DOI:10.13471/j.cnki.acta.snus.2015.06.018

基于路网拓扑的聚类分析算法研究与实现*

黄 敏[1]，李尔达[1]，袁 媛[2]，郑 健[1]
（1. 中山大学工学院//广东省智能交通系统重点实验室，广东 广州 510006；
2. 北京交通大学城市交通复杂系统理论与技术教育部重点实验室，北京 100044）

摘 要：聚类分析是一种统计分析方法，其目的是在相似的基础上对数据进行分类。该研究在基于路网拓扑的条件下，提出一种全新的算法实现交通领域中兴趣点集的聚类分析。算法为从集合中任取一个兴趣点，向邻接结点扩展，沿着路网拓扑的方向进行广度搜索，以凝聚度（α）和扩展度（β）为聚类限制条件，将满足限制条件的兴趣点与选定的兴趣点聚类。最后，把算法应用于珠江新城路网，分析行人的可达性情况。

关键词：聚类分析；路网拓扑；兴趣点；算法；行人可达性
中图分类号：TP311.13 **文献标志码**：A **文章编号**：0529－6579（2015）06－0099－－05

Research and Implementation of Clustering Analysis Algorithm Based on Road Network Topology

HUANG Min[1], *LI Erda*[1], *YUAN Yuan*[2], *ZHENG Jian*[1]
（1. School of Enginnering//Guangdong Provincial Key Laboratory of Intelligent Transportation System, Sun Yat-sen University, Guangzhou 510006, China;
2. Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Beijing 100044, China）

Abstract: Clustering analysis is a statistical analysis method, which aims to classify the data from their similarities. Based on road network topology, a new algorithm had been put forward to achieve the clustering analysis for points of interest (POI) in transportation field. The algorithm is that POI taken from point sets would be extend to the adjacent node and has a breadth of search along the direction of the network topology with limiting conditions of aggregating degree (α) and extending degree (β). POI those meet the conditions would be clustered together. At last, this algorithm is applied in the Guangzhou Pearl River Metro network to analysis pedestrian accessibility.

Key words: clustering analysis; network topology; POI; algorithm; pedestrian accessibility

“物以类聚，人以群分”。在自然科学和社会科学中，存在着大量的分类问题。聚类分析是分类问题的一种统计分析方法，其目的是在相似的基础上对数据进行分类[1]。聚类分析作为数据挖掘的核心技术，在不同的应用领域都得到了很好的发展。其中，对空间数据的聚类分析一直是研究热点之一。空间聚类分析是指通过对空间数据进行聚类，发现数据属性之间的相互关系[2]。聚类分析算法在交通领域有着广泛的应用[3]。在交通诱导系统中，通过对兴趣点的聚类分析得到的区域聚

* **收稿日期**：2015－02－04
基金项目：广东省科技计划资助项目（2014B010118002，2015B010110005）；广州市科技计划资助项目（201510010247）；高校基本科研业务费资助项目（15lgpy10）
作者简介：黄敏（1975 年生），女；**研究方向**：路网数据模型与道路交通标志标识系统；E-mail：huangm7@mail.sysu.edu.cn

类，可用于指路系统中的整体诱导等；在交通规划中，通过对兴趣点的聚类分析得到的可达性聚类，可用于分析居民出行的可达性[4]。

目前的空间聚类分析算法，大多数都是针对欧式几何空间中的数据对象[5-6]。然而在交通领域的应用中，空间对象的访问受限于道路网络。而且，很多兴趣点间的路网距离和欧式距离有很大的差别[7]。如图 1，从新华社广东分社到广东省政府，欧氏距离约为 270 m。而实际上如图 2，从新华社广东分社到广东省政府，要经过东风中路，路网距离约为 3.5 km。可见，网络距离远远大于欧式距离。因此，在交通领域中，研究基于路网拓扑的聚类分析算法更具有应用价值。

