doi: 10.12068/j.issn.1005-3026.2020.07.001

空间多样化约束下的移动 k 近邻查询

许鸿斐, 谷 峪, 于 戈
(东北大学 计算机科学与工程学院, 辽宁 沈阳 110169)

摘 要: 考虑为移动中的查询对象连续返回 k 个距离近并且满足空间多样化约束的对象,提出了空间多样化约束下的移动 k 近邻(SDC-MkNN)查询. 在此,满足空间多样化约束代表对象之间的相互距离大于距离阈值. 为了高效处理 SDC-MkNN 查询问题,提出了两种基于安全区域技术的算法. 算法均通过减少重新计算查询结果的次数来提高查询效率. 其中一种为精确算法 EA,可连续返回精确的查询结果;另一种为近似算法 ρAA,可连续返回具有近似率保障的近似查询结果. 采用真实数据集验证了所提出算法的有效性.

关 键 词: 移动 k 近邻查询;空间多样化;安全区域;基于位置的服务;查询算法
中图分类号: TP 311.13 **文献标志码**: A **文章编号**: 1005-3026(2020)07-0913-07

Moving k Nearest Neighbor Query with Spatial Diversity Constraints

XU Hong-fei, GU Yu, YU Ge
(School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China. Corresponding author: YU Ge, E-mail: yuge@mail.neu.edu.cn)

Abstract: A new type of queries, named moving k nearest neighbor query with spatial diversity constraints (SDC-MkNN), was proposed. When the query object is moving, this type of queries can continuously return the k nearest neighbors, and any two of the returned objects are satisfied with the spatial diversity constraints, which means the spatial distance between any two of the returned objects must be larger than the distance threshold. Based on the safe region technique, two algorithms were proposed to increase the query efficiency by reducing the frequency of the recomputation of query results. One is an exact algorithm (EA), which can continuously return exact query results, and the other is an approximate algorithm (ρAA), which can continuously return approximate query results with exact bounds. The proposed algorithms were verified by extensive experiments on a real dataset. The results confirm the superiority of the proposed algorithms over the baseline algorithm.

Key words: moving k nearest neighbor query; spatial diversity; safe region; location-based service; query processing algorithms

随着移动终端设备以及基于位置的服务(location-based service,LBS)的快速发展,移动查询作为 LBS 中的一个重要类别,近些年得到了广泛关注. 其中,移动 k 近邻(moving k nearest neighbor,MkNN)查询是移动查询领域的当前研究热点[1-2]. 给定一个移动对象 q 以及一个静态数据对象集合 P,当 q 移动时,MkNN 查询连续返回 q 的 k 个最近邻对象. 该查询具有广泛的应用,例如:为城市中漫步的游客连续推荐 k 个最近邻的景点或餐馆;为正在高速公路行驶的司机连续找到 k 个最近的加油站.

然而,通常情况下 MkNN 查询返回的 k 个对象间的距离较近,空间位置形成聚集且拥有相似的周边环境,这样的结果有时并不能让用户满意.

收稿日期: 2019-06-01
基金项目: 国家重点研发计划项目(2018YFB1003404);国家自然科学基金资助项目(61872070,61602103);中央高校基本科研业务费专项资金资助项目(N171605001).
作者简介: 许鸿斐(1987-),男,山西太原人,东北大学博士研究生;谷 峪(1981-),男,辽宁鞍山人,东北大学教授,博士生导师;于 戈(1962-),男,辽宁大连人,东北大学教授,博士生导师.

例如,查询结果均位于火车站附近,那里通常交通拥堵,停车困难. 为了增强查询结果的可用性,避免产生空间聚集的结果,查询结果空间多样化是一种常用的手段[3-4]. 空间多样化技术使得查询结果在满足用户需求的同时,相互之间距离较远. 正如文献[3]中描述,进行空间查询的用户更加偏好距离查询位置近,并且空间多样化的结果. 用户可以根据对象的周边环境(比如是否有步行可达的商场或体育馆)或对区域的偏好(比如旅行者喜好探索未曾去过的区域)来进行选择.

根据以上应用需求,本文提出了空间多样化约束下的移动 k 近邻(moving k nearest neighbor query with spatial diversity constraints, SDC – MkNN)查询. 当查询对象移动时,连续返回距离查询对象最近的 k 个对象,并且要求结果对象间相互距离不小于距离阈值. 在此,本文使用欧氏距离来度量对象之间的相互距离,这样可以充分体现对象间的空间位置关系. 例如,在为旅行者推荐景点时,如果使用 MkNN 查询,返回的结果由于位置接近,可能已被全部访问. 此时,用户更倾向于使用 SDC – MkNN 查询为其推荐彼此相距较远的结果.

