一种障碍空间中不确定对象的连续最近邻查询方法

李传文 谷 峪 李芳芳 于 戈
(东北大学信息科学与工程学院计算机软件与理论研究所 沈阳 110004)
(医学影像计算教育部重点实验室(东北大学) 沈阳 110004)

摘 要 近年来,基于位置的服务获得了越来越广泛的关注,其中最近邻查询是最常用的一种查询方式.测量手段的不准确性以及数据本身的性质导致不确定性在位置数据中普遍存在,这种不确定性会对最近邻查询结果产生影响.空间中障碍物的存在也给空间数据查询带来了挑战.文中研究存在障碍物的空间中不确定对象连续最近邻查询的处理方法,设计了一种剪枝策略大幅降低需要计算的不确定对象数目,并进一步提出了障碍空间中不确定对象最近邻查询安全区域的概念及安全区域生成算法.设计了安全区域的索引存储方法.实验结果表明,文章所提出的方法具有良好的效率和可扩展性.

关键词 最近邻;不确定;障碍空间;基于位置的服务
中图法分类号 TP311 **DOI 号**: 10.3724/SP.J.1016.2010.01359

A Continuous Nearest Neighbor Query Method for Uncertain Data in Obstructed Spaces

LI Chuan-Wen Gu Yu LI Fang-Fang YU Ge
(*Institute of Computer Software and Theory of Information Science and Engineering Institute, Northeastern University, Shenyang* 110004)
(*Key Laboratory of Medical Image Computing, Northeastern University, Ministry of Education, Shenyang* 110004)

Abstract In recent years, location-based services (LBS) are getting more and more attention. The nearest neighbor query is the most common query type in the LBS area. The uncertainty of data exists commonly due to the inaccuracy of measurement instructions and the data attributes itself. This uncertainty will affect the results of nearest neighbor queries. The existence of obstacles in planes also put challenges to spatial data queries. This paper studies the continuous nearest neighbor query by the existence of obstacles and the uncertainty of data. It also gives a pruning strategy that greatly reducs the number of objects which need to be calculated. Furthermore, this paper proposes the safe region concept with regard to uncertain data in obstructed spaces, an algorithm to generate the safe regions, and an indexing method for saving safe regions. Experimental results show that the proposed method has good efficiency and scalability.

Keywords nearest neighbor; uncertain; obstructed space; location-based services

1 引 言

近年来,基于位置的服务(Location Based Services,LBS)获得了越来越多的关注,并越来越广泛地应用到生产生活当中.LBS 以查询为基础,根据用户的当前位置信息为用户提供各种有用的信息,例如,最近的自动提款机在哪,附近有哪些饭店等等.

收稿日期:2010-06-11.本课题得到国家自然科学基金(60773220,60933001)、国家"八六三"高技术研究发展计划"高效的纯 XML 数据管理关键技术研究及原型系统实现"(2009AA01Z131)资助.**李传文**,男,1982 年生,博士,主要研究方向为时空数据管理、复杂事件处理等.E-mail: lichuanwen@ise.neu.edu.cn.**谷 峪**,男,1981 年生,博士,副教授,主要研究方向为时空数据管理、RFID 数据管理、数据流等.**李芳芳**,女,1977 年生,博士,讲师,主要研究方向为传感器数据管理等.**于 戈**,男,1962 年生,博士,教授,博士生导师,主要研究领域为数据库理论与技术.

LBS 包括多种查询类型及处理方法，如最近邻、KNN、Skyline、反 KNN 等. 在这些查询中，最近邻查询是最常用的一种查询方式.

现已提出若干最近邻查询处理方法[1-4]. 这些方法大部分针对确定数据(即完整且精确的位置数据)并且全都假设空间中不存在障碍物(即任意两点之间可以直线连接). 然而，在许多实际应用中还存在大量的不确定数据以及障碍物的情况. 例如，卫星图像或 GPS 定位，由于设备本身的硬件条件限制或数据传输过程中产生的噪声都会使数据中的位置信息具有不确定性. 而且，地面上移动的物体一般都会受到地理条件的限制(例如建筑、湖泊等)，两物体之间的最短距离必须考虑障碍物的因素.

本文提出障碍空间内不确定对象连续最近邻查询的处理方法. 与精确对象不同，不确定对象的位置不是一个确定点，而是一个范围，在这个范围内该对象服从某种概率分布函数. 在障碍空间中，两个点之间的距离由他们绕过障碍物的最短距离决定. 例如一辆汽车在行驶的过程中不断地查询距离自己最近的加油站的位置. 由于地图数据不精确以及 GPS 测量误差等因素，汽车自身的位置以及加油站的位置都不是绝对准确的，它们都归属于某个不确定区域. 同时，由于自然地形或建筑物的阻挡，汽车和加油站之间的距离也不仅仅是它们之间的欧氏距离.

