

道路网中针对多目标决策的兴趣点高效查询算法

李松, 杨晓龙, 靳海鹏, 张丽平

(哈尔滨理工大学计算机科学与技术学院, 150080, 哈尔滨)

摘要: 为了解决道路网中利用多目标决策技术进行兴趣点推荐和高效位置查询的问题, 针对由于数据规模增加产生大量近似数据, 导致传统多目标决策技术在道路网环境下查询效率和可用性方面较低的问题, 提出了一种道路网广义近似 Skyline 查询算法。首先基于兴趣点的维度相似性和道路网近似性构建近似集和独立点, 并根据兴趣点特性设计相应的剪枝策略; 随后, 通过近似集和独立点重构数据集, 根据剪枝策略过滤掉当查询位置移动时对查询结果无影响的兴趣点, 并构建 AA-R*-Tree 索引以提升查询效率; 最后, 根据兴趣点的近似性提出一种广义近似聚集支配算法, 通过选取代表点代替近似集进行 Skyline 计算, 减少冗余运算并优化查询结果, 最终得到满足兴趣点近似整合有序的 Skyline 结果集。实验结果表明: 所提近似查询算法在大规模数据集和大量相似数据条件下表现出较好的效率与可行性; 与 Higher-Gsky、MG-EGsky 和 GSSK-A 算法相比, 所提算法在数据规模、查询范围及路段数增加时的平均效率提升约 14%, 能够为道路网用户提供更快速有效的决策支持。

关键词: 道路网; Skyline 查询; 多目标决策; 近似查询; 兴趣点推荐

中图分类号: TP311 **文献标志码:** A

DOI: 10.7652/xjtuxb202504014 **文章编号:** 0253-987X(2025)04-0148-10

Efficient Query Algorithm of Points of Interest for Multi-Objective Decision-Making in Road Network

LI Song, YANG Xiaolong, JIN Haipeng, ZHANG Liping

(School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: To address the issue of point-of-interest (POI) recommendation and efficient location queries in road networks using multi-objective decision-making techniques, and to overcome the problem of low query efficiency and usability of traditional multi-objective decision-making techniques in road network environments due to the large amount of approximate data generated by increasing data scales, a generalized approximate skyline query (GAA-SQ) algorithm for road networks is proposed. Approximate sets and independent points are constructed based on POI dimensional similarity and road network proximity, and pruning strategies tailored to POI characteristics are designed. The dataset is then reconstructed using approximate sets and independent points. Based on pruning strategies, POIs that do not affect query results when the query location changes are filtered out, and an AA-R*-Tree index is built to enhance query efficiency. Finally, a generalized approximate aggregation domination algorithm is proposed based on the similarity of POI. By selecting representative points to substitute for the approximate set in Skyline computation, redundant calculations are reduced and query results are optimized. Ultimately, a Skyline result set

is obtained that satisfies the approximate and integrated ordering of POI. Theoretical analysis and experimental results indicate that high efficiency and feasibility are exhibited by the proposed algorithm for large-scale datasets with substantially similar data. Compared with the Higher-Gsky, MG-EGsky, and GSSK-A algorithms, the proposed algorithm achieves an average efficiency improvement of approximately 14% as data size, query range, and road segments increase, providing faster and more effective decision support for road network users.

Keywords: road network; Skyline query; multi-objective decision; approximate query; point of interest recommendation

在道路网信息处理系统中,兴趣点查询是实现地图导航和位置服务的重要基础,其主要目标是通过精准高效的查询机制,帮助用户快速找到与需求相关的地点,例如餐馆、加油站、医院、景点等,为用户提供便捷的出行和生活服务。然而,由于不同人群的自身情况各异,需求的差异性多样性不断增大,多目标决策技术在兴趣点查询中的重要性也不断提升。在兴趣点查询中引入多目标决策技术可以更全面地优化系统表现、提升个性化和用户体验,使得兴趣点查询不仅能提供更精准的查询结果,还能在多维度上更好地服务用户。Skyline 查询^[1]作为多目标决策中的重要技术有助于提高多目标决策的效率。例如,在路网兴趣点查询中,用户不仅关心位置的相关性,还会考虑其他因素(如距离、价格、评分等)。Skyline 查询用于获取数据集合中在不同维度上的最优点,广泛应用于决策支持、兴趣点选取、推荐系统等领域。

近些年,道路网环境下的 Skyline 查询是热点问题之一,Deng 等^[2]首次将 Skyline 查询引入到道路网环境中。Bai 等^[3]提出一种改进的基于位置的道路网 Skyline 查询,指在快速计算路网下的 Skyline 查询。Cai 等^[4]提出了一种速度和方向感知的 Skyline 查询方法,通过考虑用户的移动速度和方向来为用户提供 Skyline 查询区域。Li 等^[5]致力于找到一组天际线凝聚群,其中每个群体在社会凝聚力和空间凝聚力方面都不能被任何其他群体支配。Attique 等^[6]基于空间、文本和社会相关性对数据对象进行排序,在道路网 Skyline 查询问题中引入了地理社交。Liu 等^[7]提出了一种基于 Skyline Cube 的 FHL 索引,利用从低维路径推导高维路径的方法,高效解决了灵活多约束最短路径场景中的查询问题。郝晋瑶等人^[8]提出一种大规模路网图下关键词覆盖最优路径查询算法 KORL,通过路网图划分、近似支配剪枝等策略高效地搜索近似最优解。

随着 Skyline 查询的不断发展,衍生出许多致力于适配用户需求的变种问题,朱睿等^[9]提出高速

流环境下近似连续 k 代表 Skyline 查询算法,给定滑动窗口和连续查询监听窗口中的数据,查询返回窗口中组合支配面积最大的 k 个对象。Yuan 等^[10]和 Bai 等^[11]解决了不完整数据下的 Skyline 查询问题,在多维不完全数据中高效和准确的返回结果。李松等^[12]提出了一种基于范围的障碍空间连续 Skyline 查询算法,用于解决由于查询者位置的移动和空间障碍物的位置变化导致传统多目标决策技术的查询效率较低的问题。Liu 等^[13]介绍了 a-FHL 方法,旨在避免昂贵的天际线路径搜索并加快天际线路径索引的计算。Wang 等^[14]和 Zhang 等^[15]针对 Skyline 查询中存在的隐私泄露问题,提出了保护用户位置隐私和查询结果集隐私的方法。Wang 等^[16]进一步提出了一种基于位置的高效 Skyline 查询 SecSky,目的是在保护数据集隐私的同时降低计算开销。

