 1,2 之间的数据流程和调用关系，以及空间数据立方体与算法 1,2,3 之间的数据供给关系如下图 3 所示.

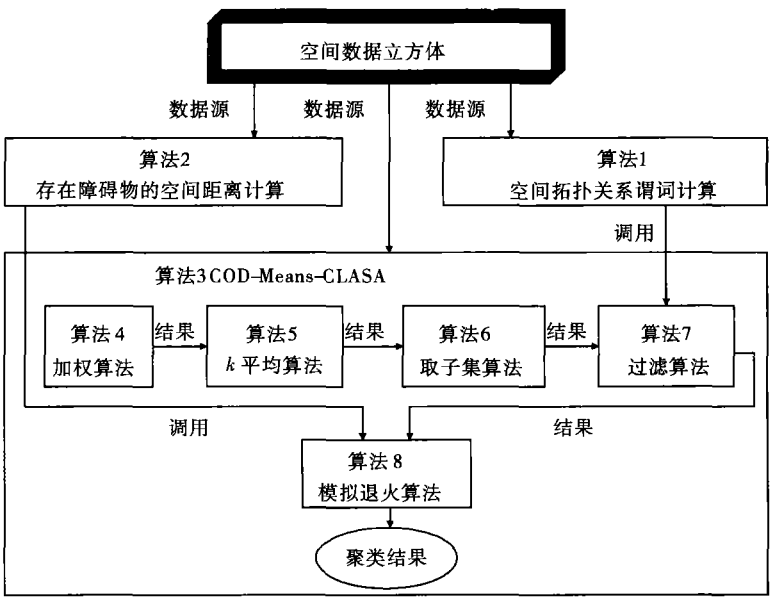

图 3　算法的数据流程框架

Fig. 3 The process of the proposed method

在图 3 中，空间数据立方体操作可以为算法 1,2,3 提供输入数据源；算法 4，算法 5，算法 6，算法 7 和算法 8 是顺序串行关系，前序算法为后续算法提供输出结果；其中算法 7 在运行过程中要调用算法 1 的空间谓词计算功能，算法 8 在运行过程中要调用算法 2 的空间距离计算功能.

3.2　COD-MEANS-CLASA 聚类算法

文献[4,17] 研究了“存在障碍物的空间聚类算法 (Clustering with Obstructed Distance, COD)”. 提了出 COD-CLARANS 的聚类方法，它将 CLARANS 算法的思想与“存在障碍物的聚类问题” 相结合，解决了存在障碍物的物流中心选址问题. 文中比较了 k-means 方法与 k-medoid 方法在解决“存在障碍物的聚类问题” 的差异. 由于 k-means 方法以聚类集合中空间对象的平均值为聚类中心点，而 k-medoid 以最接近聚类集合中心的空间对象作为中心点. 如果考虑障碍物的存在，两种方法不同的计算机制会产生不同的聚类结果，如图 4[10]. 显然，采用 k-means 方法得出的聚类中心很可能在障碍物上，对于选址问题是没有实际意义的；而采用 k-medoid 方法才能有效的选择某一空间对象为聚类中心，并解决了 k-means 方法的对此类问题的求解缺陷. 然而，COD-CLARANS 方法计算复杂度大约是 $O(n^* n)$，n 是空间对象的数量，当 n 的数值很大时，算法存在比较严重的执行效率问题.

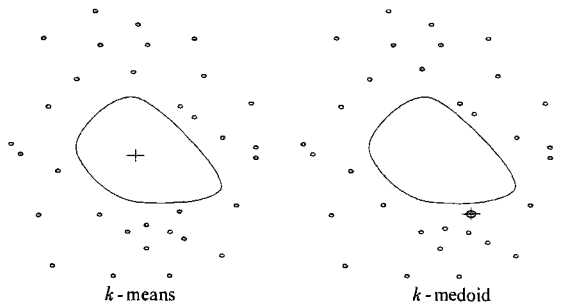

图 4 两种聚类方法的结果比较

Fig.4 The comparison of the results of the two clustering methods

研究表明[18,19],最佳的聚类结点存在于距离该聚类集合的平均值较近的位置. 基于此,本文拟利用 k-means 方法执行代价低、适用于大数据量的优点. 首先初步计算出各个聚类集合的中心点,如果在 COD 问题中,无论该中心点是直接可达到的还是在障碍物上,最佳的聚类结点很可能是距离聚类中心点的直接空间距离较近的点. 所以在每个聚类集合中,仅将与该集合平均值直接空间距离比较近的点作为最佳聚类求解的子集合,该子集合的空间对象数目将远小于它的父集合,这些子集合将极大地缩小问题的解搜索空间.