文献[5]给出了两点之间障碍距离的计算方法. 然而，当障碍空间中的对象为不确定对象时，现有的计算方法不再适用，这是因为两对不确定对象之间距离远近的比较结果不确定. 本文对最近邻的概念进行了扩展，将最近邻查询的结果从某个确定的对象扩展到一个对象集合，其中包含所有可能成为最近邻的不确定对象. 本文考虑不确定对象的概率分布，用不确定对象的分布半径作为参数，提出了一种处理方法计算最近邻可能对象的集合，并在该算法中设计了一种剪枝策略，通过不断更新剪枝条件，大幅减少需要计算的不确定对象数目.

本文进一步提出了障碍空间中不确定对象最近邻查询安全区域的概念及安全区域生成算法. 文献[6]提出了欧氏空间中确定对象的安全区域概念，所谓安全区域即空间中某些点的集合，这些点具有相同的最近邻查询(或其它 LBS 查询，如 K 近邻)结果集. 当查询点在某个结果集的安全区域内移动时，不再需要重复查询，因而事先生成安全区域可以节省大量的实时计算开销. 提出了一种基于四分树的索引方法对安全区域进行存储，该方法具有较优的查询性能.

最近几年，对于不确定数据的研究工作取得了广泛的成果[7-8]. 在不确定数据查询领域，研究者已提出了多种类型的查询及其处理方法，如一般的关系查询、top-k 查询、kNN 查询、概率 Skyline 查询等. 文献[9]提出了不确定 Voronoi 图的概念，文献[10-11]也针对特定的不确定查询提出了相应的方法. 然而，据我们所知，目前尚无对障碍空间中的不确定数据进行有效最近邻查询的方法.

目前对障碍空间数据查询的研究主要集中在可视 kNN 及反 kNN 查询[12-14](即只考虑直线可达而不考虑绕过障碍物的情况). 文献[15]提出了障碍空间中连续 kNN 查询的解决方法.

综上所述，对障碍空间中的不确定数据进行最近邻查询的研究具有现实意义和理论价值.

本文研究的主要内容如下：

(1) 基于空间障碍物及不确定对象的数据模型，提出障碍空间中不确定对象最近邻查询问题.

(2) 设计一种高效的基于障碍空间距离的算法来进行查询处理. 运用了一种剪枝技术来提高性能.

(3) 提出不确定对象分割区域的概念，并以之为基础设计一种有效的安全区域生成方法.

(4) 提出一种高效的索引方法，对安全区域进行存储. 该方法比传统的 R-tree 索引方法需要的存储代价低并且查询速度较快.

(5) 通过实验考察并证实了本文提出的方法具有良好的效率和可扩展性.

本文第 2 节介绍相关术语并给出数据模型和问题定义；第 3 节给出障碍空间中不确定对象最近邻查询的计算方法，提出对象分割区域的概念、安全区域生成算法以及索引存储方法；第 4 节给出实验结果和分析；第 5 节总结全文.

2 问题定义

2.1 数据模型

本文考虑具有 n 个查询对象的集合 $P=\{p_1, p_2, \cdots, p_n\}$，$p_i \in \mathbb{R}^2$，$2 \leq n \leq \infty$，障碍物集 $O=\{O_1, O_2, \cdots, O_n\}$ 以及一个不确定查询点 q. 集合 P 为查询对象集，其中 $\mathbb{R}^2$ 代表欧氏平面，集合 O 为障碍物集，其中障碍物具有确定的边界且相互不重叠，查询对象具有不确定性且都在障碍物之外. 为描述方便，本文假定障碍物都是凸多边形(如果有障碍物为凹多边形，可将其分解为几个凸多边形的组合).

查询点 q 及集合 P 中每个查询对象 p_i 均具有一个“不确定区域”(U_q 或 U_{p_i})以及对应的概率分布

函数(Probability Distribution Function，PDF)．不确定区域包含该对象所有可能出现的位置，对象在该区域某点的概率由概率分布函数确定．不确定区域可以为任意形状，概率分布函数也可以是任意分布(如均匀分布、高斯分布等)．若无特别说明，后文中不确定区域均是指圆形分布，定义 $\dot{p}_i$ 为 U_{p_i} 的圆心，r_{p_i} 为 U_{p_i} 的半径，对不规则区域的讨论将在3.3节给出．

令 $VIS(p)$ 为点 $\dot{p}_i$ 的可视区域，即区域 $VIS(p)$ 中任意一点与点 $\dot{p}_i$ 的连线都不与障碍物相交．在区域 $S=\mathbb{R}^2\backslash O$ 上，定义两点之间的最短距离如下．

定义 1.　障碍空间 $S=\mathbb{R}^2\backslash O$ 中两具有确定位置的点 p_i，p_j 之间的最短距离 $d(p_i,p_j)$ 为

$$d(p_i,p_j)=\min\Big(\Big(\sum_{k=1}^{m-1}|x_k,x_{k+1}|\Big)+|p_i,x_1|+|p_j,x_{m-1}|\Big),$$