现实中随着物联网与互联网协调发展,Skyline 查询结果集规模缺乏有效控制,难以提供有效决策支持,因此控制 Skyline 结果集规模也是值得注意的问题。白梅等^[17]提出一种基于最大覆盖的代表 Skyline 查询,选取 k 个最具代表性的 Skyline 查询元组,并通过贪心算法提高效率,但是没有考虑路网空间环境。李松等^[18-19]融合道路网和 K 支配的概念提出了路网 K 支配 Skyline 查询,从用户角度出发精炼 Skyline 结果集并且提高查询效率,但是没有考虑数据点间的关系以及数据点的路网分布。Wang 等^[20]提出了一种名为 TGPE 的有效算法,用以快速计算大规模数据库的前 k 个 G-Skyline 组。Nadouri 等^[21]提出通过使群体 Skyline 支配更为严格,使得一些群体保持不可比性,从而扩展群体 Skyline,但是扩展后的结果集很难符合用户期望。Vlachou 等^[22]提出一种真正平衡多个标准的决定性 Skyline 查询,寻找 Skyline 结果集中尽可能好的数据对象,但是存在效率低的问题。Xia 等^[23]提出一种 ϵ -Skyline 查询方法对 Skyline 结果集进行增大或减小并反映用户偏好,当 ϵ 为正或为负时,

ϵ -Skyline 可以小于或大于传统 Skyline。当 $\epsilon = 0$ 时, ϵ -Skyline 即是传统 Skyline, 该方法没有考虑路网环境且全部数据进行运算存在效率问题。Yin 等^[24]通过选择一个对象代表附近相似的对象, 避免用户重复扫描相似对象。该对象可以来自数据集或结果集, 但是数据集的方法存在效率问题, 而结果集的方法可用性相对较低。

随着数据规模的增长, 传统路网 Skyline 查询的效率和可用性显著降低。一方面, 数据量的增加导致大量路网距离运算, 降低了查询效率; 另一方面, 大量近似数据使得结果集难以支持多准则决策, 用户难以从大量数据中快速选择, 同时还可能遗漏与 Skyline 点仅有微小差异但符合用户兴趣的数据点。现有路网 Skyline 查询较少关注兴趣点的空间分布及维度相似性。例如, 路网上的兴趣点在价格、评分等方面可能差异极小, 但传统算法容易忽略这些细微但重要的差异。用户在到达某个兴趣点后, 通常会便利地访问周边兴趣点, 因此, 通过近似兴趣点的查询, 可以帮助用户快速切换选择, 更好地满足需求, 同时提高结果集的可解释性和实用性。

传统的近似 Skyline 查询通常直接对 Skyline 结果集或对数据集进行近似处理, 不能将扩展的兴趣点进行整合, 并且没有从兴趣点出发考虑兴趣点之间的近似关系。此外, 传统控制 Skyline 结果集规模的查询算法未考虑到在道路网环境下的距离维度, 也未考虑到兴趣点在道路网络环境中的空间分布。因此, 本文基于兴趣点的空间分布以及兴趣点之间的近似关系提出一种道路网广义近似 Skyline 查询算法。

1 基本定义

道路网广义近似 Skyline 查询是路网 Skyline 查询的变体问题, 涉及兴趣点集 P 、道路网、路网近似范围和属性近似范围。本文设定数据集 P 具有 k 个维度, 其中第 k 维为路网空间属性维度, 假定一组 $k-1$ 个非空间属性阈值 $\Delta = \{\lambda_1, \lambda_2, \dots, \lambda_{k-1}\}$ 。其中 $\lambda_i (1 \leq i \leq k-1)$ 为数据点集 P 的第 i 个非空间属性阈值, 假定 ϵ 为第 k 维路网空间属性维度阈值其相关的主要定义如下。

定义 1 非空间属性支配^[25] 设 p_1, p_2 为数据集 P 中的两个兴趣点, 对于任意的 $i, 1 \leq i \leq k-1$ 使得 $p_1(i) \leq p_2(i)$ 且满足 $p_1(j) < p_2(j) (\exists j, 1 \leq j \leq k-1)$ 则称 p_2 非空间支配 p_1 , 记作 $p_2 \prec_{k-1} p_1$ 。

定义 2 道路网支配 给定道路网环境下查询点 q 以

及兴趣点 p_1 和 p_2 , 用 $d(q, p_1)$ 表示 q 与 p_1 之间的路网距离, 若满足 2 个条件其中 1 个时:

$$(1) p_2 \prec_{k-1} p_1, \text{ 且满足 } d(q, p_1) \leq d(q, p_2);$$

$$(2) p_1(i) = p_2(i) (\forall i, 1 \leq i \leq k-1), \text{ 且 } d(q, p_1) < d(q, p_2);$$

则称 p_2 道路网支配 p_1 , 记作 $p_2 \prec p_1$ 。

定义 3 兴趣点非空间相似 给定兴趣点 p_1 和 p_2 以及设定的一组非空间属性阈值 $\Delta = \{\lambda_1, \lambda_2, \dots, \lambda_{k-1}\}$, 若满足 $|p_1(i) - p_2(i)| \leq \lambda_i (\forall i, 1 \leq i \leq k-1)$ 则称兴趣点 p_1 和 p_2 非空间相似, 记作 $\widetilde{p_1 p_2}$ 。

定义 4 兴趣点道路网近似 给定道路网环境下兴趣点 p_1 和 p_2 , $d(p_1, p_2)$ 为 p_1 和 p_2 之间的路网距离, 当且仅当满足下面两个条件时, 称 p_1 和 p_2 道路网近似, 记作 $\widetilde{p_1 p_2}$:

$$(1) \text{ 兴趣点 } p_1 \text{ 和 } p_2 \text{ 非空间相似, 即 } \widetilde{p_1 p_2};$$

$$(2) d(p_1, p_2) \leq \epsilon, \text{ 兴趣点 } p_1 \text{ 和 } p_2 \text{ 在道路网环中的路网距离小于等于距离阈值 } \epsilon.$$

定义 5 近似集与独立点 近似集由彼此非空间相似的点组成, 并且对于 $\forall p \in C, \exists p' \in C$, 存在 $\widetilde{p p'}$, 近似集视作一个整体数据对象不可分割; 那些不存在于任何近似集的兴趣点称为独立点, 记作 $\hat{p}$ 。

