时间区间上的不确定移动对象距离范围查询①

陈逸菲, 王玉芳, 赵丽玲, 陈 慧

(南京信息工程大学 信息与控制学院, 南京 210044)

摘 要: 针对目标对象与查询发出者皆为不确定移动对象的情况, 提出了一种时间区间上的距离范围查询(DRqTI). 此类查询搜索出数据集中在给定时间区间内, 到查询发出者距离不超过阈值的目标对象, 查询结果中包含对象满足查询条件的有效时间段和匹配度. 提出了基于轨迹、基于时间区间和基于距离的三种剪枝策略, 并给出了精炼和匹配度计算方法, 在此基础上设计了查询处理算法. 实验分析表明, 三种剪枝策略中基于距离的方法性能最佳, 提出的算法能有效处理 DRqTI 问题.

关键词: 移动对象; 范围查询; 时间区间; 不确定性

Distance-Based Range Queries over Uncertain Moving Objects within Time Intervals

CHEN Yi-Fei, WANG Yu-Fang, ZHAO Li-Ling, CHEN Hui

(School of Information and Control, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: Aiming at the scenarios that the query issuers and query sender are uncertain moving objects, a new type of queries named DRqTI (Distance-based Range query within a Time Interval) is defined. The DRqTI searches out the target object in the dataset, which satisfies that the distance to the query issuer does not exceed the threshold value in a given time interval. And query results contain the suitability and valid time intervals which satisfy the condition. Three pruning strategies, namely trajectory-based, time-based and distance-based rules are designed. Furthermore, algorithms that integrate the calculation method of refining and matching degree are developed. On this basis, a query processing algorithm is designed. The experimental analysis shows that distance-based method is the best among three methods and the proposed algorithm can deal with DRqTI problem efficiently.

Key words: moving objects; range query; time interval; uncertainty

基于位置的服务(Location-based Services, LBS)在服务业、交通、军事等领域具有广泛应用, 然而 LBS 应用中移动对象定位技术的局限性、连续位置存储的间断性、位置隐私保护需求等因素使得位置信息不可避免的包含了不确定性[1], 对查询处理产生了很大影响[2].

不确定对象的轨迹是三维空间中其可能位置的集合. 假设要找出“在 t_s 至 t_e 时间内, 到车辆 q 距离不超过 R 公里的警车”. q 的位置和目标对象的位置不确定, 则到 q 距离不超过 R 的空间范围也是不确定的. 本文称这类问题为时间区间上的距离范围查询(DRqTI). DRqTI 对 PUDR (Probabilistic Uncertain Distance-based Range)查询[3]在时间维度上进行了扩展, 即从时间点扩展为时间段. FDR(Fuzzy Distance-based Range)查询[4]与 DRqTI 的不同之处是, FDR 查询中距离是个模糊值, 且查询返回的结果中不包含对象关于查询条件的满足程度和具体时间段.

文献[5]和文献[6]提出了不确定移动对象的最近邻查询处理算法, 基于时间段表示查询处理结果. 这种表示方法不适合范围查询: 首先, 范围查询结果中的对象一般远多于最近邻查询, 若采用此表示方法会产生大量时间段, 结果表示分散, 用户使用不便; 此外, 同一对象可能在多个时间段中出现, 每个时间段内都要计算匹配度, 计算量大. 因此本文采用基于对象的形式[7]来表示查询结果. 可见 DRqTI 在语义和查询处理方法上均有别于现有研究.

① 基金项目:国家自然科学基金(41301407)

收稿时间:2016-05-26;收到修改稿时间:2016-07-07 [doi:10.15888/j.cnki.csa.005613]

1　问题的定义

假设移动对象采用 VBP 更新策略[8], 在任意时刻 t, 对象 o 的不确定位置在以期望位置 $o.c(t)$为圆心(下文简记为 $o.c$), 偏差阈值 $o.r$ 为半径的圆内[3,4], 记为 $o.ur(t)$, 称为 o 的不确定域(下文简记为 $o.ur$).

定义 1. 不确定移动对象 o 和不确定查询发出者 q 之间存在最近距离 $n_{qo}(t)$和最远距离 $f_{qo}(t)$, 它们是时间 t 的函数[3,4],

$$n_{qo}(t)=\begin{cases}\text{dist}(q.c,o.c)-q.r-o.r, & q.ur \cap o.ur=\varnothing\\ 0, & q.ur \cap o.ur\neq\varnothing\end{cases} \tag{1}$$

$$f_{qo}(t) = \text{dist}(q.c,o.c) + q.r + o.r \tag{2}$$

其中 dist(,)为两点之间的距离函数.