对于 SDC – MkNN 查询,一种简单的解决方法是在查询对象移动的每个时间戳,都将其作为静态查询来处理. 然而,求解空间约束下的 k 近邻问题为 NP – hard 问题,频繁地重新计算查询结果会花费大量的运行时间. 因此,如何有效地减少重复计算次数,提高查询效率是本文的挑战. 安全区域技术[1]是解决移动查询问题的常用方法,但现有技术并未考虑结果对象之间的距离约束关系.

针对以上问题,本文首先提出了支配区域的概念. 当查询对象在某个区域内移动时,如果对象集 S_i 始终优于对象集 S_j,则称该区域为 S_i 对于 S_j 的支配区域. 基于这个概念,计算出了安全区域 R_e,并且提出了一种精确算法. 为了进一步提高查询效率,仅用最优结果集构建出近似安全区域 R_ρ,并且提出了一种近似算法. 当查询对象在 R_ρ 中移动时,查询结果保持不变,并且该集合始终为当前位置最优结果集的 ρ 近似解. 从实验结果看出,近似算法进一步降低了重新计算查询结果的次数.

1 相关工作

近些年,MkNN 查询被广泛研究. 先前工作通常采用基于 Voronoi 图的方法来构建安全区域. 一个 k 阶 Voronoi 图由若干个区域组成,只要查询对象位于同一个区域中,查询结果始终为形成该区域的 k 个对象. 文献[1]提出的 V* – Diagram 方法能够快速计算出一个近似 k 阶 Voronoi 图的区域. 文献[2]提出了 influential neighbor set 方法,通过使用 safe guarding 对象维护当前 kNN. 以上方法均未考虑对象间的距离关系. 文献[5]提出了移动 k 多样化近邻查询问题,与本文查询目标类似,但其采用双目标评价函数(相关性与多样性之和)来评价对象集,通过参数来调节两者之间的比重,返回具有最大目标函数值的对象集. 然而,对于查询用户来说参数调节困难,导致查询结果可能相关性较高且仍空间聚集. 本文采用更加简洁的度量模型(距离和),并且要求查询结果相互之间满足距离约束,从而有效避免了空间聚集现象的产生.

查询结果多样化被广泛运用于时空查询领域. 文献[6]提出了 k 最近多样化邻居(kNDN)问题,找到满足多样化约束的 k 个近邻对象,使用 Gower Coefficient 函数来度量对象之间的多样性,并且提出了启发式搜索算法来解决该问题. 文献[7]研究 k 多样化近邻查询问题. 在 Hamming 空间内,根据查询对象位置以及查询半径,找到一个由 k 个对象组成的,并且多样化程度最大的集合. 集合的多样化程度由其中任意两个对象之间的距离的最小值来度量(值越大,多样化程度越高). 文献[8-9]研究角度多样性,对象之间的多样化程度是通过目标对象相对于查询对象的方向之间的差异来定义的. 文献[8]的目标是为多边形的查询对象从不同的角度找到近邻对象,查询结果能够对查询对象周围形成良好的覆盖. 文献[9]的目标是发现最优的 k 个对象能够最大化相关性的同时最小化两两之间的角度相似性. 文献[4]提出了路网中的多样化空间关键字查询问题,目标在于寻找 k 个对象能够最大化一个由相关性与多样性组成的线性方程. 多样性由对象之间的路网距离来定义. 并且提出了一种基于签名的倒排索引技术,结合基于关键字的剪枝技术来减小搜索空间. 文献[10]基于线性 skylines 解决多样化 k 近邻问题提出了有效算法,能够快速找到在邻近性与多样性方面均不被占有的结果集. 文献[11]提出了具有多样性的 top – k 最短路径(KSPD)问题. 给定查询起点与终点,KSPD 找到长度总和最小的 k 条最短路径,并且任意两条路径的相似性低于阈值约束. 以上工作均为静态查询,本文研究目标为移动状态下的空间多样化 k

近邻查询,因此,以上方法均不适用.

空间偏好查询是空间查询的一种重要类别,通过考虑目标对象的空间位置,以及目标对象周围的特征对象属性对该对象进行评分,返回前 k 个目标对象. 文献[12]给出了 top－k 空间偏好查询的定义,并且使用两类空间约束(即范围约束与最近邻约束)对目标对象进行评分. 文献[13]提出位置敏感的组偏好查询,该查询为一组用户返回一个最优空间对象,并且通过两个方面来评价对象:①所有用户到该对象的距离和;②该对象到满足所有用户偏好的最近邻点的距离和. 文献[14]提出了空间组偏好查询,该查询为一组用户返回 top－k 对象,提出了 satisfaction degree 模型,用于对候选对象进行评价. 该模型不仅考虑了候选对象与查询对象间的距离因素,还将候选对象在社交网络中的评分考虑进来. 空间偏好查询目的在于返回评分较高的目标对象,希望目标对象周围能够最大程度地满足用户偏好,但并未考虑返回的 k 个对象之间的空间距离关系.