本文在每个聚类集合中以 k-means 算法形成的聚类平均值为中心,采用 GIS 的缓冲区分析功能,以这些中心点为圆心,捕捉一定数量的空间对象,形成每个聚类集合的子集合求解空间. 最佳聚类结点的求解是 NP_Hard 问题,模拟退火算法[18~21]在求解质量方面具有强大的优势. 在以上研究的基础上,本文提算法 3,该算法以围绕每个聚类集合的平均值为中心,获取其周围的空间对象集合,然后使用模拟退火算法为解搜索策略,以期达到执行代价小和求解质量高的双重效果.

算法 3 COD_Means_CLASA

Input:

(1) 聚类中心的数目(选址的数目)——k

(2) k-means 算法执行次数 ——m

(3) 在每个集合中选择子集合的元素数目 ——r

(4) 空间数据立方体模型操作和空间度量映射得出的大量地理位置点 ——$Set()$

(5) 位置地点的权重 ——w

(6) 空间障碍物 $P(V,E)$ 图

(7) 空间环境约束因素

(8) 成本约束条件

(9) 初始温度 T_0,冷却温度 T_a

(10) 参数 t,δ,ρ,λ

Output

k 个聚类集合和 k 个最佳聚类中心点

method

Step 1 Running **Weigh Algorithm.** /运行加权算法(算法 4). 如果某个空间对象的权值是 w,则在电子地图上将这个空间对象计为 w 个,体现应急机构选址对重点救援对象的侧重性,产生经过加权的空间数据对象集合 $Weight\text{-}set()$.

Step 2 Running **K-means Algorithm.** /运行 k 平均方法(算法 5). 在 $Weight\text{-}set()$ 包含的空间对象集合中,使用 k 平均方法找出 k 个平均方差最小值的坐标点及相应的 k 个聚类集合,其中 k 平均方法的循环次数为 m.

Step 3 Running **Subset Algorithm.** /运行取子集算法(算法 6). 分别以 k 个平均点为中心,采用 GIS 软件的缓冲区分析功能逐一在每个聚类集合中选取 r 个临近的点,形成子集合 $\{Subset_1(), Subset_2(),\cdots,Subset_k()\}$.

Step 4 Running **Elimination Algorithm.** /运行过滤算法(算法 7). 计算空间环境制约因素、空间地形因素和成本因素对选址的限制,其中空间环境因素、空间地形因素可通过空间度量函数中的拓扑关系谓词的计算而实现分析. 分别从 $Subset_1(), Subset_2(),\cdots,Subset_k()$ 中进一步过滤不合格解,形成解侯选集合 $Candidateset_1(), Candidateset_2(),\cdots,Candidateset_k()$

Step 5 Running **CLASA Algorithm.** /运行模拟退火算法(算法 8). 分别在集合,$Candidateset_1(), Candidateset_2(),\cdots,Candidateset_k()$ 中任意选取 1 个点,共计 k 个点 $C=\{c_1,c_2,\cdots,c_k\}$,作为聚类中心点的初始解. 采用的模拟退火算法在这个小范围内寻找最优解.

Step 6 Return the k medoid and k clusters

3.3 COD-MEANS-CLASA 的调用算法

算法 4 Weigh Algorithm(加权算法)

Input

(1) 空间数据立方体的空间度量操作所产生救援对象位置图 ——$Set()$

(2) 位置地点的权重 ——w

Output

(3) 经过加权的空间数据对象集合 ——$Weight\text{-}set()$

Method

Weight-set( ) : = *set*( )

For each coordinate (x_i, y_i) in *Weight-set*( ) do / 对于每个在 *Weight-set*( ) 集合中的元素坐标 $(x_i, y_i) \Leftarrow w_i$ / 将集合中的每个元素赋予相应的权值.

For j = 1 to w − 1;j ++

Weight-set( ) : = $\bigcup_{j=1}^{w-1}(x_i, y_i)$ / 将点坐标按照权值的数量合并进入 *Weight-set*( ) 集合

Endfor;

Endfor;

算法 5　K-means Algorithm / 该算法为空间聚类的经典算法之一,详见文献[16][13].