其中点 x_k，x_{k+1} 为任意两可见点，m 为可见点总数，即直线段 $\overline{x_kx_{k+1}}$ 不与 O 中任何障碍物相交，$|x_k,x_{k+1}|$ 表示线段 $\overline{x_kx_{k+1}}$ 的长度．

例 1.　图1示例了一个具有两个障碍物的空间，竖条阴影部分($\triangle abc$ 及 $\triangle def$)表示障碍物，点 p_1，p_2，p_3 周围的圆形阴影表示不确定对象 p_1，p_2，p_3 的不确定区域，q 表示查询点．其中 p_1，p_2 圆心的最短距离由连线段 $p_1c\rightarrow cd\rightarrow dp_2$ 组成(即图中黑色粗线段)．这条连线段比其它所有连线段(如 $p_1b\rightarrow bd\rightarrow dp_2$ 或 $p_1c\rightarrow cf\rightarrow fe\rightarrow ep_2$ 等)短．

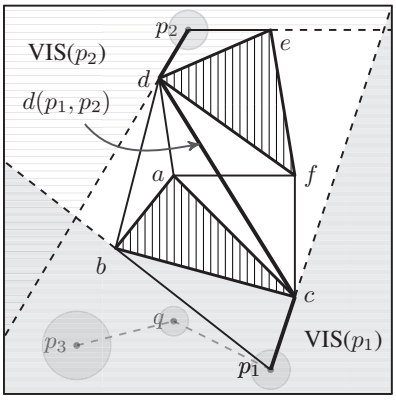


图1　不确定对象的最近邻

基于确定点之间最短距离的定义，本文提出障碍空间中不确定对象的最近邻查询如下定义．

定义 2.　查询点 q 在障碍空间 $S=\mathbb{R}^2\backslash O$ 中的最近邻定义为一个集合 $N=\{n_1,n_2,\cdots,n_m\}$，其满足

(1) $n_i\in P(1\leq m)$；

(2) $d_{\min}(q,n_i)>\forall d_{\max}(q,n_j)$，$j\neq i$，$n_j\in P$．其中 $d_{\min}(q,n)$ 与 $d_{\max}(q,n)$ 分别代表点 q 与区域 n 之间的最短与最长距离．

障碍空间中不确定对象的最近邻查询结果为一个集合，包含所有可能成为查询点最近邻的不确定对象．

例 2.　对图1中查询点 q 求其最近邻，查询点 q 到查询对象 p_1，p_2，p_3 的距离分别属于区间 $Range_{q,p_1}=[d(\dot{q},\dot{p}_1)-r_q-r_{p_1},d(\dot{q},\dot{p}_1)+r_q+r_{p_1}]$，$Range_{q,p_2}=[d(\dot{q},\dot{p}_2)-r_q-r_{p_2},d(\dot{q},\dot{p}_2)+r_q+r_{p_2}]$，$Range_{q,p_3}=[d(\dot{q},\dot{p}_3)-r_q-r_{p_3},d(\dot{q},\dot{p}_3)+r_q+r_{p_3}]$．通过观察可知，两区间 $Range_{q,p_1}$，$Range_{q,p_2}$ 存在交集且都小于 $Range_{q,p_3}$，因此查询对象 p_1 和 p_2 都有可能成为查询点 q 的最近邻，p_3 不可能成为查询点 q 的最近邻，故查询结果为 $P_n=\{p_1,p_2\}$．

2.2　问题描述

本文假设空间中查询对象和障碍物的信息保存在某种常用空间索引当中(如R-tree等)，且不存在特殊的空间数据结构(如Voronoi图或 k 阶Voronoi图等)．这是因为特殊的空间数据结构维护代价大，并且应用范围较窄．例如Voronoi图或 k 阶Voronoi图只适用于查询对象和障碍物固定情况下的最近邻或 k 近邻查询，当 k 变化或空间信息变化时这些数据结构的更新操作需要高昂的代价．

障碍空间中不确定对象的最近邻查询与传统的最近邻查询主要区别在两点：首先，空间中两点之间的距离不是直线距离，而是绕过障碍物的最短距离；其次，两点间的最短距离不是一个确定的值，而是一个范围．

基于上述特点，障碍空间中对不确定对象的最近邻查询可以分为3步：(1)找到可能成为查询点最近邻的所有可能结果的集合；(2)找到该集合的安全区域，所谓安全区域即与查询点有相同的查询结果的点的集合；(3)将查询结果和安全区域的信息返回给查询点．下节主要讨论不确定对象连续最近邻查询的前两步．

3　障碍空间中不确定对象最近邻查询

3.1　不确定对象最近邻查询

本节假定查询点具有确定位置，查询点为不确定区域的情况将在3.3节讨论．下面提出计算距离某确定点最近的不确定对象集的算法，其中采用文献[15]提出的算法计算障碍空间中两确定点之间最短距离 $d(p_1,p_2)$．

算法 1. 不确定对象最近邻生成.