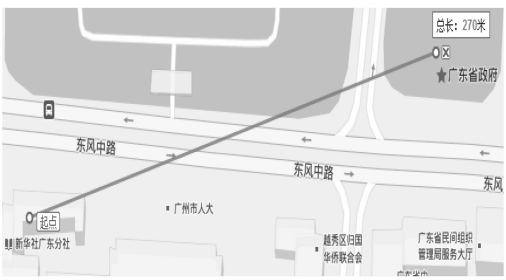

图 1 对象间的欧式距离

Fig. 1 The Euclidean distance between objects

图 2 对象间的网络距离

Fig. 2 The network distance between objects

Yiu 等[8]首次提出在空间网络中对象聚类的问题，并提出相应的解决方案来扩展存在的聚类方法，将其运用到空间网络中的对象上，但算法复杂，实用性不高。唐良[9]在其博士论文上也曾提过空间网络的聚类分析算法，但并没有考虑到空间上如何限制聚类的扩展，适用性不强。本文针对道路网络的特点，建立了实用性和适用性都较高的基于路网拓扑的聚类分析算法模型。并利用 C#语言及 ArcEngine 二次开发组件实现算法的可视化，应用于广州珠江新城，研究居民出行的可达性。

1 基本模型和问题定义

本节引入道路网络以及对象之间的网络距离等定义[9]，然后根据道路网络的聚类特点，给出了对象与聚类的网络距离定义。

定义 1 道路网络与兴趣点。

为了降低道路网络复杂性，可以将路网的交叉口建模为网络图的结点，将交叉口之间的弧段建模为网络图的边[10]。一个道路网络可以表示为一个无向带权图 $G = (V, E, W)$，其中：V 是结点的集合；E 是弧段集合；W 为路段对应权值的正实数集合。

确定兴趣点间的路距离，首先要建立兴趣点与路网的关系。

兴趣点 p（$p \in P$）位于弧段（v_i，v_j））上，用一个三元组（v_i，v_j，pos）来表示兴趣点与路网的关系。这里，v_i 和 v_j 为兴趣点所在弧段（v_i，v_j））的起终点（v_i 为起点，v_j 为终点），pos 为兴趣点 p 与结点 v_i 在弧段上的距离。如图 3，$p_1 = (A, B, 30)$，$p_2 = (A, B, 75)$。