2　问题定义

本文假设欧式空间中包含一个静态二维数据对象集合 P. 对于 $\forall p_i,p_j\in P$,使用 $d(p_i,p_j)$ 代表 p_i 与 p_j 之间的欧式距离. 空间多样化约束下的 k 近邻查询 q 包括 3 个参数:查询对象位置 λ、查询结果的数量 k 和空间多样化约束 σ,即 $q=\langle\lambda,k,\sigma\rangle$. 下文中 q 也可作为一个移动的查询对象.

定义 1　空间多样化约束下的 k 近邻(SDC－kNN)查询. 给定对象集合 P 及查询 q,SDC－kNN 查询目标为找到 P 中的一个大小为 k 的子集 S,S 中任意两对象之间距离不小于 σ,并且 S 中所有对象到 q 的距离和最小,即

$$\begin{aligned}&|S|=k,\\ &d(p_i,p_j)\geqslant\sigma, \qquad \forall p_i,p_j\in S,\\ &f(q,S)\leqslant f(q,S'), \qquad \forall S'(|S'|=k),\\ &\qquad\qquad \forall p_i,p_j\in S',d(p_i,p_j)\geqslant\sigma.\end{aligned}$$

式中:$|S|$表示 S 中包含的对象个数;S'代表任意一个大小为 k 且其中任意两对象之间距离不小于 σ 的对象集;函数 $f(q,S)$代表 q 与 S 中所有对象的距离和,即

$$f(q,S)=\sum_{p_i\in S}d(q,p_i).$$

定理 1　SDC－kNN 查询问题是 NP－hard 问题.

证明　SDC－kNN 查询问题可以映射为最大独立集(MIS)问题. MIS 的目标为发现图中的最大节点集,并且要求节点集中任意两个节点之间没有边相连. MIS 问题已被证明为 NP－hard 问题.

首先构建一个图 G. 集合 P 中的每个对象映射为 G 中的一个节点,如果两个对象之间距离小于 σ,那么就在其对应的节点之间加入一条边. 给定一个 SDC－kNN 查询 q,假设所有节点到查询点的距离都相同且 $k=|P|$,那么查询 q 等同于在 G 中寻找最大独立集. 因此,SDC－kNN 查询问题为 NP－hard 问题. 证毕.

当查询 q 移动时,即 q 在不同的时间可能位于不同的地点,该查询变成了空间多样化约束下的移动 k 近邻查询.

定义 2　空间多样化约束下的移动 k 近邻(SDC－MkNN)查询. 给定对象集合 P 以及移动查询 $q=\langle\lambda,k,\sigma\rangle$,当查询对象移动到 λ' 时,SDC－MkNN 查询连续返回 SDC－kNN 查询 $q'=\langle\lambda',k,\sigma\rangle$ 的结果.

当一个 SDC－MkNN 查询到来时,首先将其当作静态 SDC－kNN 查询来处理. 由定理 1 可知,静态 SDC－kNN 查询为 NP－hard 问题,如果在每个时间戳都根据查询对象的当前位置重新计算,这将花费大量的代价. 因此,在第 3 节展示了基于安全区域技术的算法,有效减少重复计算次数,提高查询效率.

3　基于安全区域技术的算法

解决静态 SDC－kNN 查询问题可以通过列举集合 P 中所有大小为 k 的对象集,找到满足多样化约束的最优结果集,时间复杂度为 $O(|P|^k)$. 因此本文重点在于计算当前结果集的安全区域.

3.1　精确算法

首先引入支配区域的概念,基于这个概念使用函数 $f(\cdot)$ 值最优的两个对象集(top－2)计算出一个安全区域. 当查询对象在该区域内移动时,查询结果保持不变.

1) 支配区域. 给定查询 q 以及两个对象集 S_i 与 S_j,如果 q 在某个区域内移动时,S_i 始终优于 S_j,称该区域为 S_i 对于 S_j 的支配区域.

定义 3　支配区域. 给定一个 SDC－MkNN 查询 $q,S_i,S_j\subseteq P$,为满足多样化约束、大小为 k 的两个集合,且满足 $f(q,S_i)<f(q,S_j)$. S_i 对于 S_j 的支配区域表示为 $D(S_i,S_j)$. 当 q 移动到 q' 时,

如果 q' 位于 $D(S_i,S_j)$ 中,$f(q',S_i) < f(q',S_j)$ 始终成立.