定义 2. 已知不确定移动对象数据集 D, 不确定查询发出者 q, 查询时间区间$[t_s,t_e]$以及距离阈值 R, 则 DRqTI 的形式化定义为 DRqTI$(q,[t_s,t_e],R)$= $\{(o,[t_{os},t_{oe}],md_o)|o\in D,[t_{os},t_{oe}]\subseteq[t_s,t_e]\wedge md_o>0\}$, 其中,

$$md_o=\frac{\left[f_{qo}(t),n_{qo}(t),R\right]_{t_{os}}^{t_{oe}}}{\left[f_{qo}(t),n_{qo}(t)\right]_{t_{os}}^{t_{oe}}} \tag{3}$$

t_{os} 和 t_{oe} 是在$[t_s,t_e]$内 $n_{qo}(t)\le R$ 的起点和终点. $[]_{t_{os}}^{t_{oe}}$ 表示在$[t_{os},t_{oe}]$内相关曲线包围的面积.

任意时刻对象 o 和 q 之间可能的距离在$[n_{qo}(t),f_{qo}(t)]$内, 因此将在$[t_{os},t_{oe}]$内曲线 $n_{qo}(t)$、$f_{qo}(t)$和水平线 $d(t)=R$ 包围的区域面积与 $n_{qo}(t)$和$f_{qo}(t)$包围的区域面积之比, 定义为 o 在此时间段内关于查询条件的匹配度. 下面进一步说明公式(3)的计算.

定义 3. 如果 $n_{qo}(t)=R$ 有 2 个根 t_{n1}、$t_{n2}(t_{n1}<t_{n2})$, 则称 $t_{n1}\in[t_s,t_e]$为正的外部事件点, $t_{n2}\in[t_s,t_e]$为负的外部事件点, 分别记作 t_{n+}、t_{n-}.

定义 4. 如果 $f_{qo}(t)=R$ 有 2 个根 t_{f1}、$t_{f2}(t_{f1}<t_{f2})$, 则称 $t_{f1}\in[t_s,t_e]$为正的内部事件点, $t_{f2}\in[t_s,t_e]$为负的内部事件点, 分别记作 t_{f+}、t_{f-}.

图 1 中 t_1 和 t_4 分别是对象 a 的正负外部事件点, t_1 是 $n_{qa}(t)$大于 R 和小于 R 的分界点, 表示 a 从不可能满足查询变成可能满足查询; t_4 刚好相反. 可见对象满足查询条件的时间段$[t_{os},t_{oe}]$与外部事件点密切相关. t_2 和 t_3 分别是对象 a 的正负内部事件点, t_2 是 $f_{qa}(t)$大于 R 和小于 R 的分界点, a 到 q 的距离从可能小于 R 变成一定小于 R; t_3 刚好相反. 内部事件点是对象匹配度计算的分界点. 以对象 a 为例, 根据定义 2 有:

$$md_a=\frac{[f_{qa}(t),n_{qa}(t),R]_{t_1}^{t_4}}{[f_{qa}(t),n_{qa}(t)]_{t_1}^{t_4}}$$
$$=\frac{\int_{t_1}^{t_2}(R-n_{qa}(t))\mathrm{d}t+\int_{t_2}^{t_3}(f_{qa}(t)-n_{qa}(t))\mathrm{d}t+\int_{t_3}^{t_4}(R-n_{qa}(t))\mathrm{d}t}{\int_{t_1}^{t_4}(f_{qa}(t)-n_{qa}(t))\mathrm{d}t}$$

类似可计算得到在 $[t_s,t_e]$内 md_b 和 md_c 分别为 0 和 1, 表示对象 b 和 c 在 $[t_s,t_e]$内到 q 的距离不可能小于和一定小于 R, 可见定义 2 中关于匹配度的定义是符合语义的.