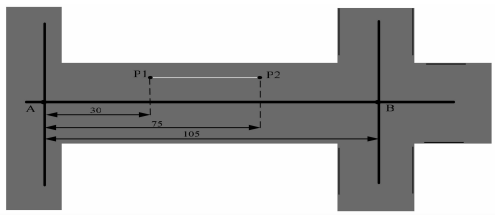

图 3 兴趣点与路网的关系

Fig. 3 The relationship between POI and network

定义 2 对象之间的网络距离 d_D。

1）同一弧段上的对象之间的直接距离：

$$d_D(p,q) = | \text{pos}_1 - \text{pos}_2 \quad (1)$$

2）结点之间的网络距离。d_D（v_i，v_j）为从结点 v_i 到结点 v_j 的最短路径距离。最短路径距离用 Dijkstra 算法计算[11]。

3）不同弧段上的兴趣点之间的网络距离：

$$d_D(p,q) = d_D(p,v_i) + d_D(v_i,v_j) + d_D(v_j,q) \quad (2)$$

定义 3 兴趣点与聚类间的最小距离 d_N。

最小距离 d_N 为兴趣点与聚类不同兴趣点间最小的网络距离。

设有兴趣点 p，聚类 C 的兴趣点为 $q_1, q_2, \cdots, q_m$，则兴趣点 p 与聚类 C_x 的网络距离表示为：

$$d_N(p,C) = \min_{i\in\{1,2,\cdots,m\}} d_D(p,q_m) \quad (3)$$

定义 4　兴趣点与聚类间的最大距离 d_M。

最大距离 d_M 为兴趣点与聚类不同兴趣点间最大的网络距离。

设有兴趣点 p，聚类 C 的兴趣点为 q_1，q_2，…，q_m，则兴趣点 p 与聚类 C 的最大距离表示为：

$$d_M(p,C) = \max_{i\in\{1,2,\cdots,m\}} d_D(p,q_m) \quad (4)$$

2　基于路网拓扑的空间聚类算法模型

2.1　算法介绍

现有的空间聚类算法中的一些概念，如聚类中心、聚类特征等，都难以在道路网络中进行定义[11]。具体来说，给定道路网络中的一组对象，由于可能在网络边上不存在一个唯一点，与聚类中所有对象之间的平均距离最小，所以不能使用网络距离来定义它们的聚类中心。

因此，本文提出一种沿着路网拓扑方向进行的广度搜索，利用凝聚度（α）和扩展度（β）为聚类限制条件的聚类分析算法，α 和 β 的取值直接影响到聚类的结果。其中，定义兴趣点与聚类间的最小距离的限制值为凝聚度 α，α 控制对象之间在网络上的凝聚程度；兴趣点与聚类间的最大距离的限制值为扩展度 β，β 控制整个聚类在网络上的扩展程度。

2.2　算法步骤

基于路网拓扑的空间聚类算法需要一个列表 Q 来保存结点周围待遍历的邻接路段信息，即项（v_1，v_2，w）。这里，w 为弧段（v_1，v_2）的权值。

任取一个兴趣点 p_i（v_a，v_b，pos_i）的聚类分析主要步骤如下：

Step1 创建聚类，设定参数。

创建新聚类 C，将兴趣点 p_i 加入 C。设定限制值 α 和 β，进行 Step2。

Step2 由兴趣点 P_i 沿所在路段的方向扩展搜索至路段结点。

设 p_k 为兴趣点 p_i 在结点 v_b 方向上的邻接兴趣点。

1）若存在兴趣点 p_k，则判断 d_N（p_k，C）$\leqslant\alpha$ 且 d_M（p_k，C）$\leqslant\beta$，符合则将 p_k 加入 C，继续向 v_b 方向扩展，直到结点 v_b，进行 Step3；不符合则直接转到 Step5；

2）若不存在兴趣点 p_k，则直接转到 Step3。

Step3 遍历结点 v_b 的邻接结点，构造列表 Q。

设 v_b 有邻接结点 i 个，遍历 v_b 邻接结点 v_c^1，v_c^2，……，v_c^i，设 p_k 为路段（v_b，v_c^i）上邻接 v_b 的兴趣点。设当前操作结点 v_c^j，初始设定 $j=1$，进行以下循环判断：

1）若边（v_b，v_c^j）上存在兴趣点，则将（v_b，v_c^j，$w(v_b,v_c)$）插入到 Q 队首。$j=j+1$，继续进行循环判断；

2）若边（v_b，v_c^j）上无兴趣点，比较（v_b，v_c）和凝聚度 α 的值：小于凝聚度 α，则将（v_b，v_c，w（v_b，v_c））插入到 Q 队首，大于凝聚度 α，不操作。$j=j+1$，继续进行循环判断。

至到 $j=i$，结束循环判断，进行 Step4。

Step4 取出 Q 队首项，继续进行聚类。

判断队列 Q 是否为空：

1）若 Q 为空，结束整个算法；

2）不为空，取出 Q 队首项。由 v_b 向 v_c 遍历路段　（v_b，v_c）上的兴趣点，将满足聚类限制条件，即兴趣点与聚类间最小距离小于 α，且兴趣点与聚类间最大距离小于 β 的兴趣点加入聚类 C 中。