算法 6　Subset Algorithm(取子集合算法)

Input

(1)*k-means* 算法得出的 k 个中心点——*k mediods*

(2)*k-means* 算法得出的 k 个聚类集合——*k clusters*

(3)r 个备选中心点

Output

(4) 形成子集合{$Subset_1$( ), $Subset_2$( ),…, $Subset_k$( )}

Method

For i = 1 to k do

$Subset_i$ ( ) : = GIS buffering function (*mediod*$_i$, r, *culster*$_i$) / 采用 GIS 软件的缓冲区分析功能,分别以 k 个中心点(*mediod*) 为圆心,在相应的 k 个聚类集合(*cluster*) 中每个集合选取 r 个临近的点, 形成解侯选集合{$Subset_1$( ), $Subset_2$( ),…, $Subset_k$( )}

End for;

算法 7　Elimination Algorithm(过滤算法)

Input

(1) 算法 6 产生的子集合{$Subset_1$( ), $Subset_2$( ),…, $Subset_k$( )}

(2) 空间环境制约因素(例如选址的出口道路宽度大于 60M——*Touch*(x, *wide-road*),选址不能临近铁路 ——*Not-close-to*(x, *railway*) 等)

(3) 成本因素(例如动迁成本小于6万元/平方米)

Output

(4) 形成解侯选集合{$Candidateset_1$( ), $Candidatestet_2$( ),…, $Candidateset_k$( )}

Method

For i = 1 to k do

$Cadidateset_i$( ) := $Subset_i$( )

For each *mediod* in $Cadidateset_i$( ) do

Calling **algorithm 1** / 计算空间拓扑关系谓词

If *Mediod. spatial predicate* : = *Environmental factor restriction* Then / 基于空间拓扑关系谓词判断选址地点的空间环境因素

$Cadidateset_i$( ) := $Cadidateset_i$( )-{*mediod*}/ 在集合中将该点去除

Else if *Mediod. cost* := *Cost factor* Then / 基于空间拓扑关系谓词判断选址地点的成本因素

$Cadidateset_i$( ) := $Cadidateset_i$( )-{*mediod*}/ 在集合中将该点去除

Endif

Endfor;

Return $Cadidateset_i$( )

Endfor;

Metropolis 等人在 1953 年提出了模拟退火算法[16],模拟退火是物理退火过程的计算机模拟,物理退火是先将固体加热到一定温度,使之融化,然后在缓慢降温,达到凝固点,形成晶体. 模拟退火算法的基本思想是把某类优化问题的求解与统计热力学中的热平衡问题进行对比,试图通过模拟高温物体退火过程的方法,来找到优化问题的全局最优或近似全局最优解. 在物体的降温退火过程中,其能量转移服从玻尔兹曼(Boltzmann)分布规律公式(3)

$$P(E) = \exp(-E/kT) \tag{3}$$

在公式中,$P(E)$—— 系统处于低能 E 的概率;k—— 玻尔兹曼常数;T—— 系统温度. 为方便起见,通常将 k 算入 T 中. 随着温度 T 的降低,物体处于高能状态的概率就逐渐减小,最后当温度下降到充分低时,物体的能量状态概率为 1,稳定在低能状态. 即随着温度的降低,系统的活动性逐渐降低,使玻尔兹曼分布达到最低状态,最终以概率 1 稳定在全局最小区域. 另外,当降温过程中能量偶尔上升时,即 $P(E)$ 大于一个给定的概率值时,该算法允许一定限度内容纳次优解,以便跳出局部最优区域,提高组合优化问题的解搜索质量.

模拟退火过程的降温过程要足够长以使系统达到稳定状态，特别是当温度接近凝固点时，如果降温过快，晶体就会形成瑕疵. 模拟退火算法实现的关键是选择合适的“冷却程序”. 冷却程序包括初始状态的选择，Metropolis 过程稳定性的检验、降温策略、循环次数、终止条件的设置等. 下面介绍基于模拟退火算法的救援机构位置选取方法，见算法 8.

算法 8　CLASA Algorithm(模拟退火算法)

Input:

(1) 初始的聚类中心点 $C = \{c_1, c_2, \cdots, c_k\}$

(2) 初始温度 T_0，冷却温度 T_a

(3) 解集合 $S = \{Cadidateset_1(), Cadidateset_2(), \cdots, Cadidateset_k()\}$

(4) 参数 t, δ, ρ, λ

Output

k 个聚类集合和 k 个最佳聚类中心点

method

Repeat

Begin

Generate (C' form S) / 在解集合 S 中，通过对初始解 C 中的聚中心点的交换，随机生成另一个侯选解 C'.