下面对支配区域 $D(S_i,S_j)$ 进行数学推导. 已知 $f(q',S_i) = \sum_{p_x\in S_i} d(q',p_x)$, $f(q',S_j) = \sum_{p_y\in S_j} d(q',p_y)$.

根据三角不等式,$\forall\ p\in P, d(q,p) - d(q,q') \leqslant d(q',p)$. 用 $d(q,p_y) - d(q,q')$ 替换 $f(q',S_j)$ 中的 $d(q',p_y)$,得

$$f(q',S_j) \geqslant \sum_{p_y\in S_j}\{d(q,p_y) - d(q,q')\}.$$

若要使得 $f(q',S_i) < f(q',S_j)$ 成立,可以让 $f(q',S_i) < \sum_{p_y\in S_j}\{d(q,p_y) - d(q,q')\}$.

于是可得

$$\sum_{p_x\in S_i} d(q',p_x) + k\cdot d(q,q') < \sum_{p_y\in S_j} d(q,p_y). \tag{1}$$

在此,假设 $S_i = \{p_{x1},p_{x2},\cdots,p_{xk}\}$, $S_j = \{p_{y1},p_{y2},\cdots,p_{yk}\}$,且引入变量 $\theta_{i,j} = (f(q,S_j) - f(q,S_i))/k$,并构建以下不等式组:

$$\left.\begin{aligned} &d(q',p_{x1}) + d(q,q') < d(q,p_{x1}) + \theta_{i,j},\\ &d(q',p_{x2}) + d(q,q') < d(q,p_{x2}) + \theta_{i,j},\\ &\quad\vdots \qquad\qquad\quad \vdots\\ &d(q',p_{xk}) + d(q,q') < d(q,p_{xk}) + \theta_{i,j}. \end{aligned}\right\} \tag{2}$$

显然,若不等式组(2)成立,则不等式(1)一定成立. 因此发现,对于不等式

$d(q',p_{xm}) + d(q,q') < d(q,p_{xm}) + \theta_{i,j}, m\in[1,k]$,$q'$ 可能的位置构成了一个椭圆型区域 $E^m_{i,j}$. 该椭圆以 p_{xm} 和 q 为焦点,$d(q,p_{xm}) + \theta_{i,j}$ 为长轴:

$$E^m_{i,j} = \{q' \mid d(q',p_{xm}) + d(q,q') < d(q,p_{xm}) + \theta_{i,j}\}. \tag{3}$$

通过对所有椭圆区域取交集,可以得到 q' 的一个可行区域. 只要 q' 位于该区域中,不等式(1)始终成立. 从而保证 $f(q',S_i) < f(q',S_j)$ 始终成立. 根据定义3可知,该区域是 S_i 对于 S_j 的支配区域,即

$$D(S_i,S_j) = \bigcap_{m\in[1,k]} E^m_{i,j}. \tag{4}$$

2) 安全区域. 给定查询 q,假设 $S_1,S_2,\cdots,S_N$ 为所有满足空间约束的对象集合,且 $f(q,S_1) < f(q,S_2) < f(q,S_3) <,\cdots, < f(q,S_N)$. 根据定义3,可以构建 S_1(当前最优解)的一个安全区域 R_e,即

$$R_e = D(S_1,S_2)\cap D(S_1,S_3)\cap\cdots\cap D(S_1,S_N). \tag{5}$$

通过定理2可以得出 $R_e = D(S_1,S_2)$,即 q 点在支配区域 $D(S_1,S_2)$ 中移动时,最优解 S_1 保持不变.

定理 2 $R_e = D(S_1,S_2)$.

证明 对于支配区域 $D(S_1,S_n)$, $n\in[3,N]$,可以通过构建如下不等式组进行求解. 在此,假设 $S_1 = \{p_{x1},p_{x2},\cdots,p_{xk}\}$, $\theta_{1,n} = (f(q,S_n) - f(q,S_1))/k$.

$$\left\{\begin{aligned} &d(q',p_{x1}) + d(q,q') < d(q,p_{x1}) + \theta_{1,n},\\ &d(q',p_{x2}) + d(q,q') < d(q,p_{x2}) + \theta_{1,n},\\ &\quad\vdots \qquad\qquad\quad \vdots\\ &d(q',p_{xk}) + d(q,q') < d(q,p_{xk}) + \theta_{1,n}. \end{aligned}\right.$$

由于 $f(q,S_2) < f(q,S_n)$,所以 $\theta_{1,2} < \theta_{1,n}$. 根据式(3)可知 $E^m_{1,2}\subseteq E^m_{1,n}$, $m\in[1,k]$. 进一步根据式(4)可以得到 $D(S_1,S_2)\subseteq D(S_1,S_n)$. 因此,$R_e = D(S_1,S_2)$. 证毕.