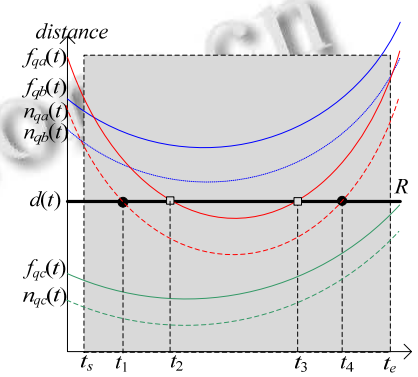


图 1　距离曲线图

处理 DRqTI 查询的关键之一就是确定以上两类事件点. $n_{qo}(t)=R$ 经过整理后为 $A_nt^2+B_nt+C_n=0$ 形式, 外部事件点的分布与该方程的解相关. 方程的两个根 t_{n1}、t_{n2} 有 6 种分布可能, 其与外部事件点及 t_s 和 t_e 的关系见表 1, 其中"-"表示不存在.

表 1　外部事件点分布关系

情况	大小关系	t_{n+}	t_{n-}	t_{os}	t_{oe}	备注
i	$t_s<t_{n1}<t_{n2}<t_e$	t_{n1}	t_{n2}	t_{n1}	t_{n2}	
ii	$t_{n1}<t_s<t_{n2}<t_e$	-	t_{n2}	t_s	t_{n2}	
iii	$t_s<t_{n1}<t_e<t_{n2}$	t_{n1}	-	t_{n1}	t_e	
iv	$t_{n1}<t_s<t_e<t_{n2}$	-	-	t_s	t_e	$n_{qo}(t)$在 $d(t)=R$ 下方
v	$t_{n1}<t_{n2}<t_s<t_e$	-	-	-	-	$n_{qo}(t)$在 $d(t)=R$ 上方
vi	$t_s<t_e<t_{n1}<t_{n2}$	-	-	-	-	$n_{qo}(t)$在 $d(t)=R$ 上方

根据内部事件点分布, 概率计算有 6 种情况, 由于篇幅问题, 不再展开讨论. 由 $n_{qo}(t)$和 $f_{qo}(t)$的定义可知, $f_{qo}(t)-n_{qo}(t)$和 $R-n_{qo}(t)$是可积的初等函数, 因此积分计算的时间复杂度为 $O(1)$.

2　DRqTI查询处理

以 TPR 树作为不确定移动对象的索引[3,4], 在此基础上采用以下规则分别进行剪枝处理.

规则 1. 设 TPR 树的结点 E 在 t_s 和 t_e 时刻的外包矩形分别为 $\text{MBR}(E)_{t_s}=[l_i(t_s),u_i(t_s)]_{i=1}^{d}$,

$MBR(E)_{t_e}=[l_i(t_e),u_i(t_e)]_{i=1}^{d}$, 其中 $l_i(t)$、$u_i(t)$表示 t 时刻 E 的外包矩形在第 i 维的下限和上限, d 表示维度. 于是 E 在 $[t_s,t_e]$ 内扫过的空间区域的外包矩形为 $MBR(E)_{t_s}^{t_e}=[\min(l_i(t_s),l_i(t_e)),\max(u_i(t_s),u_i(t_e))]_{i=1}^{d}$. 设 $tr(q)_{t_s}^{t_e}$ 是在$[t_s,t_e]$内 q 的不确定轨迹在 x-y 平面上的投影, R 是距离阈值, 如果 $MBR(tr(q)_{t_s}^{t_e}\oplus R)\cap MBR(E)_{t_s}^{t_e}=\varnothing$, 则 E 的所有子结点都可以被剪枝, $\oplus$表示求 Minkovski 和[9].

证明: 如图 2 所示, $tr(q)_{t_s}^{t_e}\oplus R$ 是在$[t_s,t_e]$内每一时刻到 q 的距离可能小于 R 的空间点构成的集合. $MBR(E)_{t_s}^{t_e}$ 包含了结点 E 在$[t_s,t_e]$内扫过的空间区域. $tr(q)_{t_s}^{t_e}\oplus R\subseteq MBR(tr(q)_{t_s}^{t_e}\oplus R)$, 如果 $MBR(tr(q)_{t_s}^{t_e}\oplus R)\cap MBR(E)_{t_s}^{t_e}=\varnothing$, 则 $tr(q)_{t_s}^{t_e}\oplus R\cap MBR(E)_{t_s}^{t_e}=\varnothing$. 所以在$[t_s,t_e]$内的任意时刻, E 的外包矩形到 q 的最近距离大于 R, 于是 E 及其子结点可以被排除.