$\triangle E = E(C') - E(C) / E(C)$ 是解 C 为聚类中心点集的、考虑障碍物的平方误差值总和，它是通过调用算法 2 计算得出的.

If $\triangle E < 0$　　/ 如果 $\triangle E$ 小于零，即 C' 平方误差总和比 C 小

Then $C := C'$ and $n := n + 1$ / 则 C' 是比 C 更好的解，并累计 $\triangle E < 0$ 的数量 n.

Else if $e^{-\Delta E} / T_t \geqslant random[0,1)$/ 给定一个参数 $random[0,1)$，如果玻尔兹曼分布概率不小于该参数.

Then

$C := C'$ // 则在该概率值下，C' 被接受，从而跳出了局部最优区域，提高解质量.

End;

$t := t + 1$;

CALCULATE_LENGTH (n) / 这里 n 是指在求解循环中 $\triangle E < 0$ 的次数. 如果 $n > \rho$，ρ 是一给定参数，则转到验证 stop criterion.

CALCULATE_CONTROL(T_t); / 计算控制温度 T_t，即退火进度表的设计，在本例中 $T_t = T_0\delta^t$. 其中 t 为循环次数，δ 为某一常数，$0 < \delta < 1$.

Until stop criterion;/ 本例的终止规则有两条，满足任意一条则终止：① 给定一个 medoids 组合数目 λ，求解的循环次数 t 大于 λ；② 控制温度 $T_t < T_a$，T_a 是给定的最终冷却温度. 否则随着 t 的增加，控制温度 T_t 将降低，然后继续返回求解.

Return $\{C\}$ / k 个聚类集合和 k 个最佳聚类中心点

End;

3.4　实验分析

(1) 实验数据　为了验证算法的可行性和有效性，本文采用 GIS 应用程序分别生成 10 000 和 12 000 个点数据，以不同的北京市区域电子地图为母版地图，叠加在一起展开实验分析. 每个点数据以 (x,y) 坐标给出，代表潜在救援对象的位置. 城市中的障碍物包括公园、湖泊、社区等，用实体顶点和顶点连线 (V,G) 表示，如下图 5 所示. 以该数据源为基础，比较运行两种算法的效率.

图 5　两个实验数据集合

Fig. 5 The datasets of two experiments

表 3　两种算法的实验 2 结果

Table 3 The results of two algorithms

COD-CLARANS		COD-MEANS-CLASA	
算法所计算的总距离 (10^5 m)	应急救援的平均距离 (10^3 m)	算法所计算的总距离 (10^5 m)	应急救援的平均距离 (10^3 m)
9.72	8.02	4.05	7.76
9.90	8.09	4.08	7.77
9.82	8.04	3.99	7.71
9.95	8.11	4.06	7.76
9.79	8.07	4.03	7.75
9.83	8.08	4.01	7.75
9.82	8.07	4.07	7.76
9.77	8.03	4.08	7.77
9.79	8.04	4.09	7.78
9.80	8.05	4.00	7.72

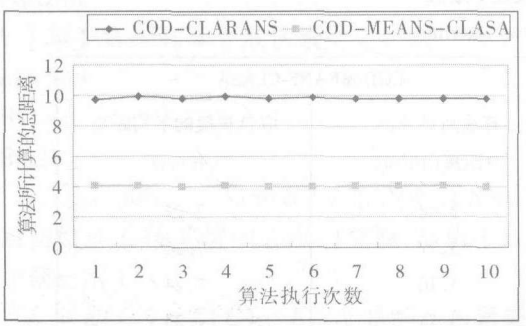

图 8　两种算法所计算的总距离(实验 2)

Fig. 8 Total distances calculated by two algorithms(2)

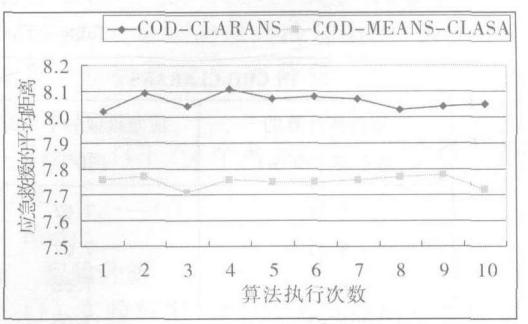

图 9　两种算法的平均应急救援距离(实验 2)

Fig. 9 Average response distances of two algorithms(2)

(3) 实验结果　实验 1 和实验 2 得出的选址中心点和该中心点覆盖的聚类集合分别如下图 10,图 11 所示. 黑色框图为一个聚类集合,打"×"的地方就是应急救援中心的选址地点.