下面给出 R_e 的一个举例. 图1中存在6个空间二维对象 $o_1,o_2,\cdots,o_6$. 给定查询 $q = \langle(0,0),3,3\rangle$,通过计算可得 $S_1 = \{o_1,o_3,o_5\}$ 与 $S_2 = \{o_1,o_4,o_5\}$,为满足空间多样化约束($\sigma = 3$)且具有最小距离和的 top-2 最优结果集,根据式(3),计算得到如图1所示的3个椭圆形区域 $E^1_{1,2}$, $E^2_{1,2}$, $E^3_{1,2}$. 从而3个椭圆的交集为安全区域 R_e(深色区域)

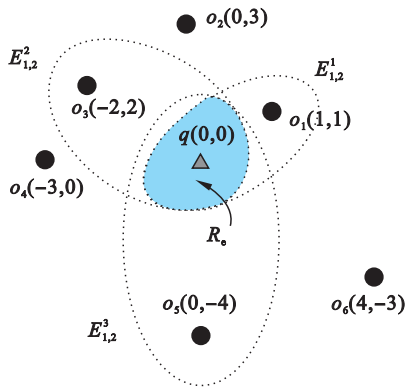

图 1 R_e 的一个举例

Fig. 1 An example of R_e

3) 精确算法. 通过使用安全区域 R_e,提出了一种精确算法(exact algorithm,EA),如表1所示. 当查询到来时,首先求得当前位置的 top-2 查询解 S_1 与 S_2,之后使用函数 $C(\cdot)$ 计算 S_1 的一个安全区域 R_e,并返回 S_1;此后进行查询维护,当查询对象移动到新的位置 q' 时,检查 q' 是否位于 R_e 中,如果成立,则最优解 S_1 将保持不变,否则计算 q' 的 top-2 查询结果集以及新的安全区域 R_e.

4) 复杂度分析. EA 的主要代价在于使用函数 $T(\cdot)$ 计算 top-2 最优结果集,其时间复杂度为 $O(|P|^k)$. 若对象离开安全区域 R_e, $T(\cdot)$ 需要

被调用. 本文用 q_e 表示 R_e 边界点,那么 q 与 q_e 之间的最小距离为 $d_e = \frac{\theta_{1,2}}{2} = \frac{f(q,S_2) - f(q,S_1)}{2k}$. 假设查询对象的移动速度为 v 且沿直线运动,那么函数 $T(\cdot)$ 重复计算的频率为

$$f_r = \frac{v}{d_e} = \frac{v \cdot 2k}{f(q,S_2) - f(q,S_1)}.$$

因此,EA 的时间复杂度为 $O(|P|^k \cdot f_r)$.

表 1　精确算法

Table 1　Exact algorithm

```
输入: SDC - MkNN 查询 q
输出: 最优结果集 S1
1    S1,S2 ← T(·);(计算 top - 2 最优结果)
2    C(S1, S2);
3     返回 S1;
4    While query continues do
5        q moves to q′;
6        If q′ ∈ Re then
7            返回 S1;
8           continue;
9        If q′ ∉ Re then
10           q ← q′,S1,S2 ← T(·);
11           C(S1, S2);
12            返回 S1;
13           continue;
function  C(S1, S2)
1    For  m = 1 to k
2       Re ← Re ∩ E^m_{1,2}
```

3.2　近似算法

为了进一步提高查询效率,提出了一种近似算法. 该算法使用近似安全区域 R_ρ 来减少重复计算频率. 区别于 R_e, R_ρ 仅使用最优结果集 top - 1 来构建. 当查询对象在 R_ρ 内移动时,先前最优结果为当前位置精确结果的 ρ 近似($\rho < 1$,由用户设定).

1) 近似安全区域. 假设 q 的最优结果集为 S,当查询对象移动到 q' 时的最优结果集为 S_{new}. 那么对于 $\forall\ p \in P$,根据三角不等式可得

$$f(q',S_{new}) = \sum_{p_x \in S_{new}} d(q',p_x) \geqslant \sum_{p_x \in S_{new}} \{d(q,p_x) - d(q,q')\} \geqslant f(q,S_{new}) - k \cdot d(q,q').$$

由于 S 为 q 时的最优集,所以 $f(q,S_{new}) \geqslant f(q,S)$ 成立. 因此,

$$f(q',S_{new}) \geqslant f(q,S) - k \cdot d(q,q').$$

若 $\rho \cdot f(q',S) \leqslant f(q,S) - k \cdot d(q,q')$ 成立,则 $\rho \cdot f(q',S) \leqslant f(q',S_{new})$ 成立,即 S 为 S_{new} 的 ρ 近似解. q' 需满足:

$$d(q,q') \leqslant \frac{f(q,S) - \rho \cdot f(q',S)}{k}. \quad (6)$$

由不等式(6) 构建的区域称为近似安全区域 R_ρ.