Step5 扩展邻接结点，遍历该结点的邻接结点，扩充列表 Q。

设路段（v_b，v_c）上邻接 v_b 的兴趣点为 p_t。

1）若 d_D（p_t，v_c）$\leqslant\alpha$，则转到 Step3 对 v_c 的邻接结点进行遍历，扩充优先队列 Q；

2）若 d_D（p_t，v_c）$>\alpha$，则转到 Step4 继续进行聚类。

下面介绍算法流程图，如图 4 所示。其中，遍历 v_b 的邻接结点的算法流程图如图 5 所示。

3　应用示例

根据本文提出的基于路网拓扑的聚类分析算法，以广州市珠江新城路网为实例，利用 C#语言以及 ArcEngine 二次开发组件构建应用分析模块，实现后界面如图 6 所示。

这一节将利用基于路网拓扑的聚类分析算法分析行人的可达性情况。可达性是交通规划中评价交通网络系统性的一项有效的综合性指标[12]。可达性的主要定义的其中一种是，一定空间范围内可获得或接近的目标对象的数量多少。因此，研究的具体的作法是，选择两个地物点，通过聚类分析分别计算一定范围内，行人可获得的公共交通站点的数量。以此比较这两个地物点的可达性。

HCM2000 中规定行人的步行速度取决于行人中老年人的比例，推荐 1.22 m/s 作为标准步行速度值[13]。一般来说，行人步行可接受时间为 5 ~ 10 min，因此取步行适宜距离 L =600 m。