定义 6 位置最佳点存在 $p \in C$, 位置最佳点 p 满足 $\min(\max(d(p, p_i)))$, $p_i \in C, p \neq p_i$, p_i 为近似集 C 中任意与 p 不相同的点, 位置最佳点是近似集中距离最远兴趣点最近的兴趣点。近似集中兴趣点共享同一距离维度值, 该距离维度值由位置最佳点的路网维度值确定。

定义 7 近似数据打分函数记作

$$\begin{cases} F(p_i) = \psi_i / \sum_{j=1}^r \psi_j \bar{\omega}_i \prod_{j=1}^{k-1} (p_i(j) / \lambda_j)^{\zeta_j} \\ p_i(j) \neq 0; \lambda_j \neq 0 \end{cases} \quad (1)$$

式中: ψ_i 为近似集内第 i 个兴趣点的历史访客数量; r 为近似集中兴趣点数量; r 为近似集整体历史访客数量; $k-1$ 为非空间属性数; $p_i(j)$ 为第 i 个兴趣点在第 j 个属性上的取值; λ_j 为非空间第 j 个属性维度上的阈值; ζ_j 为第 j 属性的属性优劣值, 当第 j 属性上取值越大优势越大时, ζ_j 取值为 1, 反之 ζ_j 取值为 -1; $\bar{\omega}_i$ 为用户评价指标, 近似集中分数最高的点被称为近似集的代表点。

定义 8 广义近似聚集支配 给定道路网中两个数据对象 o_1 和 o_2 (数据对象可以是近似集或者独立点), 若 o_1 广义近似聚集支配 o_2 , 记作 $o_1 \ll o_2$ 。

(1) 给定兴趣点 $\hat{p}_1$ 和 $\hat{p}_2$ 是数据集 O 中的两个

独立点,若满足 $\hat{p}_1 < \hat{p}_2$,则有 $\hat{p}_1 \ll \hat{p}_2$;反之若满足 $\hat{p}_2 < \hat{p}_1$,则有 $\hat{p}_2 \ll \hat{p}_1$;

(2)给定独立点 $\hat{p}$ 和近似集 C 是数据集 O 中的两个数据对象,兴趣点 p_i 是近似集 C 的代表点,若满足 $p_i < \hat{p}$,则有 $C \ll \hat{p}$;反之,若满足 $\hat{p} < p_i$,则有 $\hat{p} \ll C$;

(3)给定近似集 C 和近似集 C' 是数据集 O 中的两个数据对象,兴趣点 p_i 和 p_j 分别是近似集 C 和近似集 C' 的代表点,若满足 $p_i < p_j$,则有 $C \ll C'$;反之,若满足 $p_j < p_i$,则有 $C' \ll C$ 。

定义 9 道路网广义近似 Skyline 查询 给定一个查询点 q 和一个 k 维兴趣点集 P ,以及道路网路段集合 R 。预先对兴趣点集 P 进行道路网近似计算形成包含近似集和独立点的数据集 O 。道路网广义近似 Skyline 查询结果返回数据集 O 的子集,其中所有独立点或近似集不被数据集 O 中其他独立点或近似集广义近似聚集支配。这些独立点或者近似集被称为道路网广义近似 Skyline 对象,记为 $GAA\text{-}SQ(q,O,R)$,道路网广义近似 Skyline 查询简称为 $GAA\text{-}SQ$ 查询。

道路网广义近似 Skyline 查询首先对道路网环境中存在的大量邻近相似的兴趣点进行聚集,形成近似集和独立点,然后对近似集的独立点进行道路网广义近似聚集支配判断,得到结果集,下面给出具体示例。给定一个查询点 q ,表 1 表示道路网中宾馆数据对象的属性表。其包括宾馆的非空间属性以及与查询点 q 的路网距离属性。表 1 中,给定三维属性兴趣点 $p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8$ 的属性值,分别对应价格和舒适度,第 3 维则为兴趣点到查询点 q 的路网距离属性,表示为 $d(q, p)$ 。

表 1 兴趣点数据对象属性示例

Table 1 Example of attributes of a point of interest object

宾馆对象	宾馆价格/元	舒适度	$d(q, p)$
p_1	187	3	819
p_2	300	4	815
p_3	200	3	834
p_4	190	3	410
p_5	197	4	438
p_6	200	4	441
p_7	210	2	937
p_8	220	3	925

首先进行道路网近似计算,给定非空间阈值 $\Delta = \{20, 1\}$,价格阈值为 20 元,舒适度阈值为 1,路网

距离阈值 $\epsilon = 50$ 。图 1 展示了表 1 中 8 个宾馆之间的相邻情况,其中兴趣点 p_4, p_5, p_6 组成一个近似集 C_1 ,而兴趣点 p_7, p_8 构成了另一个近似集 C_2 。独立点则为 p_1, p_2, p_3 。虽然兴趣点 p_1 与兴趣点 p_4, p_5, p_6 非空间相似,但由于它们之间的道路网距离大于距离阈值,所以 p_1 不属于近似集 C_1 。同样,虽然 p_2, p_3 之间的道路网距离小于距离阈值,但因为它们不满足兴趣点的非空间相似,所以也无法形成近似集。在对道路网近似计算得到的独立点和近似集进行道路网广义近似 Skyline 查询时, p_2, p_3 被 $\{p_4, p_5, p_6\}$ 广义近似聚集支配,而 p_1 广义近似聚集支配 $\{p_7, p_8\}$ 。因此,道路网广义近似 Skyline 查询的结果为 $\{p_1, \{p_4, p_5, p_6\}\}$ 。

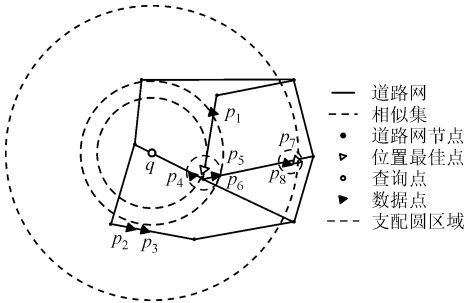


图 1 兴趣点道路网近似示例

Fig.1 Example of a point of interest road network approximation

2 道路网广义近似 Skyline 查询

本文主要分为 2 个主要过程:兴趣点道路网近似计算预处理过程和道路网广义近似 Skyline 查询过程。首先对兴趣点集 P 进行道路网近似计算形成包含近似集和独立点的数据集 O ,然后对数据集 O 进行广义近似聚集 Skyline 查询。