2) 近似算法. 基于近似安全区域 R_ρ,提出了一种近似算法(ρ-approximate algorithm,ρAA),如表 2 所示. 该算法可连续返回一个 ρ 近似解.

表 2　近似算法 ρAA

Table 2　Approximate algorithm ρAA

```
输入: SDC - MkNN 查询 q
输出: 近似结果集 S
1    S ← T(·); (计算当前最优结果)
2     返回 S;
3    While query continues do
4       q moves to q′;
5       If  q′ ∈ Rρ then
6          返回 S;
7            continue;
8       Else
9          q ← q′,S ← T(·);
10           返回 S;
11           continue;
```

3) 复杂度分析. 在 ρAA 中,同样使用函数 $T(\cdot)$计算当前最优结果. 当 q 移动到 q' 时,若 $d(q,q') > \frac{f(q,S) - \rho \cdot f(q',S)}{k}$,需要重新计算最优解,即调用函数 $T(\cdot)$(复杂度为 $O(|P|^k)$). 假设查询对象速度为 v 且沿直线运动,$T(\cdot)$ 的重复计算频率 $f_r = \frac{vk}{f(q,S) - \rho \cdot f(q',S)}$,那么,$\rho$AA 的复杂度为 $O(|P|^k \cdot f_r)$.

4　实验结果

实验数据集取自文献[15]中所使用的真实数据集,包含 New York 50 334 个 Foursquare 签到地点. 本节对所提出的算法 EA,ρAA 以及一个基于采样的算法 BASE 在查询效率上进行对比. BASE 算法在查询对象移动的每个时间戳都重新计算查询结果. 每组实验随机生成 20 条轨迹,每条轨迹包含 100 个时间戳. 实验结果显示了算法的平均 CPU 时间以及重新计算查询结果的次数. 默认参数为 $k = 3$,多样化距离约束 $\sigma = 2$ km,近似率 $\rho = 0.7$,距离间隔为 100 m(代表连续两个时间戳之间查询对象的移动距离).

所有算法均由 C + + 编程实现. 实验在 PC 机上执行,处理器为 3.40 GHz Intel Core i7 - 6700 CPU,内存 16 GB,64 位 Windows 操作系统.

图 2 显示了当参数 k 变化时,三种算法的运行时间以及重复计算次数的变化. 由于三种算法

均包含静态 SDC – kNN 查询计算,随着 k 的增加计算代价显著增加. 由于BASE算法在每个时间戳都需要计算查询结果,所以计算次数始终为 100. EA 与 ρAA 使用了安全区域技术减少了重复计算次数,运行时间明显好于 BASE. 由于 ρAA 使用参数 ρ 得到了面积较大的安全区域 R_ρ,从而具有最佳性能.

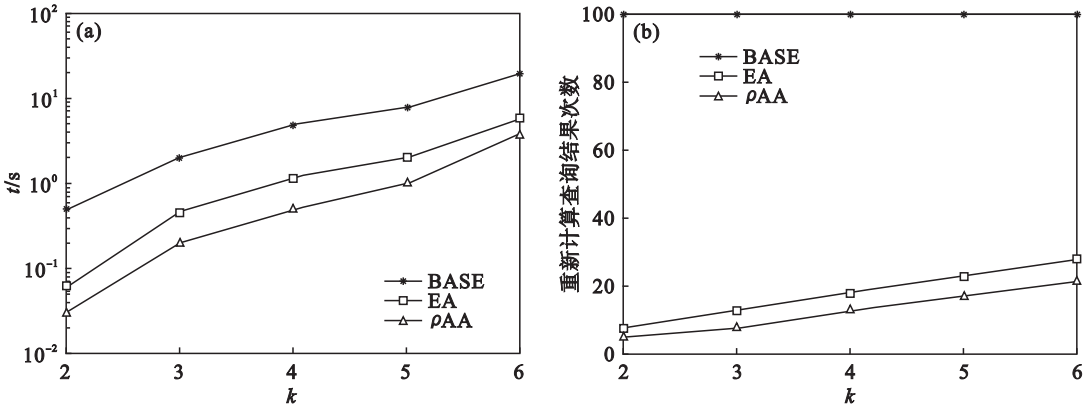

图 2 查询结果数量 k 对三种算法性能的影响

Fig. 2 Effects of k on the performances of the three algorithms

(a)—执行时间;(b)—重复计算次数.

图 3 显示了距离阈值 σ 对三种算法效率的影响. 可以看出,随着 σ 的增加算法执行时间均有所下降. 原因在于 σ 的增加,满足距离阈值约束的对象集数量下降,从而提升了 SDC – kNN 查询的计算效率.