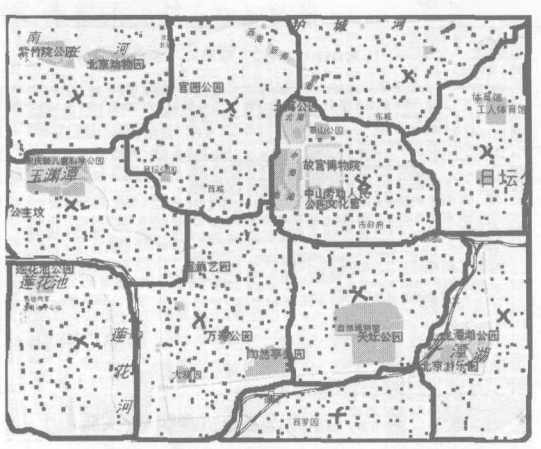

图 10　实验 1 的聚类结果

Fig. 10 The clustering result of experiment 1

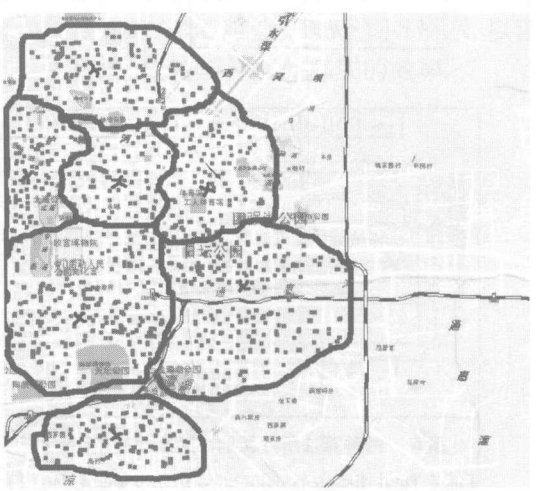

图 11　实验 2 的聚类结果

Fig. 11 The clustering result of experiment 2

4 结　论

本文研究了面向应急服务机构选址的空间聚类方法.以潜在的海量救援对象为分析主线,在空间数据立方体操作的基础上,首先对救援对象的位置数据进行分析.然后提出COD-MEANS-CLASA算法,以救援距离最短为启发信息,综合考虑了空间障碍因素和空间环境因素,实现了传统线形优化模型无法解决的NP-hard选址问题,探索出基于数据驱动的城市应急机构选址方案.通过与其他空间聚类方法进行对比,实验结果证明了本文提出方法的有效性和优越性.

参 考 文 献:

[1]Vladimir Marianov, Charles Revelle. The queueing maximal availability location problem: A model for the sitting of emergency vehicles[J]. European Journal of Operational Research, 1996, 4(93): 110—120.

[2]方　磊, 何建敏. 城市应急系统优化选址决策模型和算法[J]. 管理科学学报, 2005, 1(8): 12—15.
Fang Lei, He Jianmin. Optimal location model and algorithm of urban emergency systems[J]. Journal of Management Sciences in China, 2005, 1(8): 12—15. (in Chinese)

[3]贾传亮, 池　宏. 基于多阶段灭火过程的消防资源布局模型[J]. 系统工程, 2005, 9(23): 12—15.
Jia Chuanliang, Chi Hong. The allocation model of fire resource based on multistage fire protection process[J]. Systems Engineering, 2005, 9(23): 12—15. (in Chinese)

[4]Tung A K H, Hou J, Han J. Spatial Clustering in The Presence of Obstacles[R]. Proceeding of 2001 International Conference On Data Engineering, 2001. 359—367.

[5]Korte G P E, Koret G P. The GIS Book: Understanding the Value and Implementation of Geographic Information Systems [M]. USA: Delmar Publishers, January 2002. 36—62.

[6]刘彦呈. 海上溢油应急反应基于GIS的模拟训练系统研究[J]. 系统仿真学报, 2004, 11(5): 18—22.
Liu Yan-cheng. The research and development of the simulation training system of marine oil spill crisis response[J]. Acta Simulata Systematica Sinica, 2004, 11(5): 18—22. (in Chinese)

[7]邹　亮. 基于GIS的灾害疏散模拟及救援调度[J]. 自然灾害学报, 2006, 6(15): 141—145.
Zou Liang. GIS-based evacuation simulation and rescue dispatch in disaster[J]. Journal of Natural Disasters, 2006, 6 (15): 141—145. (in Chinese)

[8]Marc Bonazountas. A decision support system for managing forest fire casualties [J]. Journal of Environmental Management, Available online, 2008, 22(8): 23—32.