2.1 道路网近似计算

本节首先对数据集进行道路网近似计算,将路网中相邻且相似的兴趣点聚集成近似集,并提出了三种剪枝规则,以提前过滤部分兴趣点。通过选取近似集中位置最佳的兴趣点,并结合近似数据评分函数对其进行排序,实现了 Skyline 查询结果集的多样化。此举有效减少了大量近似点之间的支配判断,从而提升了查询效率。

道路网近似计算主要通过聚集道路网中近邻且相似的兴趣点形成近似集,运用 DBSCAN 算法^[26-27]聚集那些满足道路网近似的兴趣点,并进一步进行非空间相似性判断,筛选出符合条件的兴趣

点构建近似集。近似集的构建不仅多样化了查询结果集,还减少了近似点间的冗余支配判断,同时有效约减了部分兴趣点。

定理 1 已知存在兴趣点 p 属于近似集 C , 如果兴趣点 p 的 ϵ 路网距离内存在其他兴趣点 p' , 兴趣点 p' 满足与近似集 C 中任意兴趣点非空间相似, 则将兴趣点 p' 加入近似集 C 。

证明 已知存在兴趣点 p 属于近似集 C , 兴趣点 p 为近似集 C 中一点, p' 满足与近似集 C 中任意兴趣点非空间相似, 并且兴趣点 p' 与兴趣点 p 之间的路网距离小于距离阈值 ϵ , 所以兴趣点 p' 与兴趣点 p 之间满足兴趣点道路网近似, 所以可以将兴趣点 p' 加入近似集 C , 使得新的集合 C 仍然为近似集。

剪枝规则 1 设兴趣点 p_1 和 p_2 属于兴趣点集 P , 且兴趣点 p_2 不属于任何近似集, 若满足 $d(p_1, p_2) < \epsilon$, $|p_1(i) - p_2(i)| > \lambda_i (\exists i, 1 \leq i \leq k-1)$, 并且存在 p_2 非空间支配 p_1 即 $p_2 \prec_{k-1} p_1$, 则将兴趣点 p_1 剪枝。

证明 设兴趣点集 P 中存在两个兴趣点 p_1 和 p_2 , p_2 不属于任何近似集, 满足 $d(p_1, p_2) < \epsilon$, 路网距离阈值 ϵ 相对于 $d(p_1, p_2)$ 可忽略, 因此, p_1 和 p_2 可以视为位于路网中的同一位置, 因为 $|p_1(i) - p_2(i)| > \lambda_i (\exists i, 1 \leq i \leq k-1)$, p_1 和 p_2 不在同一个近似集中。所以当查询点变化时, 始终有 $p_2 \prec p_1$, 兴趣点 p_1 可被剪枝。

剪枝规则 2 设兴趣点 p 和近似集 C , 兴趣点 p 和近似集 C 中兴趣点均属于兴趣点集 P , 若满足 $d(p, p_j) < \epsilon (\exists p_j \in C)$, 兴趣点 p 与近似集 C 构成的新集合不为近似集, 并且近似集 C 中存在兴趣点 p_i 非空间支配 p 即 $p_i \prec_{k-1} p$, 则将兴趣点 p 剪枝。

证明 因为满足 $d(p, p_j) < \epsilon (\exists p_j \in C)$, 由剪枝规则 1 的证明可知, 兴趣点 p 和近似集 C 可以视为在道路网中的同一位置, 又因为兴趣点 p 与近似集 C 构成的新集合不为近似集, 并且近似集 C 中存在兴趣点 p_i 非空间支配 p 即 $p_i \prec_{k-1} p$, 所以当查询点变化时, 始终有 $C \prec p$, 所以兴趣点 p 可被剪枝。

剪枝规则 3 设兴趣点 p 和近似集 C 中兴趣点均属于兴趣点集 P , 若满足 $d(p, p_j) < \epsilon (\exists p_j \in C)$, 兴趣点 p 与近似集 C 构成的新集合不为近似集, 并且 $p \prec_{k-1} p_i, \forall p_i \in C$, 则将近似集 C 剪枝。

证明 同剪枝规则 2 证明, 兴趣点 p 和近似集 C 可以视为在道路网中的同一位置, 又因为兴趣点 p 与近似集 C 构成的新集合不为近似集, 并且 $p \prec_{k-1} p_i$,

$\forall p_i \in C$, 所以当查询点变化时, 始终有 $p \prec C$, 所以兴趣点 p 可被剪枝。

基于定理 1, 以及剪枝规则 1~3, 本节进一步给出道路网近似计算算法 O-RNAC, 构造可以多样化查询结果的近似集, 减少近似点之间的支配判断, 同时可以约减部分兴趣点, 如算法 1 所示。

算法 1 O-RNAC(P, Λ, ϵ, R)

输入: 兴趣点集 P 、非空间阈值 $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_{k-1}\}$ 、路网距离阈值 ϵ 、道路网路段集合 R ;

输出: 道路网近似计算包含近似集 C 和独立点 $\hat{p}$ 的数据集 O 。

1. $O \leftarrow \emptyset$;
2. for 兴趣点集 P 未被访问的兴趣点 p do
3. 创建近似集 C , 近似集 $C \leftarrow p$;
4. 候选集 $N \leftarrow$ 兴趣点 p 的 ϵ 邻域内未被访问过的兴趣点;
5. for every point $p' \in N$ do
6. if $\widetilde{p'p_i}, \forall p_i \in C$ then
7. $C \leftarrow C + p'$;
8. $N \leftarrow N +$ 兴趣点 p' 的 ϵ 邻域内未被访问过的兴趣点;
9. else if $p_i \prec_{k-1} p', \exists p_i \in C$ then
10. 兴趣点 p' 被剪枝;
11. else if $p' \prec_{k-1} p_i, \forall p_i \in C$ then
12. 候选集 $N \leftarrow$ 兴趣点 p' 的 ϵ 邻域内未被访问过的兴趣点;
13. 近似集 $C \leftarrow p'$;
14. if $C.length > 1$ then
15. $\min(\max(d(p, p_i))), p_i \in C, p \neq p_i$ 计算近似集 C 的位置最佳点;
16. 通过近似数据打分函数对近似集 C 进行排序, 确定近似集代表点;
17. $O \leftarrow O + C$;
18. else if $C.length = 1$ then
19. 标记 C 中唯一点 p 为独立点 $\hat{p}$;
20. $O \leftarrow O + \hat{p}$;
21. $C \leftarrow \emptyset$;
22. return O ;