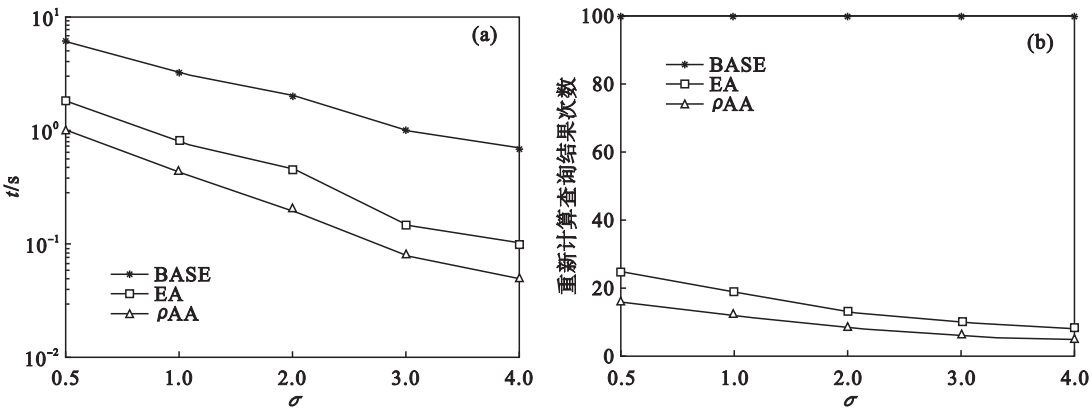

图 3 距离阈值 σ 对三种算法性能的影响

Fig. 3 Effects of σ on the performances of the three algorithms

(a)—执行时间;(b)—重复计算次数.

图 4 显示了不同近似率 ρ 对 ρAA 的影响. 从图 4a,图 4b 中发现,随着 ρ 的增加,ρAA 的运行时间以及重复计算次数都相应增加. 原因在于 ρ 的增加,导致安全区域 R_ρ 面积缩小. 尤其当 $\rho = 1$ 时,ρAA 产生精确结果,但其代价高于 EA. 从图 4c 中可知,在 ρ 变化的所有情况下,ρAA 返回结果的精度(即,S 为算法返回的结果,S^* 为精确结果)均大于 ρ,从而证明了 ρAA 的有效性.

最后,比较了本文提出的算法与文献[5]中 PCPM 算法的性能. 正如在相关工作中所描述的,文献[5]与本文查询目标类似,但其采用双目标评价函数(相关性与多样性之和)来评价对象集,相关性为查询对象与结果对象的距离,多样性为结果对象间的相互距离. PCPM 可以连续返回精确的查询结果. 从图 5 中可以看出,EA 与 ρAA 在执行时间和返回对象间平均距离方面均优于 PCPM 算法. 特别是结果对象间的平均距离,均大于默认约束阈值 2 km,因结果对象间的平均距离越大,空间多样化越好,从而证明了本文提出的算法的有效性.

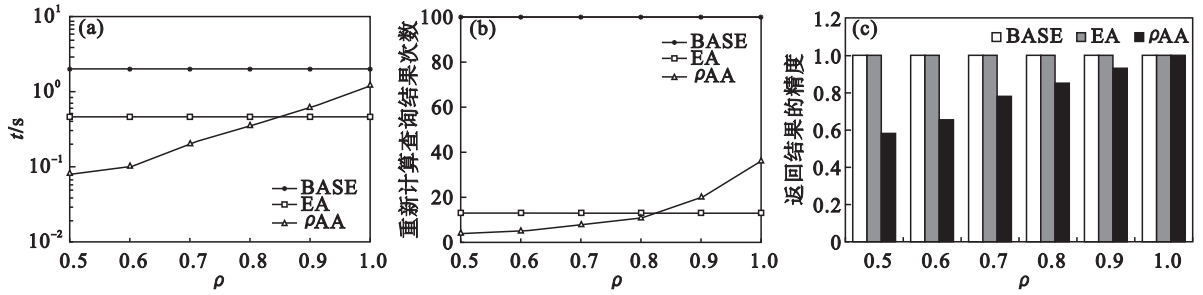

图 4　近似率 ρ 对三种算法性能的影响

Fig. 4　Effects of ρ on the performances of the three algorithms

(a)—执行时间;(b)—重复计算次数;(c)—返回结果的精度.