输入:查询点 q,查询对象集 P

输出:最近邻结果集 P_n

1. $bound \leftarrow \infty, PS \leftarrow \varnothing, PU \leftarrow P$
2. $l_{\max} \leftarrow 0, P_n \leftarrow \varnothing$
3. 根据 O 构建可见图 $G(V,E)$
4. While($PU \neq \varnothing$)
5. $p_n \leftarrow$ 从 PU 中取出 q 欧氏距离最近邻
6. If($d(p_n,q)-r_{p_n} > bound$)
7. Return
8. If($d(p_n,q)-r_{p_n} < l_{\max}$)
9. $P_n \leftarrow P_n + p_n$
10. $l_{\max} \leftarrow d(p_n,q)-r_{p_n}$
11. RefreshCandidates(p_n)
12. $PS \leftarrow PS+p_n$, $PU \leftarrow PU-p_n$
13. $bound \leftarrow \min(bound, d(p_n,q)+r_{p_n})$
14. Return

算法 1 的工作过程如下:首先构造区域 S 的可见图 $G(V,E)$,其中 E 包含 O 中障碍物的所有顶点及 P 中所有不确定对象的圆心,V 包含 E 中所有顶点之间的可见线段.图 1 中示例了一个可见图 $G(V,E)$,其中 E 包括障碍物顶点 a,b,c 及不确定对象 $\dot{p}_1$,$\dot{p}_2,\dot{p}_3$ 等,V 包括图中所有实线段($\overline{p_2 d},\overline{dc},\overline{fc}$ 等).

集合 PS,PU 分别保存已查询点和未查询点,每次循环从 PU 当中选择一个离 q 欧氏距离最近的点 p_n.因为查询对象已经保存在某种空间索引中,所以选择点 p_n 只需要很小的运行开销(例如,R-tree 索引查找最近点 p_n 需要的查询复杂度接近常数).

算法 1 第 8~10 行检查是否将点 p_n 加入到结果集 P_n 中.$l_{\max}$ 保存结果集 P_n 的上界,如果 p_n 到 q 的最小值大于 $l_{\max}$,则 p_n 不可能是 q 的最近邻,否则 p_n 属于结果集 P_n.如果 p_n 是结果之一,则需要更新 $l_{\max}$ 并且调用子算法 RefreshCandidates 将结果集中不满足条件的结果去掉.

算法 2. 候选集刷新(RefreshCandidates).

输入:新加入查询对象 p_n

输出:最近邻结果集 P_n

1. $max \leftarrow d(p_n,q)-r_{p_n}$
2. Foreach($p \in P_n$)
3. If ($d(p,q)-r_p > max$)
4. $P_n \leftarrow P_n - p$

为提高性能,算法 1 采用边界量 $bound$ 对象的计算条件.第 13 行对 $bound$ 赋值,第 6、7 行采用 $bound$ 对查询对象进行剪枝.因为如果查询对象 p_1 到查询点的最近距离比另一查询对象 p_2 到查询点的最远距离还大,那么 p_1 绝不可能成为查询点的最近邻,所以查询对象 p_1 可以被安全剪枝.

当 PU 未被剪枝掉的查询对象都被检查过之后,P_n 包含的即是所有可能成为查询点最近邻的查询的集合.

3.2 最近邻查询结果的安全区域

3.2.1 不确定对象的区域分割

首先,本文将欧氏空间中两点之间平分线的概念在障碍空间中进行推广,提出两不确定对象之间区域分割的概念.

定义 3. 两不确定对象 p_1,p_2(其不确定区域为 U_{p_1} 及 U_{p_2})将障碍空间 $S=\mathbb{R}^2 \backslash O$ 分为 3 个分割区域:$B_{\bar{p}_1,p_2}$、B_{p_1,p_2} 及 $B_{\bar{p}_2,p_1}$,其满足

(1) $B_{\bar{p}_1,p_2} \cup B_{p_1,p_2} \cup B_{\bar{p}_2,p_1} = S$;

(2) $d_{\max}(x,p_i) < d_{\min}(x,p_j), x \in B_{\bar{p}_i,p_j}$;

(3) $\begin{matrix} d_{\max}(x,p_i) > d_{\min}(x,p_j) \\ d_{\max}(x,p_j) > d_{\min}(x,p_i) \end{matrix}, x \in B_{p_i,p_j}$.

例 3. 图 2 示例了障碍空间中两不确定对象 p_1 和 p_2 的分割区域.其中上方横线阴影部分为 $B_{\bar{p}_2,p_1}$,该区域内所有点到不确定对象 p_2 的距离肯定大于到不确定对象 p_1 的距离.下方阴影部分为 $B_{\bar{p}_1,p_2}$,该区域内所有点到不确定对象 p_1 的距离肯定大于到不确定对象 p_2 的距离.而中间未标阴影的部分为 B_{p_1,p_2},p_1 和 p_2 都有可能成为距离该区域内的点最近的不确定对象.