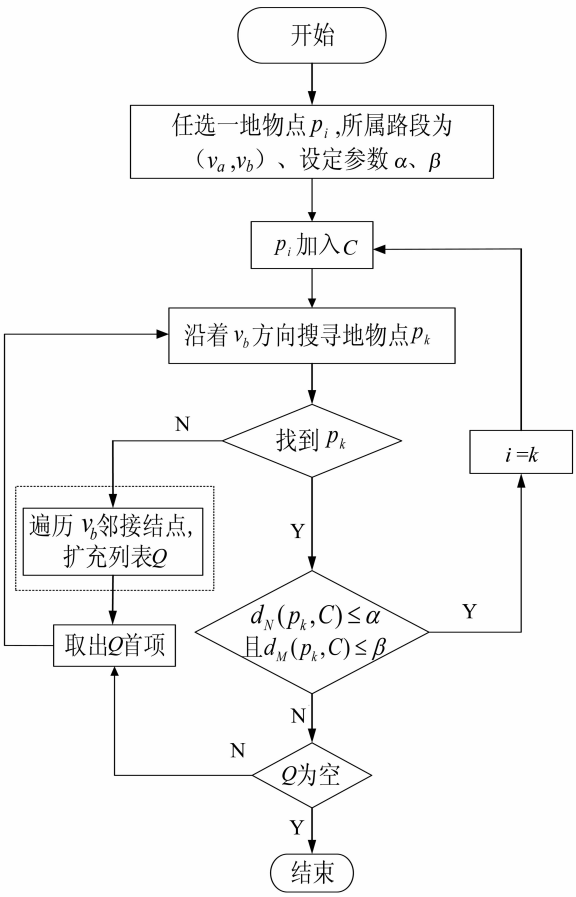

图 4 算法流程图

Fig. 4 Algorithm flow chat

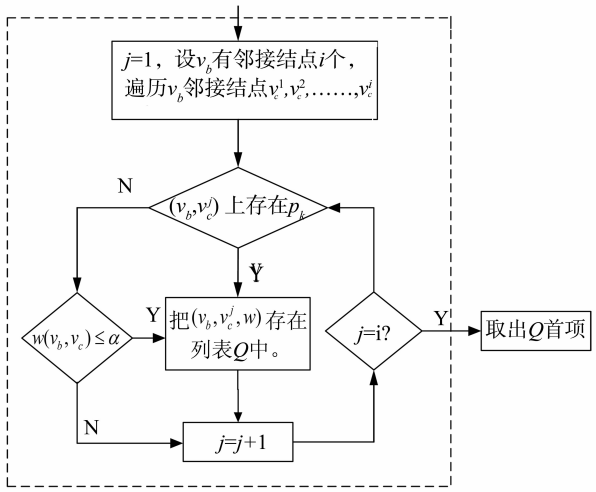

图 5 遍历邻接结点算法流程图

Fig. 5 Algorithm flow chat of extending to the adjacent node

选取珠江新城的广州大剧院和富力盈凯广场，设置限制值$\alpha=600$ m，运行程度得到的结果如图 7 和图 8。

由结果可得，在广州大剧院的 600 m 范围内，行人可到达的公共交通站点共有 3 个，分别是华穗路公交车站、广州大剧院西门公交车站、海心沙公园公交车站。而在富力盈凯广场的 600 m 范围内，行人可到达的公共交通站点共有 4 个，分别是珠江新城地铁站、华穗路口公交车站、友谊国金店公交车站和市政务服务中心公交车站。