道路网近似计算 O-RNAC 由算法 1 实现。访问数据集中未被访问过的兴趣点, 将该兴趣点道路网距离阈值 ϵ 领域内存在的其他未被访问的兴趣点加入候选集 N 。遍历候选集 N 中兴趣点, 如果该兴

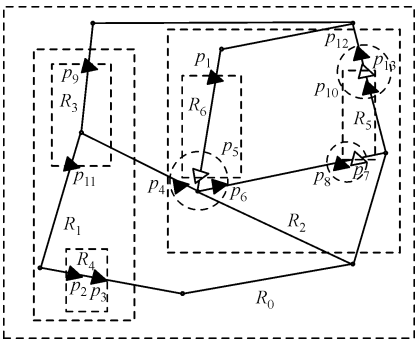
趣点与当前近似集合并后仍为近似集,则将该兴趣点加入当前近似集,并将该兴趣点 ϵ 领域内存在的其他未被访问的兴趣点加入候选集 N ;如果该兴趣点支配当前近似集,并且该兴趣点与当前近似集构成的集合不为近似集,则删除该近似集,将候选集清空,从该兴趣点开始重新创建近似集;如果近似集有且只有一个兴趣点存在,则将该兴趣点标记为独立点,加入结果集;如果近似集存在多于一个的兴趣点,则将近似集加入结果集,并且选取近似集位置最佳点以及对近似集内部兴趣点进行打分排序。最后返回包含近似集和独立点的数据集 O 。

当道路网数据集 P 数据点增加时,对预处理数据集 O 进行更新操作。若此数据点 p 与 ϵ 邻域内的独立点 p' 非空间相似,则将数据点 p 和数据点 p' 形成近似集加入数据集 O ;若此数据点 p 与近似集内所有点非空间相似,且近似集内存在点在数据点 p 的 ϵ 邻域内,则将数据点 p 加入该近似集,更新数据集 O ;若数据点 p 的 ϵ 邻域内不存在与之非空间相似的数据点,则将数据点 p 标记为独立点加入数据集 O 。更新数据集 O 。

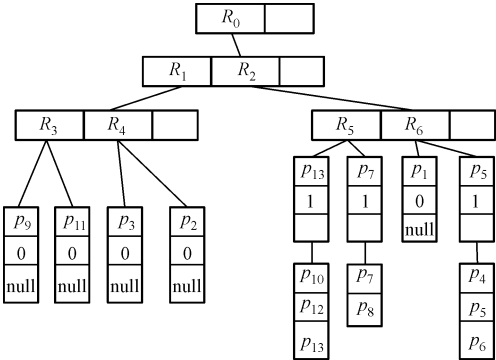
2.2 AA-R*-Tree 索引

传统的 R^* -Tree 无法有效存储路网中存在的大量近似集,且近似集内兴趣点的重复计算也增加了计算开销。为了更高效地存储这些近似集并减少计算开销,本文提出了 AA- R^* -Tree 索引结构。利用近似集位置最佳点替代近似集进行索引构建,可以有效存储近似集结构以及进一步减少计算开销。图 2 展示了所提 AA- R^* -Tree 索引结构的示例。将道路网中的近似集视为一个数据对象,在近似集中选取位置最佳点。用位置最佳点的路网距离维度代替整个近似集中所有其他点的路网维度属性,以近似集位置最佳点代替该近似集进行 R^* -Tree 索引的构建。

AA- R^* -Tree 索引结构的示例如图 2 所示,其中兴趣点 p_5 、 p_7 、 p_{13} 分别是其所在近似集的位置最佳点。在 AA- R^* -Tree 的叶子节点中构建三元组,第 1 位存储兴趣点,第 2 位存储该兴趣点的类型,0 表示该兴趣点为独立点,1 表示该兴趣点为近似集的位置最佳点,第 3 位则存储该点代表的近似集, null 表示空集。在近似集内,兴趣点按评分顺序存储,剪枝操作时将近似集视为一个整体进行处理。同一个近似集中的数据对象的道路网距离属性相同,由近似集位置最佳点的路网属性确定。



(a)索引兴趣点分布图



(b)索引结构图

图 2 AA- R^* -Tree 索引结构示例

Fig. 2 Example of AA- R^* -Tree Index

2.3 道路网广义近似 Skyline 查询

本节通过 2.1 节计算得到的数据集 O 进行广义近似聚集支配计算,将每个近似集视为一个数据对象进行道路网广义近似 Skyline 查询。本节以定理的形式给出广义近似聚集支配中独立点之间、近似集之间以及独立点与近似集之间支配关系的近似简化判断算法,查询返回满足广义近似聚集支配的 Skyline 结果集。

定理 2 给定兴趣点 $\hat{p}_1$ 和 $\hat{p}_2$ 是数据集 O 中的两个独立点,若满足 $\hat{p}_1 < \hat{p}_2$,则有 $\hat{p}_1 \ll \hat{p}_2$;反之若满足 $\hat{p}_2 < \hat{p}_1$,则有 $\hat{p}_2 \ll \hat{p}_1$ 。

证明 由广义近似聚集支配定义可知,独立点之间的支配判断符合传统道路网支配判断,若满足独立点 $\hat{p}_1$ 道路网支配 $\hat{p}_2$,则一定满足独立点 $\hat{p}_1$ 广义近似聚集支配 $\hat{p}_2$,记作 $\hat{p}_1 \ll \hat{p}_2$ 。

定理 3 给定独立点 $\hat{p}$ 和近似集 C 是数据集 O 中的两个数据对象,兴趣点 p 是近似集 C 的代表点,如果 $\hat{p} \sim p$,若 $d(q, \hat{p}) < d(q, p)$,则 $\hat{p} \ll C$,其中 q 为查询点,反之若 $d(q, p) < d(q, \hat{p})$,则 $C \ll \hat{p}$;如果 $\hat{p}$ 和 p 不满足非空间相似,若 $p < \hat{p}$,则有 $C \ll \hat{p}$,反正若 $\hat{p} < p$,则有 $\hat{p} \ll C$ 。

证明 如果 $\hat{p} \sim p$,则 $\hat{p}$ 和 p 在非空间属性下相似相等的,而 $d(q, \hat{p}) < d(q, p)$, $\hat{p}$ 的距离维度占优,所以

$\hat{p} \ll p$ 进而 $\hat{p} \ll C$, 同理当 $d(q, p) < d(q, \hat{p})$, 有 $C \ll \hat{p}$; 如果 $\hat{p}$ 和 p 不满足非空间相似, 若 $p < \hat{p}$, 通过广义近似聚集支配定义可得 $p \ll \hat{p}$, p 为近似集 C 的代表点, 所以可以得出 $C \ll \hat{p}$, 同理若 $\hat{p} < p$, 则有 $\hat{p} \ll C$.