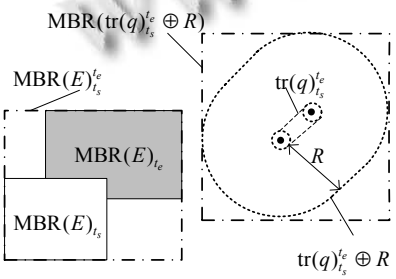


图 2 规则 1 证明示例

引理 1. 假设在时间段$[T^*_s,T^*_e]$内, 结点 E 的外包矩形 MBR(E)与 $q.ur\oplus R$ 相遇, 在时间段$[T_s,T_e]$内 MBR(E)与 MBR($q.ur\oplus R$)相遇, 则有$[T^*_s,T^*_e]\subseteq[T_s,T_e]$.

证明: 如图 3 所示, 在任意时刻 $q.ur\oplus R\subseteq$MBR($q.ur\oplus R$), 因此 MBR($q.ur\oplus R$)必先于 $q.ur\oplus R$ 与 MBR(E)相遇, 且晚于 $q.ur\oplus R$ 与 MBR(E)相离. 如果 MBR($q.ur\oplus R$)与 MBR(E)不相遇, 则 $q.ur\oplus R$ 与 MBR(E)也不会相遇. 于是$[T^*_s,T^*_e]\subseteq[T_s,T_e]$成立.

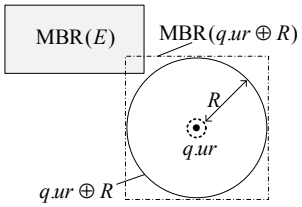


图 3 引理 1 证明示例

规则 2. 设结点 E 的外包矩形 MBR(E)与 MBR($q.ur\oplus R$)的相遇时间段为 $[T_s,T_e]$, 若 $[T_s,T_e]\cap[t_s,t_e]=\varnothing$, 则 E 及其子结点可以被排除.

MBR(E)和 MBR($q.ur\oplus R$)都是时变矩形, 大小和形状随时间发生变化. 根据引理 1, 规则 2 显然成立.

引理 2. 在任意时刻, MBR(E)与 $q.ur$ 之间的最近距离不小于 MBR(E,$q.r$)与 $q.c$ 之间的最近距离, 其中 MBR(E,$q.r$)由 MBR(E)在各维度上向外扩展 $q.r$ 得到.

证明: ① 当 $q.c$ 位于图 4(a)阴影部分, 即 MBR(E,$q.r$)–$q.ur\oplus$MBR(E) 内, 此时 MBR(E,$q.r$)与 $q.c$ 的最近距离为 0, 而 MBR(E)与 $q.ur$ 的最近距离大于 0, 见图 4(b); ② 当 $q.c$ 位于 $q.ur\oplus$MBR(E)内, 即图 4(a)中的圆角矩形内时, MBR(E)与 $q.ur$ 的最近距离和 MBR(E,$q.r$)与 $q.c$ 的最近距离都为 0; ③当 $q.c$ 在其他位置时, 即位于 MBR(E,$q.r$)外时, 两个距离不为 0 且相等. 引理 2 成立.

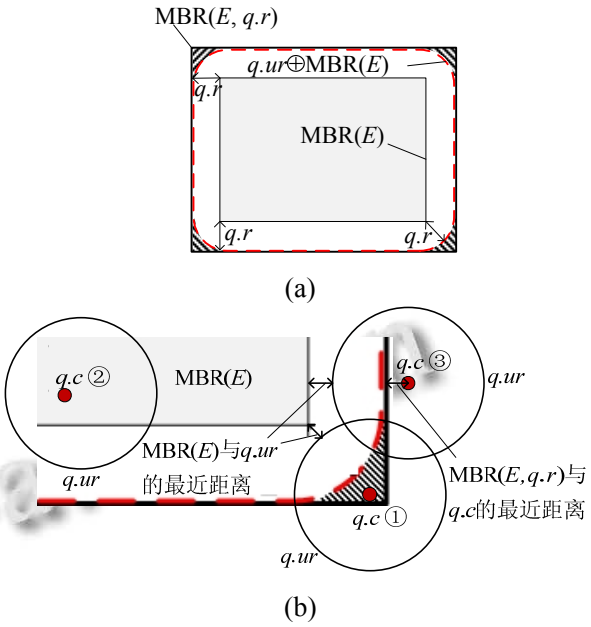


图 4 引理 2 证明示例. (a) MBR(E)、$q.ur\oplus$MBR(E)与 MBR(E, $q.r$); (b)距离关系.