[9]Michael J Kevany. GIS in the World Trade Center attack-trial by fire [J]. Computers, Environment and Urban Systems, 2003, 27(6): 571—583.

[10]Kwan Mei-Po, Lee Jiyeong. Emergency response after 9/11: The potential of real-time 3D GIS for quick emergency response in micro-spatial environments[J]. Computers, Environment and Urban Systems, 2005, 29(2): 93—113.

[11]Andre Zerger, David Ingle Smith. Impediments to using GIS for real-time disaster decision support[J]. Computers, Environment and Urban Systems, 2003, 27(2): 123—141.

[12]Al-Sabhan W, Mulligan M. A real-time hydrological model for flood prediction using GIS and the WWW[J]. Computers, Environment and Urban Systems, 2003, 27(1): 9—32.

[13]史忠植. 知识发现[M]. 北京: 清华大学出版社, 2002. 125—167.
Zhongzhi Shi. Knowledge Discovery[M]. Beijing: The Press of Tsinghua University, 2002. 125—167. (in Chinese)

[14]Raymond T Ng, Han Jiawei. CLARANS: A method for clustering objects for spatial data mining[J]. IEEE Transaction on Knowledge and Data Engineering. 2002, 14(5): 1003—1016

[15]Martin Ester, Hans-Peter Kriegel. Clustering for mining in large spatial databases[J]. Special Issue on Data Mining. KI-Journal, 1998, 9(1): 332—338.

[16]Han Jiawei. Data Mining: Concepts and Technique[M]. Canada: Morgan Kaufmann, 2000. 23—56; 77—106.

[17]Zaiane O R, Lee Chi-Hoon. Clustering spatial data in the presence of obstacles: A density-based approach[J]. International Database Engineering and Applications Symposium (IDEAS'02). Edmonton, Canada July, 2002, 8(9): 214—224.

[18]Chu S C, Roddick J F, Pan J S. A Comparative Study and Extensions to K-medoids Algorithms[R]. In Fifth International Conference on Optimization: Techniques and Application, Hong Kong, China, 2001. 1708—1717.

[19]Chu S C, Roddick J F, Pan J S. An Efficient K-medoids-based Algorithm Using Previous Medoid Index, Triangular Inequality Elimination Criteria and Partial Distance Search[R]. The 4th International Conference on Data Warehousing and Knowledge Discovery, Aix-en-Provence, France, 2002. 63—72.

[20]洪家容. 归纳学习——算法、理论、应用[M]. 北京: 科学出版社, 1996. 1—25.
Hong Jiarong. Induction Learning: Method, Theory and Application[M]. Beijing: Science Press, 1996. 1—25. (in Chinese)

[21]Kirkpatrick S, Gelatt C D, Vecchi Jr MP. Optimization by simulated annealing[J]. Science, 1983, 200(45): 671—680.

[22]Han Jiawei, Kamber M, Tung A K H. Spatial Clustering Methods in Data Mining: A Survey[R]. Simon Fraster University, Computer Science Technical Report, 2000. 1—28.

[23]Ester M, Kriegel H P, Sander J, *et al.* A Density-Based Algorithm for Discovering Clusters in Large Spatial Database With Noise[R]. Proceeding of Knowledge Discovery and Data Mining(KDD), 1996. 226—231.

[24]Tung A K H, Han J. Constraint-based Clustering in Large Databases[R]. Proceeding of 2001 International Conference On Database Theory, 2001. 405—419.

Spatial clustering mining method for site selection problem of emergency response center

FAN Bo

School of International and Public Affairs, Shanghai Jiaotong University, Shanghai 200030, China

Abstract: This article uses urban emergency response management as the background. Spatial clustering method is adopted to solve the site selection problem of emergency response center. Firstly, a new data model of emergency management for clustering analysis is given. Secondly, a new spatial clustering method named COD-MEANS-CLASA algorithm is proposed to realize the site selection of emergency center. It has the advantages of applying k-means algorithm to reduce result space, using CLASA as result-searching strategy. On the basis of GIS functions, we design a deeper analytical function by incorporating spatial obstacle factors and spatial environmental factors. The experiments have proved that the algorithm does better in both performance efficiency and result quality.

Key words: spatial clustering; k-means algorithm; CLASA algorithm; site selection; emergency response center