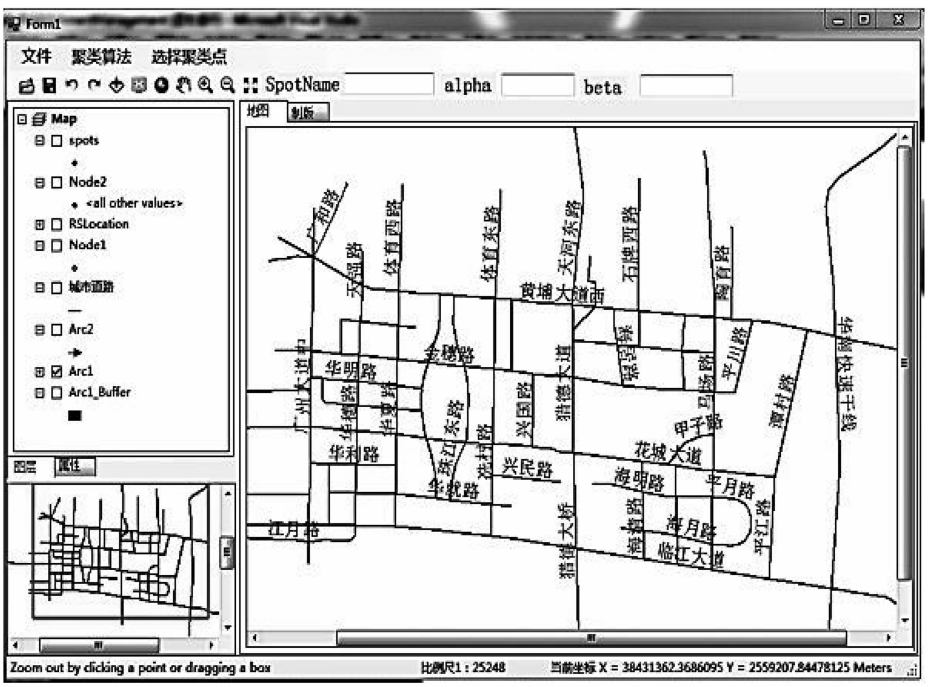

图 6 程序界面

Fig. 6 The program interface

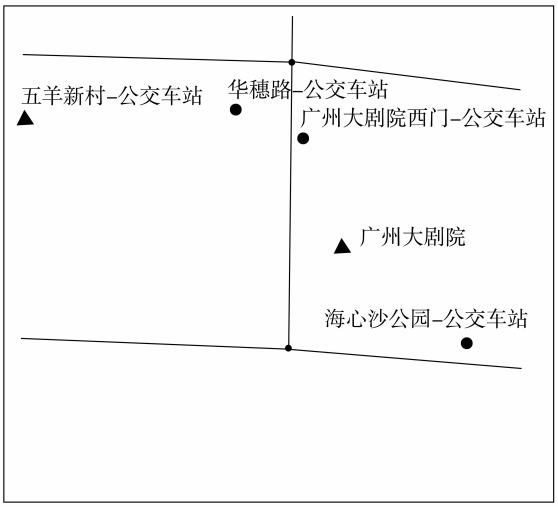

图 7　广州大剧院聚类结果

Fig. 7　The clustering result of Guangzhou Theater

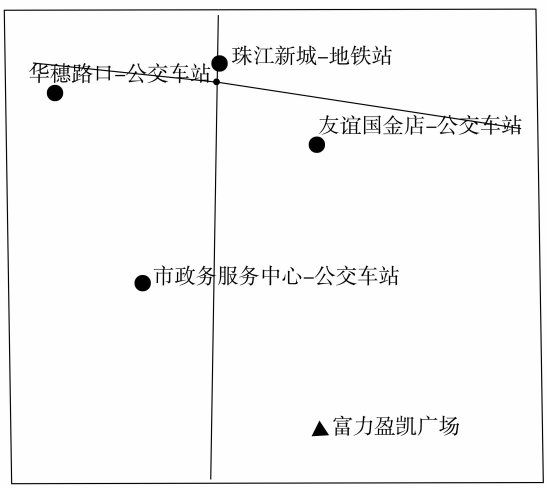

图 8　富力盈凯广场聚类结果

Fig. 8　The clustering result of Decimating Surplus Kay Square

行人在富力盈凯广场 600 m 范围内可获得的公共交通站点比在广州大剧院多。因此，从这个角度出发，600 m 范围内行人在富力盈凯广场的可达性比在广州大剧院强[14]。

4　结　语

基于路网拓扑的空间聚类分析算法，解决了以往大多数空间聚类算法不适用于道路网络的交通问题。其次，利用 C#语言和 ArcGIS，将模型应用在珠江新城路网，通过算法，成功聚类出可用于分析城市居民可达性的区域。这表明了基于路网拓扑的空间聚类分析算法是可行且具有应用价值的。但是，本聚类算法，并没有考虑标志性地物点的具体属性，下一步将考虑地物点的具体属性，进行有选择的聚类。

参考文献：

[1]　许丽利. 聚类分析的算法及应用[D]. 长春：吉林大学，2010.

[2]　刘生鑫. 空间数据聚类分析算法研究及实现[D]. 武汉：中国地质大学，2010.

[3]　李桃映. 交通领域中的聚类分析方法研究[D]. 大连：大连海事大学，2010.

[4]　刘贤腾. 空间可达性研究综述[J]. 城市交通，2007，5(6)：36－43.

[5]　席景科，谭海樵. 空间聚类分析及评价方法[J]. 计算机工程与设计，2009，30(7)：1712－1715.

[6]　JAIN A K，MURTY M N，FLYNN P J. Data clustering：A review[J]. ACM Computing Surveys，1999，31(3)：264－323.

[7]　沙志仁，余志，黄敏，等. 面向指路标志系统的非平面交通路网模型[J]. 测绘科学技术学报，2011，28(6)：442－445.

[8]　YIU M L，MAMOULIS N. Clustering objects on a spatial network[C]. SIGMOD，2004：443－454.

[9]　唐良. 城市道路交通指路标志智能设计系统的研究与实现[D]. 合肥：中国科学技术大学，2008.

[10]　黄敏，李敏，钮中铭，等. 基于指引可达性的指路标志布设优化模型[J]. 中山大学学报：自然科学版，2014，53(3)：19－23.

[11]　张开广，孟红玲，亢金轩. 基于路径的聚类分析[J]. 测绘科学技术学报，2006，23(2)：145－148.

[12]　陈继东，孟小峰，赖彩凤. 基于道路网络的对象聚类[J]. 软件学报，2007，18(2)：332－344.

[13]　Transportation Research Board . Highway capacity manual 2000[M]. Washington D C：National Research Council，2000.

[14]　肖国荣，余志，黄敏. 基于偏离指数的指路标志优化模型研究[J]. 中山大学学报：自然科学版，2008，47(1)：38－41.