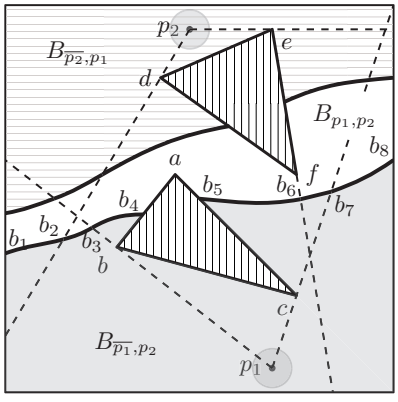


图 2 分割区域

下面给出分割区域的部分性质及构造方法.首先考虑构造 $B_{\bar{p}_1,p_2}$、B_{p_1,p_2} 及 $B_{\bar{p}_2,p_1}$ 的边界.由定义 3 知,$x \in B_{\bar{p}_i,p_j}$ 的条件为

$$d_{\max}(x,p_i) < d_{\min}(x,p_j) \tag{1}$$

其中 $d_{\max}$ 和 $d_{\min}$ 分为代表两点之间的最大距离和最小距离.则式(1)可以表示为

$$d(x,\dot{p}_i)+r_{p_i} < d(x,\dot{p}_j)-r_{p_j} \tag{2}$$

即 $B_{\bar{p}_i,p_j}$ 和 B_{p_i,p_j} 之间的边界可以表示为

$$l_{p_i,p_j} = \{x \mid d(x,\dot{p}_i)+r_{p_i}+r_{p_j} = d(x,\dot{p}_j)\} \tag{3}$$

以图 2 为例对定义 3 描述的分割区域的性质进行分析. 基于 l_{p_1,p_2} 与 l_{p_2,p_1} 在构造方式上的对称性，只需对边 l_{p_1,p_2} 的性质进行讨论. 观察图 2 可知，边 l_{p_1,p_2} 由一系列曲线构成，其中一部分位于点 $\dot{p}_1$ 和点 $\dot{p}_2$ 共同的可见区域 $VIS(\dot{p}_1)\cap VIS(\dot{p}_2)$ 内，其余部分则与点 $\dot{p}_1$ 和点 $\dot{p}_2$ 中至少一个不可见. 下面对这两种情况分别进行讨论.

(1) 在点 $\dot{p}_1$ 和点 $\dot{p}_2$ 共同的可见区域 $VIS(\dot{p}_1)\cap VIS(\dot{p}_2)$ 内，边 l_{p_1,p_2} 可定义为

$$l_{p_1,p_2}\cap VIS(\dot{p}_1)\cap VIS(\dot{p}_2)=\{x\mid |x,\dot{p}_1|+r_{p_1}+r_{p_2}=|x,\dot{p}_2|\}\cap VIS(\dot{p}_1)\cap VIS(\dot{p}_2) \tag{4}$$

令点 $\dot{p}_1$ 和点 $\dot{p}_2$ 的坐标分别为 (x_1,y_1) 和 (x_2,y_2)，并采用文献[9]提出的不确定对象之间距离的分析方法，令

$$\begin{cases}\cos\theta=(x_2-x_1)/|\dot{p}_1,\dot{p}_2|,\\ \sin\theta=(y_2-y_1)/|\dot{p}_1,\dot{p}_2|\end{cases} \tag{5}$$

则式 $|x,\dot{p}_1|+r_{p_i}+r_{p_j}=|x,\dot{p}_j|$ 可转换为

$$\frac{x_\theta^2}{\alpha^2}-\frac{y_\theta^2}{\beta^2}=1 \tag{6}$$

其中

$$\alpha=(r_{p_i}+r_{p_j})/2,\ \beta=\sqrt{\left(\frac{|\dot{p}_1,\dot{p}_2|}{2}\right)^2-\alpha^2};$$

$$x_\theta=\left(x-\frac{x_1+x_2}{2}\right)\cos\theta+\left(y-\frac{y_1+y_2}{2}\right)\sin\theta;$$

$$y_\theta=\left(\frac{x_1+x_2}{2}-x\right)\sin\theta+\left(y-\frac{y_1+y_2}{2}\right)\cos\theta.$$

式(6)指出，l_{p_1,p_2} 中满足式(4)的部分由双曲线的一段构成. 图 2 中 $VIS(\dot{p}_1)\cap VIS(\dot{p}_2)$ 为曲线段 $\overline{b_1b_2}$ 所在的阴影区域，而曲线段 $\overline{b_1b_2}$ 则是 l_{p_1,p_2} 满足式(6)的部分.