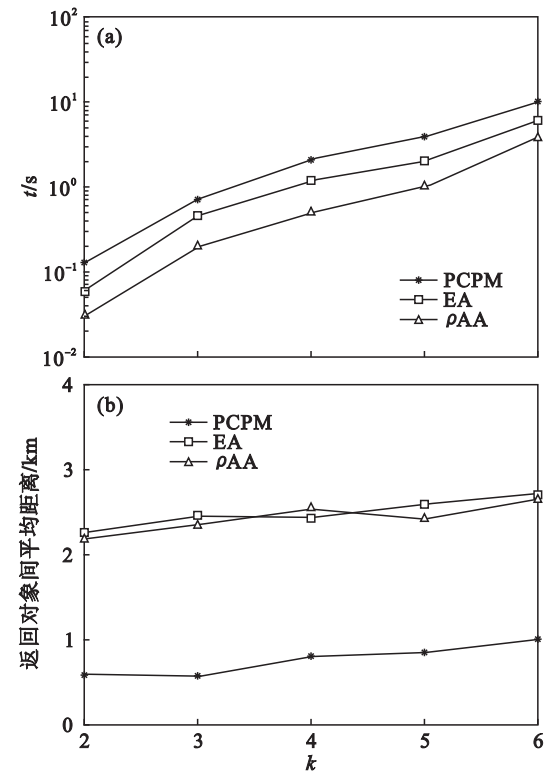

图 5　与 PCPM 算法对比

Fig. 5　Comparison with PCPM algorithm

(a)—执行时间;(b)—返回对象间平均距离.

5　结　　语

本文提出了空间多样化约束下的移动 k 近邻(SDC－MkNN)查询. 该查询可以连续地返回 k 个距离查询对象近但相互之间距离较远的对象. 基于安全区域技术思想,提出了精确算法(EA)以及近似算法(ρAA),通过减少重新计算查询结果的次数提高查询效率. 通过实验证实了两种算法均能够有效地处理 SDC－MkNN 查询. 相较于基本算法,EA 以及 ρAA(当 $\rho=0.7$ 时)可以分别节省 60% 与 80% 以上的查询时间.

参考文献:

[1] Nutanong S, Zhang R, Tanin E, et al. Analysis and evaluation of V*-diagram: an efficient algorithm for moving KNN queries[J]. *The VLDB Journal*, 2010, 19(3): 307－332.

[2] Li C W, Gu Y, Qi J Z, et al. Processing moving KNN queries using influential neighbor sets[J]. *Proceedings of the VLDB Endowment*, 2014, 8(2): 113－124.

[3] Tang J Y, Sanderson M. Spatial diversity, do users appreciate it? [C]//Proceedings of the 6th Workshop on Geographic Information Retrieval. Zurich, 2010: 18－19.

[4] Zhang C Y, Zhang Y, Zhang W J, et al. Diversified spatial keyword search on road networks[C]//Proceedings of the 17th International Conference on Extending Database Technology. Athens, 2014: 367－378.

[5] Gu Y, Liu G L, Qi J Z, et al. The moving k diversified nearest neighbor query[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2016, 28(10): 2778－2792.

[6] Jain A, Sarda P, Haritsa J R. Providing diversity in k-nearest neighbor query results[C]//Proceedings of the 8th Advances in Knowledge Discovery and Data Mining. Sydney, 2004: 404－413.

[7] Abbar S, Amer-Yahia S, Indyk P, et al. Diverse near neighbor problem [C]//Proceedings of the 29th Symposium on Computational Geometry. Rio de Janeiro, 2013: 207－214.

[8] Lee K C, Lee W C, Leong H V. Nearest surrounder queries [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(10): 1444－1458.

[9] Kucuktunc O, Ferhatosmanoglu H. λ-diverse nearest neighbors browsing for multidimensional data [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2013, 25(3): 481－493.

[10] Costa C F, Nascimento M A, Schubert M. Diverse nearest neighbors queries using linear skylines[J]. *GeoInformatica*, 2018, 22(4): 815－844.

[11] Liu H P, Jin C Q, Yang B, et al. Finding top-k shortest paths with diversity [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2018, 30(3): 488－502.

[12] Yiu M L, Dai X Y, Mamoulis N, et al. Top-k spatial preference queries [C]//Proceedings of the 23rd IEEE International Conference on Data Engineering. Istanbul, 2007: 1076－1085.

[13] Li M, Chen L S, Cong G, et al. Efficient processing of location-aware group preference queries[C]//Proceedings of the 25th ACM International Conference on Information and Knowledge Management. Indianapolis, 2016: 559－568.

[14] Zhang Z, Jin P Q, Tian Y, et al. Efficient processing of spatial group preference queries [C]//Proceedings of the 24th International Conference on Database Systems for Advanced Applications. Chiang Mai, 2019: 642－659.

[15] Bao J, Zheng Y, Mokbel M F. Location-based and preference-aware recommendation using sparse geo-social networking data[C]//Proceedings of the 20th International Conference on Advances in Geographic Information Systems. Redondo Beach, 2012: 199－208.