规则 3. 如果在$[t_s,t_e]$内, MBR(E,$q.r$)与 $q.c$ 之间最近距离的最小值大于 R, 则 E 及其子结点可以被排除.

根据引理 2, 规则 3 显然成立.

可见, 规则 1~3 分别从不同的角度对结点进行剪枝.

规则 4. 如果在$[t_s,t_e]$内, $n_{qo}(t)$始终在 $d(t)=R$ 上方, 则对象 o 可排除.

证明: 在$[t_s,t_e]$内 $n_{qo}(t)$始终在 $d(t)=R$ 上方, 说明 q 与 o 之间的最近距离必然一直大于 R, 显然由公式(3)

可知 md_o=0, 不符合定义 2, 因此可以排除.

规则 4 说明只有外部事件点分布符合表 1 中情况 i~iv 的对象才可能满足查询, 其他对象都被排除. 在规则 1~4 的基础上, 得到了查询处理算法, 见图 5.

在算法的第 13 行分别使用规则 1~3, 得到的对应查询处理方法分别记为 DRqTI-r, DRqTI-t 和 DRqTI-d. 下一节将对三种方法的性能进行对比分析.

输入: TPR 树的根结点 *root*, 查询发出者 *q*, 距离阈值 *R*, 时间区间[t_s,t_e]
输出: 结果队列 *RQ*, 数据元素形为<*o*,t_{os},t_{oe}, md_o >

```
1. 初始化队列 Queue;
2. 将 root 插入 Queue;
3. while (Queue 非空){
4.     从 Queue 中取出第一个元素 E;
5.     if (E 是叶子结点)
6.         for (E 中每一个对象 o)
7.             if (规则 4 不能排除 o) {
8.                 计算出 tos 和 toe;
9.                 计算 mdo;
10.                将<o, tos, toe, mdo>插入 RQ; }
11.            else      // E 是中间结点
12.                for (E 中每一个项 e)
13.                    if (规则 i 不能排除 e)   //i∈{1,2,3}
14.                        将 e 插入 Queue; }
15. return RQ;
```

图 5　查询处理算法

3　实验分析

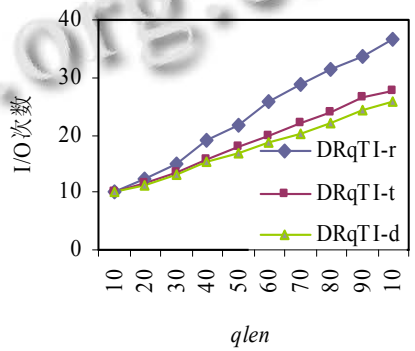
采用移动对象数据产生器[10]产生数据集, 随机从数据集中选出 100 个对象作为查询发出者, 交通网地图采用 Oldenburg. 实验中用到的参数含义及取值见表 2, 下划线表示参数的默认值.

表 2　实验参数说明

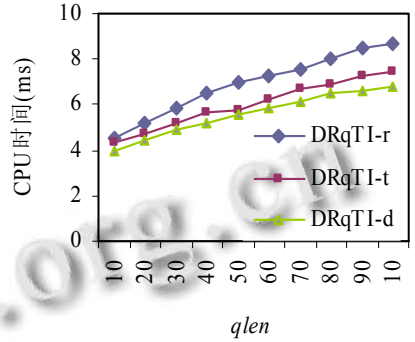
参数	说明	取值
#obj	数据集大小	10K,20K,30K,40K
r_{max}	最大不确定域半径	50,100,150,...,500
r	不确定域半径	r_{max}/4, r_{max}/2, 3r_{max}/4, r_{max}
R	距离阈值	1000,1500,2000,...,5000
qlen	时间区间[t_s,t_e]的长度	10,20,...,50,...,100

图 6~图 9 分别给出了三种方法的性能随参数 *qlen*、*R*、r_{max} 和 *#obj* 的变化情况. 可见 CPU 时间和 I/O 次数基本上随这 4 个参数增加而上升. 根据定义 2 和规则 1~4 分析, 这是显然的.