(2) 在点 $\dot{p}_1$ 和点 $\dot{p}_2$ 的非可见区域($\dot{p}_1$ 和 $\dot{p}_2$ 中至少一个点不可见)，l_{p_1,p_2} 由那些在该位置可见的点 $v(v\in E)$ 决定. 例如，图 2 中曲线段 $\overline{b_2b_3}$ 位于 $VIS(\dot{p}_1)\cap VIS(D)$，因此 $\overline{b_2b_3}$ 由点 D 和点 $\dot{p}_1$ 决定；曲线段 $\overline{b_3b_4}$ 位于 $VIS(B)\cap VIS(D)$，因此 $\overline{b_2b_3}$ 由点 D 和点 B 决定. 以 $\overline{b_3b_4}$ 为例，在 $VIS(B)\cap VIS(D)$ 中，边 l_{p_1,p_2} 可定义为

$$l_{p_1,p_2}\cap VIS(B)\cap VIS(D)=\{x\mid |x,B|+r_{p_1}-|\dot{p}_1,B|=|x,D|-r_{p_2}-|D,\dot{p}_2|\}\cap VIS(B)\cap VIS(D) \tag{7}$$

对比式(7)和式(4)可知，式(7)也可转换为双曲线的形式，即将点 B 和点 D 的坐标分别记为 (x_1,y_1) 和 (x_2,y_2)，且将式(6)中 α 和 β 分别替换为

$$\alpha=(r_{p_i}+r_{p_j}-|\dot{p}_1,B|+|D,\dot{p}_2|)/2$$

$$\beta=\sqrt{\left(\frac{|B,D|}{2}\right)^2-\alpha^2} \tag{8}$$

将式(8)代入式(6)可知，边 l_{p_1,p_2} 在点 $\dot{p}_1$ 和点 $\dot{p}_2$ 的非可见区域 $VIS(B)\cap VIS(D)$ 也是由双曲线构成. 根据双曲线的性质知，点 D 和点 B 分别是该双曲线的两个焦点.

定义 4.　作为双曲线焦点决定分割线部分曲线段形状的两点称为该曲线段的两个决定点. 如图 2 中曲线段 $\overline{b_3b_4}$ 的决定点为点 D 和点 B.

综合上述讨论，得到两不确定对象 p_1，p_2 在障碍空间 $S=\mathbb{R}^2\backslash O$ 中的分割边的性质.

定理 1.　两相邻分割区域的分割边 l_{p_i,p_j} 由一系列曲线段 $l_{p_i,p_j}=c^1_{p_i,p_j}\cup c^2_{p_i,p_j}\cup\cdots\cup c^n_{p_i,p_j}$ 组成，每一条曲线段都是某双曲线的一部分.

证明.　略(参考上文相关讨论).

观察图 2 可知，组成每条分割边的曲线段是部分相邻的，即某些曲线段邻接在一起而与另外的曲线段之间被障碍物分隔. 其中相邻的曲线段有如下性质.

定理 2.　相邻曲线段 $c^k_{p_i,p_j}$ 与 $c^{k+1}_{p_i,p_j}$ 具有一个共同的决定点. 令 M 为曲线段 $c^k_{p_i,p_j}$ 与 $c^{k+1}_{p_i,p_j}$ 的共同决定点，N^k 及 N^{k+1} 为二者各自的另一决定点，则式 $c^k_{p_i,p_j}\in VIS(N^{k+1})$ 与 $c^{k+1}_{p_i,p_j}\in VIS(N^k)$ 有且仅有一个成立. 令未成立的决定点为 $N^{k'}$，则曲线段 $c^k_{p_i,p_j}$ 与 $c^{k+1}_{p_i,p_j}$ 之间的交点位于 $VIS(M)$ 中 $VIS(N^{k'})$ 的边界处.

证明.　假设曲线段 $c^k_{p_i,p_j}$ 与 $c^{k+1}_{p_i,p_j}$ 没有相同的决定点，设曲线段 $c^k_{p_i,p_j}$ 与 $c^{k+1}_{p_i,p_j}$ 的决定点分别为 $N^{k,1}$，$N^{k,2}$，$N^{k+1,1}$ 和 $N^{k+1,2}$，则 $c^k_{p_i,p_j}\in VIS(N^{k,1})\cap VIS(N^{k,2})$，$c^{k+1}_{p_i,p_j}\in VIS(N^{k+1,1})\cap VIS(N^{k+1,2})$. 因为曲线段 $c^k_{p_i,p_j}$ 与 $c^{k+1}_{p_i,p_j}$ 相交，所以 $VIS(N^{k,1})\cap VIS(N^{k,2})$ 与 $VIS(N^{k+1,1})\cap VIS(N^{k+1,2})$ 相交，这与 $N^{k,1}$，$N^{k,2}$，$N^{k+1,1}$，$N^{k+1,2}$ 四点互异相矛盾. 所以相邻线段 $c^k_{p_i,p_j}$ 与 $c^{k+1}_{p_i,p_j}$ 必有一个共同的决定点.