定理 4 给定近似集 C 和近似集 C' 是数据集 O 中的两个数据对象, 兴趣点 p_1 是近似集 C 的代表点, 兴趣点 p_2 是近似集 C' 的代表点, 如果 $\widetilde{p_1 p_2}$, 当 $d(q, p_1) < d(q, p_2)$, 则 $C \ll C'$, 若 $d(q, p_2) < d(q, p_1)$, 则 $C' \ll C$; 当 p_1 与 p_2 不满足非空间相似时, 若 $p_1 < p_2$, 则可以得出 $p_1 \ll p_2$, 进而得出 $C \ll C'$, 反之若 $p_2 < p_1$ 则可以得出 $C' \ll C$.

证明 同定理 3 的证明, 代表点代表了所在的近似集, 当两个代表点非空间相似时, 代表两个近似集非空间相似, 距离查询点越近的近似集占优; 当两个代表点不相似时, 由定理 3 的证明同样可以得出, 若 $p_1 < p_2$, 则 $C \ll C'$, 反之亦然。

考虑道路网环境下兴趣点空间分布情况以及维度相似性的情况下进行 Skyline 查询, 基于定理 2~4 以及 AA-R*-Tree 索引结构, 将近似集视为一个整体数据对象, 对算法 1 计算得到的数据集 O 进行道路网广义近似聚集支配近似查询计算, 快速得到满足数据近似整合有序的 Skyline 结果集, 进一步给出 GAA-SQ 查询算法, 如算法 2 所示。

算法 2 GAA-SQ (q, O, R)

输入: 查询点 q 、包含近似集 C 和独立点 $\hat{p}$ 的数据集 O 、道路网路段集合 R ;

输出: 道路网广义近似 Skyline 查询的结果集 S 。

1. 以 O 中的独立点、近似集、道路网路段创建 AA-R*-Tree 索引结构;
2. $S \leftarrow \emptyset$;
3. 遍历数据集 O 中的数据对象 o_i 进行广义近似聚集 Skyline 支配判断;
4. if $o_i \in O, o_j \in O$ 且 o_i 与 o_j 均为独立点 then
5. 独立点间进行道路网支配判断;
6. if $\hat{p} \in O, C \in O$ 并且 $\hat{p}$ 为独立点, C 为近似集 then
7. if 近似集 C 的代表点 p 与 $\hat{p}$ 非空间相似 then
8. 判断 $d(q, \hat{p})$ 与 $d(q, p)$ 的大小;
9. else
10. 代表点 p 与 $\hat{p}$ 进行道路网支配判断;
11. if $C \in O, C' \in O$ 且 C 与 C' 均为近似集 then
12. if 近似集 C 的代表点 p_1 与近似集 C' 的代表点 p_2 非空间相似 then

13. 判断 $d(q, p_1)$ 与 $d(q, p_2)$ 的大小;
14. else
15. 代表点 p_1 与 p_2 进行道路网支配判断;
16. 更新结果集 S ;
17. return S ;

道路网广义近似 Skyline 查询通过算法 2 实现, 算法 2 通过 2.1 节计算得到的数据集 O 进行道路网广义近似 Skyline 查询。分别进行独立点之间、近似集之间以及独立点与近似集之间支配关系的广义近似支配简化判断算法, 得到结果集 S , 结果集 S 中的数据对象均不被 S 中其他的数据对象广义近似聚集支配, 最终, 结果集 S 就是道路网广义近似 Skyline 查询的结果集。

3 实验比较与分析

本节对本文所提出的 GAA-SQ 算法进行了实验与性能评估。由于道路网广义近似 Skyline 查询是本文研究的一个新问题, 现有的研究成果无法直接应用, 因此为了构造对比算法, 本文对文献[24]中的两个算法进行了适当修改, 使其适用于道路网环境, 并针对给定查询问题, 通过反复调用和寻找相近的数据集进行调整。修改后的算法分别命名为 Higher-Gsky 算法和 MG-EGsky 算法; 此外, 本文还对文献[6]中的 GSSK-A 算法进行了适当修改, 引入近似集的概念, 从而形成与本文提出的 GAA-SQ 算法的对比算法。

3.1 数据集及实验环境设置

本文实验采用真实道路网数据集, 从加利福尼亚道路网数据集以及加利福尼亚的数据点 (<http://www.cs.utah.edu/~lifeifei/SpatialDataset.htm>) 中一片区域抽取 10 000 个节点对象和 5 000 个线段对象(线段作为道路网线段)。查询用户采用随机生成的方式。本文使用道路网近似计算处理数据集中近似兴趣点, 并使用 AA-R*-Tree 索引结构组织处理后的数据集。

实验环境: 8 GB 内存、Core(TM) i7-12700H CPU@2.70 GHz 处理器、500 GB 硬盘, Windows 11 操作系统上用 IntelliJ IDEA 2021.3.2 开发平台 Java 语言实现。实验需要验证 4 个算法的查询性能, 数据集最大维度为 4。每个实验采取单一变量原则, 其余变量为默认值, 实验结果取 50 次实验运行的平均值。

3.2 算法对比

图 3 给出了 4 种算法在不同数据规模下 CPU

运行时间的变化对比情况。CPU 运行时间随着数据规模的增大而不断增加,本文 GAA-SQ 算法的增长趋势比其他 3 个算法小,主要是因为数据规模增大,数据集中近似集的数量和规模也会增大,利用近似集和代表点可以减少大量近似点之间的比较判断,AA-R*-Tree 索引结构可以减少大量近似点间重复的道路网距离计算开销。Higher-Gsky 算法和 MG-EGsky 算法计算得到结果集之后需要不断计算寻找最优替代点,时间消耗较大。