从各组实验可知 DRqTI-d 方法的性能最优, 而 DRqTI-r 方法最差. 主要原因是 DRqTI-r 方法并不是直接采用结点 *E* 和 *q* 在[t_s,t_e]内扫过的空间区域作为剪枝条件, 而是通过它们扫过的空间区域的外包矩形来剪枝. 如图 2 所示, 这会产生较大的无效搜索区域, 随着 *R*、r_{max} 和 *qlen* 的增加, 无效搜索区域越来越大, 访问的无关结点数量增加; 在地图不变的情况下, *#obj* 的增加虽然不会直接影响无效搜索区域的大小, 但是对象密度增加, 也使得搜索区域内无关结点增加, 引发了更多额外的 I/O 操作. 从而计算量也相应增加, 消耗更多的 CPU 时间.



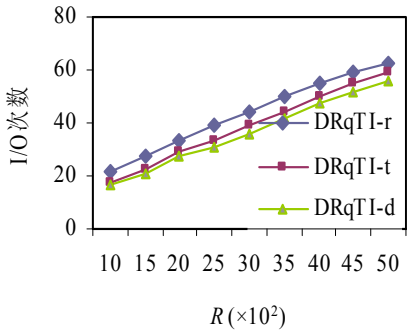
(a) I/O 次数



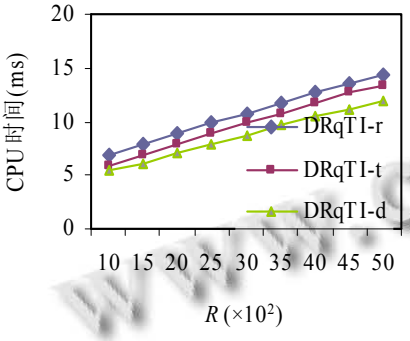
(b) CPU 时间

图 6　算法性能 vs. *qlen*

DRqTI-t 利用规则 2 对 TPR 树中的结点 *E* 进行剪枝, 计算 MBR(*E*) 与 MBR(*q.ur*⊕*R*) 相遇的时间段 [T_s,T_e], 根据引理 2 可知该时间区间一定包含 MBR(*E*) 与 *q.ur*⊕*R* 相遇的时间段. 换句话说 DRqTI-t 方法中也存在无效的搜索区域, 即 MBR(*q.ur*⊕*R*)−*q.ur*⊕*R*, 见图 3, 这也使得部分无效结点无法被排除. 而 DRqTI-d 方法通过计算 MBR(*E*,*q.r*)与 *q.c* 之间最近距离的最小值来剪枝, MBR(*E*,*q.r*)是由 MBR(*E*)在各维度上向外扩展 *q.r* 得到的, 因此无效部分仅是 MBR(*E*,*q.r*)−*q.ur*⊕MBR(*E*), 即图 4(a)中阴影部分, 是三种方法中最小的. 因此 DRqTI-d 在三种方法中剪枝效果最佳, 性能也最好.