假设式 $c^k_{p_i,p_j}\in VIS(N^{k+1})$ 与 $c^{k+1}_{p_i,p_j}\in VIS(N^k)$ 都成立，因为 $c^k_{p_i,p_j}\in VIS(N^k)$ 与 $c^{k+1}_{p_i,p_j}\in VIS(N^{k+1})$ 是隐含成立的. 所以曲线段 $c^k_{p_i,p_j}$ 与 $c^{k+1}_{p_i,p_j}$ 所处区域为 $VIS(M)\cap VIS(N^k)\cap VIS(N^{k+1})$. 这显然是不可能的.

假设式 $c^k_{p_i,p_j}\in VIS(N^{k+1})$ 与 $c^{k+1}_{p_i,p_j}\in VIS(N^k)$ 都不成立，说明 $VIS(N^k)$ 和 $VIS(N^{k+1})$ 恰好相切，且切线通过曲线段 $c^k_{p_i,p_j}$ 与 $c^{k+1}_{p_i,p_j}$ 的交点，这与定理 2 的前提条件相矛盾.　　证毕.

算法 3. 分割边生成.

输入：查询对象 p_1，p_2，可见图 $G(V,E)$

输出：分割边 l_{p_1,p_2}

1. $P_1 \leftarrow \dot{p}_1$，$P_2 \leftarrow \dot{p}_2$
2. While($P_1 \neq \varnothing$)
3. $n_1 \leftarrow \min\{p \mid d(p,p_1), p \in P_1\}$
4. While($P_2 \neq \varnothing$)
5. $n_2 \leftarrow \min\{p \mid d(p,p_2), p \in P_2\}$
6. If($VIS(n_1) \bigcap VIS(n_2) \neq \varnothing$)
7. $seg_{n_1,n_2} \leftarrow$ 根据式(6)及(8)计算双曲线方程
8. $l_{p_1,p_2} \leftarrow l_{p_1,p_2} \bigcup seg_{n_1,n_2}$
9. $P_2 \leftarrow P_2 - n_2 + vis(n_2)$
10. $P_1 \leftarrow P_1 - n_1 + vis(n_1)$

例 4. 以图 2 为例说明定理 1 和定理 2. 两不确定对象 p_1 和 p_2 之间分割区域 $B_{\dot{p}_1,p_2}$ 和 B_{p_1,p_2}，一条分割边由曲线段 $\overline{b_1b_2}$，$\overline{b_2b_3}$，$\overline{b_3b_4}$，$\overline{b_5b_6}$，$\overline{b_6b_7}$ 和 $\overline{b_7b_8}$ 组成，它们的决定点分别为 $\{\dot{p}_1,\dot{p}_2\}$，$\{\dot{p}_1,d\}$，$\{b,d\}$，$\{c,d\}$，$\{c,e\}$ 和 $\{\dot{p}_1,e\}$. 相邻曲线段的共用决定点依次为 $\dot{p}_1$，d，c 和 e，且有 $\overline{b_1b_2} \in VIS(d)$，$\overline{b_2b_3} \in VIS(b)$，$\overline{b_6b_7} \in VIS(d)$ 以及 $\overline{b_7b_8} \in VIS(c)$.

基于定理 1 和定理 2，给出障碍空间中两不确定对象之间分割边的生成算法. 由于分割区域是由分割边直接产生的，所以该算法即是障碍空间中两不确定对象的分割区域的产生算法.

算法 3 中 P_1 和 P_2 集合中分别保存了从 p_1 方向和 p_2 方向开始查找的点集，并将两集合中的元素按距离各自起始点由近到远的顺序依次配对. 第 6～8 行中如果两点的可视区域有交集的话，则计算该区域内以两点为决定点(即双曲线焦点)的双曲线，并将其加入到边集中. 每一组配对计算结束后，第 9、10 行分别对 P_1 和 P_2 集合进行刷新，将 n_2 和 n_1 查询对象从 P_1 和 P_2 集合中除去，并将它们的可见点加入到 P_1 和 P_2 集合中. 其中 $vis(n)$ 代表在点 n 可见并且从未被加入到集合中的节点 $v(v \in E)$.

3.2.2 安全区域生成

令 q 为查询点，查询对象集 P_n 为 q 的可能最近邻集，p 为 P_n 中任意对象. 通过观察图 2 知，任意查询对象 p 附近存在一个区域，在这个区域内 p 始终属于查询点 q 的可能最近邻集.

定理 3. 在查询对象 p 附近存在一个区域，当查询点 q 处于该区域内时，p 属于 q 的可能最近邻集 P_n. 该区域称为查询对象 p 的支配区域，记为 Dom(p).

证明. 假设不存在该区域. 令 p' 为 P_n 中任意对象，当 q 无限接近 p 时，均存在 $d(\dot{p}',q)+r_{p'} < d(\dot{p},q)-r_p$，因此 p' 无限接近 p. 这意味着 P_n 中的所有对象都与 p 重合，显然这是不成立的.

所以对任意 p，必然存在区域 Dom(p). 证毕.

算法 4. 支配区域生成.