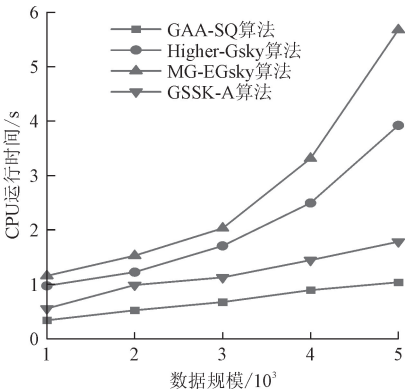


图 3 数据规模对 CPU 运行时间的影响
Fig. 3 Effect of data size on CPU execution time

图 4 展示了不同距离阈值 ϵ 下各算法的性能分析。随着 ϵ 的增大,GAA-SQ 和 GSSK-A 算法的 CPU 运行时间逐渐减少,其中 GSSK-A 算法的减少幅度较为缓慢。Higher-Gsky 和 MG-EGsky 算法的 CPU 运行时间则随着 ϵ 的增大而增加,且 MG-EGsky 算法的增长幅度较为缓慢。这是因为, ϵ 的增大使得近似集规模扩大,从而减少了大量近似点之间的比较判断。然而,随着 ϵ 的增加,MG-EGsky 算法在计算最优替代点时的计算量显著增加,导致时间消耗逐步上升。相比之下,Higher-Gsky 算法从结果集中直接选取

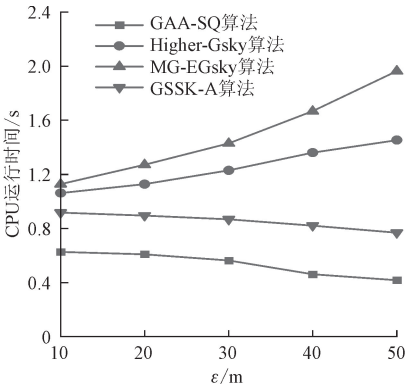


图 4 距离阈值 ϵ 对 CPU 运行时间的影响
Fig. 4 Effect of distance threshold ϵ on CPU execution time

最优替代点,由于距离阈值 ϵ 的增加对候选集规模的影响较小,因此其时间消耗呈现出缓慢增长的趋势。

图 5 展示了 4 种不同算法在不同道路网路段数的条件下 CPU 运行时间的变化。从图中可见,随着道路网数量的不断增加,4 个算法的 CPU 运行时间都有不同程度的增加。其中,GAA-SQ 算法通过近似集和 AA-R*-Tree 索引结构可以减少大量近似集中兴趣点的道路网距离计算,使得在道路网路段数不断增加时,CPU 运行时间可以缓慢增加。GSSK-A 算法引入了近似集的概念,也可以一定程度上减慢 CPU 运行时间的增加幅度,Higher-Gsky 算法和 MG-EGsky 算法随着道路网路段数的增加,计算结果集与结果集最优替代点的开销会不断加大。

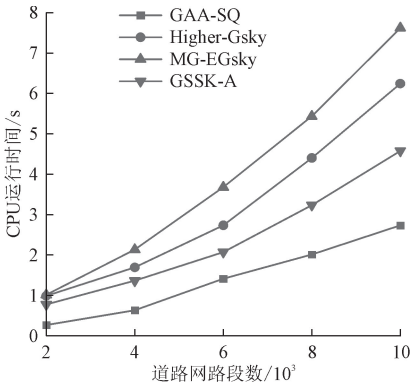


图 5 道路网路段数对 CPU 运行时间的影响
Fig. 5 Effect of road network line segment number on CPU execution time

如图 6 所示,分析算法用于推荐系统的准确率与距离阈值 ϵ 之间的关系,可以更好地说明算法的有效性。GSSK-A 算法引入近似集的概念,其结果与本文算法类似,因此在此不作进一步讨论。从图 6 可以发现,本文所提 GAA-SQ 算法在推荐准确

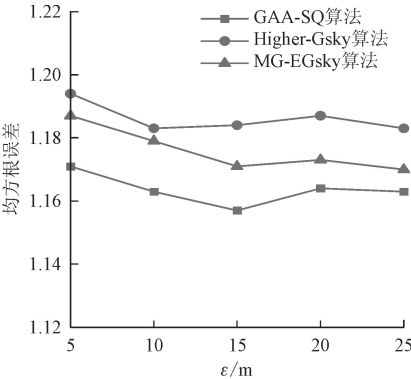


图 6 距离阈值 ϵ 对算法推荐效果的影响
Fig. 6 Influence of distance threshold d_e on algorithm recommendation effect

率上相较于 Higher-Gsky 算法和 MG-EGsky 算法具有显著优势,当距离阈值 $\epsilon \in [10, 20]$ 时推荐效果较好,这表明距离阈值 ϵ 并非越大越好,也说明近似集的范围并非越大越好。随着近似集范围的扩大,距离维度的影响变得难以忽略,从而导致算法准确率下降。

4 结束语

在路网兴趣点查询中引入多目标决策技术可以更全面地优化系统表现、提升个性化和用户体验。随着地图导航和位置服务的发展,路网近似 Skyline 查询具有更重要的价值。已有的 Skyline 查询无法直接处理道路网环境下的近似 Skyline 查询问题,针对已有的不足,本文提出了道路网广义近似 Skyline 查询算法。所提算法根据路网兴趣点集的特征构建近似集,整合数据集并提高查询效率,提高 Skyline 查询的可用性。本文的研究成果增强了兴趣点查询和基于位置服务的技术基础,能有效支持路网下基于位置的兴趣点查询推荐。

今后的研究重点主要集中在近似集规模约减控制上,使得道路网环境下的近似 Skyline 查询输出结果动态可控。

参考文献:

- [1] WU Dingming, ZHANG Zhaofen, JENSEN C S, et al. Efficient skyline keyword-based tree retrieval on attributed graphs [J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(11): 6056-6070.
- [2] DENG Ke, ZHOU Xiaofang, SHEN Hengtao. Multi-source skyline query processing in road networks [C]//2007 IEEE 23rd International Conference on Data Engineering. Piscataway, NJ, USA: IEEE, 2007: 796-805.
- [3] BAI Mei, WANG Qibo, CHANG Shihan, et al. Location-based skyline query processing technology in road networks [J]. The Journal of Supercomputing, 2024, 80(3): 3183-3211.
- [4] CAI Zhi, CUI Xuerui, SU Xing, et al. Speed and direction aware skyline query for moving objects [J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(4): 3000-3011.
- [5] LI Qiyang, ZHU Yuanyuan, YE Junhao, et al. Skyline group queries in large road-social networks revisited [J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(3): 3115-3129.
- [6] ATTIQUE M, AFZAL M, ALI F, et al. Geo-social top- k and skyline keyword queries on road networks [J]. Sensors, 2020, 20(3): 798.
- [7] LIU Ziyi, LI Lei, ZHANG Mengxuan, et al. FHL-cube: multi-constraint shortest path querying with flexible combination of constraints [J]. Proceedings of the VLDB Endowment, 2022, 15(11): 3112-3125.
- [8] 郝晋瑶, 牛保宁, 康家兴. 大规模路网图下关键词覆盖最优路径查询优化 [J]. 软件学报, 2020, 31(8): 2543-2556.
- HAO Jinyao, NIU Baoning, KANG Jiaying. Optimization of keyword-aware optimal route query on large-scale road networks [J]. Journal of Software, 2020, 31(8): 2543-2556.
- [9] 朱睿, 宋杲尧, 王斌, 等. 高速流环境下近似连续 k 代表轮廓查询算法 [J]. 软件学报, 2023, 34(3): 1425-1450.
- ZHU Rui, SONG Fuyao, WANG Bin, et al. Approximate continuous k representative skyline query algorithm over high-speed streaming data environment [J]. Journal of Software, 2023, 34(3): 1425-1450.
- [10] YUAN Dengke, ZHANG Liping, LI Song, et al. Skyline query under multidimensional incomplete data based on classification tree [J]. Journal of Big Data, 2024, 11(1): 72.
- [11] BAI Mei, HAN Yuxue, YIN Peng, et al. S_IDS: an efficient skyline query algorithm over incomplete data streams [J]. Data & Knowledge Engineering, 2024, 149: 102258.
- [12] 李松, 王冠群, 郝晓红, 等. 面向推荐系统的多目标决策优化算法 [J]. 西安交通大学学报, 2022, 56(8): 104-112.
- LI Song, WANG Guanqun, HAO Xiaohong, et al. A multi-objective decision optimization algorithm for recommendation system [J]. Journal of Xi'an Jiaotong University, 2022, 56(8): 104-112.
- [13] LIU Ziyi, LI Lei, ZHANG Mengxuan, et al. Approximate skyline index for constrained shortest pathfinding with theoretical guarantee [C]//2024 IEEE 40th International Conference on Data Engineering (ICDE). Piscataway, NJ, USA: IEEE, 2024: 4222-4235.
- [14] WANG Zuan, DING Xiaofeng, JIN Hai, et al. Efficient secure and verifiable location-based skyline queries over encrypted data [J]. Proceedings of the VLDB Endowment, 2022, 15(9): 1822-1834.
- [15] ZHANG Guopeng, CHEN Xuebin, JIA Yuanli, et al. Support personalized weighted local differential privacy skyline query [J]. Security and Communication Networks, 2022, 2022(1): 9075470.

[16] WANG Zuan, DING Xiaofeng, LU Junfeng, et al. Efficient location-based skyline queries with secure R-tree over encrypted data [J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35 (10): 10436-10450.

[17] 白梅, 王习特, 李冠宇, 等. 基于最大覆盖的代表 Skyline 问题的优化算法研究 [J]. 计算机学报, 2020, 43(12): 2276-2297.

BAI Mei, WANG Xite, LI Guanyu, et al. Research on optimization algorithms of k-maximum coverage skyline queries [J]. Chinese Journal of Computers, 2020, 43(12): 2276-2297.

[18] 李松, 窦雅男, 郝晓红, 等. 道路网环境下 K-支配空间 Skyline 查询方法 [J]. 计算机研究与发展, 2020, 57(1): 227-239.

LI Song, DOU Yanan, HAO Xiaohong, et al. The method of the K-dominant space skyline query in road network [J]. Journal of Computer Research and Development, 2020, 57(1): 227-239.

[19] 李松, 宾婷亮, 郝晓红, 等. 道路网多用户偏好 Top-k 天际线查询方法 [J]. 计算机研究与发展, 2023, 60(10): 2348-2358.

LI Song, BIN Tingliang, HAO Xiaohong, et al. Multi-User preference top-k skyline query method based on road network [J]. Journal of Computer Research and Development, 2023, 60(10): 2348-2358.

[20] WANG Kangao, HAN Xixian, WAN Xiaolong, et al. Efficient computation of top-k G-skyline groups on large-scale database [J]. Information Sciences, 2024, 670: 120614.

[21] NADOURI S, HADJALI A, SAHNOUN Z. RG-SKY: a fuzzy group skyline relaxation for combinatorial decision making [J]. Computer Science and Information Systems, 2022, 19(2): 887-912.

[22] VLACHOU A, DOULKERIDIS C, ROCHA-JUNIOR J B, et al. Decisive skyline queries for truly balancing multiple criteria [J]. Data & Knowledge Engineering, 2023, 147: 102206.

[23] XIA Tian, ZHANG Donghui, TAO Yufei. On skylining with flexible dominance relation [C]//2008 IEEE 24th International Conference on Data Engineering. Piscataway, NJ, USA: IEEE, 2008: 1397-1399.

[24] YIN Bo, WEI Xuetao, LIU Yonghe. Finding the informative and concise set through approximate skyline queries [J]. Expert Systems with Applications, 2019, 119: 289-310.

[25] AWASTHI A, BHATTACHARYA A, GUPTA S, et al. K-dominant skyline join queries: extending the join paradigm to K-dominant skylines [C]//2017 IEEE 33rd International Conference on Data Engineering (ICDE). Piscataway, NJ, USA: IEEE, 2017: 99-102.

[26] SALTOS R, WEBER R. Generalized black hole clustering algorithm [J]. Pattern Recognition Letters, 2023, 176: 196-201.

[27] HUANG Xiaogang, MA Tiefeng, LIU C, et al. GriT-DBSCAN: a spatial clustering algorithm for very large databases [J]. Pattern Recognition, 2023, 142: 109658.

(编辑 刘杨 陶晴)