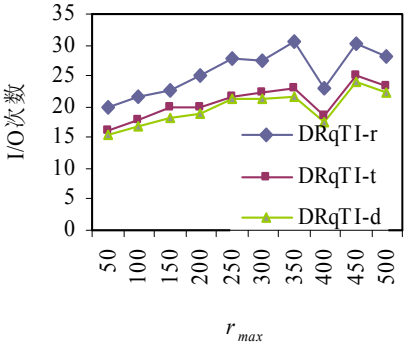


(a) I/O 次数

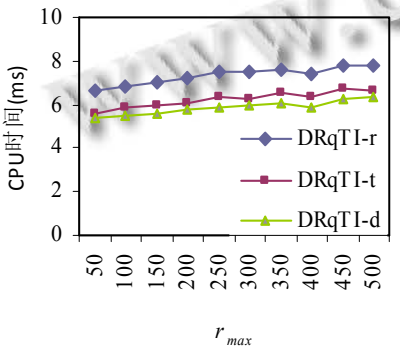


(b) CPU 时间

图 7 算法性能 vs. R

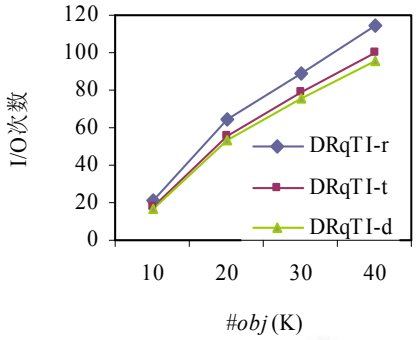


(a) I/O 次数

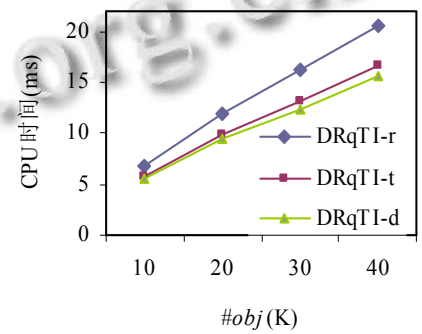


(b) CPU 时间

图 8 算法性能 vs. r_{max}



(a) I/O 次数



(b) CPU 时间

图 9 算法性能 vs. *#obj*

4 结语

针对查询发出者和目标位置均不确定的情况, 提出了 DRqTI 问题; 给出了最近、最远距离及内外事件点的定义; 在此基础上提出了基于轨迹、基于时间和基于距离的剪枝方法, 通过过滤、精炼和匹配度计算等步骤实现查询; 由实验验证了算法的有效性.

公式(3)定义的匹配度计算简单效率高, 符合 DRqTI 语义, 不会引起误报和漏报; 但是不能反映对象位置在不确定域内的分布特征[2], 也就是说无法区分不确定性呈均匀分布还是高斯呈分布等情况[3,4]. 文献[11]给出了两个不确定对象之间距离分布的表示模型, 文献[12]指出如果用二维随机变量表示不确定移动对象的可能位置, 则两个移动对象之间距离的分布可以表示为它们概率密度函数的卷积. 这些模型可以计算出更准确的匹配度, 但是计算量较大, 势必增加查询处理时间. 如何在匹配度计算的效率和准确性之间找到折衷, 这将是我们下一步的工作.

参考文献

1 周傲英,金澈清,王国仁,李建中.不确定性数据管理技术研

究综述.计算机学报,2009,32(1):1–16.

2 Wang Y, Li X, Li X. A survey of queries over uncertain data. Knowledge Information Systems, 2013, 37(3): 485–530.

3 Chen Y, Qin X, Liu L. Uncertain distance-based range queries over uncertain moving objects. Journal of Computer Science and Technology, 2010, 25(5): 982–998.

4 Chen Y, Qin X, Liu L, Li B. Fuzzy distance-based range query over uncertain moving objects. Journal of Computer Science and Technology, 2012, 27(2):376–396.

5 Huang YK, Liao SJ, Lee C. Evaluating continuous k-nearest neighbor query on moving objects with uncertainty. Information Systems, 2009, 34(4): 415–437.

6 Trajcevski G, Tamassia R, Cruz IF, Scheuermann P, Hartglass D, Zamierowski C. Ranking continuous nearest neigbors for uncertain trajectories. The VLDB Jounral, 2011, 20(5): 767–791.

7 Zheng K, Trajcevski G, Zhou X, Scheuermann P. Probabilistic range queries for uncertain trajectories on road networks. In: Anastasia A, Sihem AY, Jignesh MP, Tore R, Pierre S, Julia S, eds. Proc. of the 14th International Conference on Extending Database Technology. NewYork. ACM. 2011.283–294.

8 Čivilis A, Jensen CS, Pakalni S. Techniques for efficient road network-based tracking of moving objects. IEEE Trans. on Knowledge and Data Engineering, 2005, 17(5): 698–713.

9 Berg MD, Kreveld MV. 邓俊辉译.计算几何算法与应用.北京:清华大学出版社, 2005.

10 Brinkhoff T. A framework for generating network-based moving objects. GeoInformatica, 2002, 6(2): 153–180.

11 Hung E, Xiao L. An efficient representation model of distance distribution between two uncertain objects. http://www4.comp.polyu.edu.hk/~csehung/paper/wmwa14-lncs.pdf. [2013-09-01].

12 Trajcevski G, Tamassia R, Ding H, Scheuermann P, Cruz IF. Continuous probabilistic nearest-neighbor queries for uncertain trajectories. In: Kersten ML, Novikov B, Teubner J, Polutin V, Manegold S, eds. Proc. of the 12th International Conference on Extending Database Technology: Advances in Database Technology. NewYork. ACM. 2009. 874–885.