输入：查询对象 p，查询对象集 P，障碍区域 S

输出：支配区域 Dom(p)

1. Dom(p) $\leftarrow S$，$bound \leftarrow \infty$
2. Foreach($p' \in P, p' \neq p$)
3. If($bound < d(p',p) - r_{p'}$)
4. Return
5. 生成 p 针对 p' 的近分割边 $b_{p,p'}$
6. 将 Dom(p)中位于 $b_{p,p'}$ 外的部分去掉
7. $bound \leftarrow d\dot{p}', p + r_{p'}$

对任意对象 p，定理 3 指出的 Dom(p)具有如下性质

$$\mathrm{Dom}(p) = \{x \mid d(\dot{p}',x) - r_{p'} > d(\dot{p},x) + r_p\} \quad p' \in P,\ p' \neq p \tag{9}$$

即对 Dom(p)中的任何一点 x 及 P 中其他任何对象 p'，点 $\dot{p}$ 到 x 的距离的最大值小于 $\dot{p}'$ 到 x 距离的最小值. 结合 3.2.1 节的讨论可知，Dom(p)的边界由对象 p 和其它对象之间的分割边构成. 且该分割边为 p 与其它对象 p' 每一对分割边中距离对象 p 较近的一条，下文将这种距离 p 较近的分割边称为 p 针对 p' 的近分割边.

Dom(p)生成算法中 p 针对其它对象的近分割边的计算需调用算法 3.

算法 4 中 $bound$ 的作用与算法 1 作用相同，用于对不必要计算的对象进行剪枝. 算法第 5 行对 P 中每个未被剪枝的对象调用算法 3 生成分割边并用该分割边对 Dom(p)切割，逐渐逼近 Dom(p)结果. 当所有对 p 有作用的对象 p' 都计算过后，Dom(p)取得最终结果.

将查询点 q 的可能最近邻结果集 P_n 中每个对象 p 的支配区域 Dom(p)取交集

$$\mathrm{Dom}(P) = \bigcap_{p \in P} \mathrm{Dom}(p) \tag{10}$$

则 Dom(P)即为查询点 q 的安全区域，q 在此区域内运动时，其最近邻结果集 P_n 保持不变.

3.2.3 安全区域索引

采用四分树的方式对安全区域进行索引保存. 四分树每个内部结点都有 4 个子结点，每个内部结点将当前空间均等的分为 4 个区域. 设 T_θ 为结点对应索引区域内安全区域所占面积比的阈值，即如果该索引区域内某安全区域所占面积比超过 T_θ，即可近似认为该索引区域全部都是该安全区域. 如果某区域包含的任何安全区域所占比例都小于 T_θ，那么就把该区域继续均等的划分为 4 个区域，依此类推，

直到每个区域都包含一个所占比例大于 T_θ 的安全区域数为止.

四分树每个叶子节点上都存储了该节点对应的安全区域的最近邻结果集. 查询时只需根据查询点的位置到四分树进行遍历，即可以 $O(\log n)$ 的复杂度取得最近邻结果集.

3.3　不规则不确定区域及不确定查询点

当查询对象 p 的不确定区域 U_p 不规则时，可以通过构造圆形区域 C_p，使其满足 $U_p \subset C_p$ 并且 $r(C_p) \leq \forall\, r(C'_p)$，$U_p \subset C_p$，其中 $r(C_p)$ 代表圆 C_p 的半径. 即 C_p 为包含不确定区域 U_p 的最小圆形区域. 在计算时，以 C_p 代替 U_p 进行最近邻查询及安全区域生成.

这种采用最小包含圆形区域替换的方法会对最终查询结果引入误差，即可能将本来没有可能成为查询点 q 最近邻的结果包括到结果集内. 因此结果集 P_n 生成后需要对 P_n 中的结果采用真实不确定区域进行复查.

以行驶的汽车为例，由于 GPS 定位误差，它作为查询点的位置是不精确的. 当查询点 q 位置不确定时，设区域 U_q 为 q 的不确定区域. 观察算法 1 知，该算法支持对不确定查询点的计算，即以区域 U_q（U_q 为圆形）或包含 U_q 的最小外接圆（U_q 为其它非规则区域）的圆心 $\dot{q}$ 和半径 c_q 代入算法 1，即可解出 q 的最近邻.

在对不确定查询点 q 进行安全区域检测时，将与 U_q 存在重合部分的安全区域的最近邻结果集取并集，所得结果即为对不确定查询点 q 的最近邻结果集.

综上，本文提出的障碍空间中不确定对象的最近邻查询方法可以应用到不规则不确定区域及不确定查询点的情况.

4　实验结果与分析

实验主要考察提出方法的预处理性能和查询处理效率. 实验运行环境为 Xeon E5580×2 CPU、32GB 内存和运行 64 位 Windows XP 系统的 HP Z800 工作站. 本文方法采用 C++语言实现，Visual Studio 2008 编译(/GL，/O2 优化).

实验空间设定为 20000×20000 单位的平面. 采用 R*-tree 作为平面内不确定对象的存储结构. 采用两个真实数据集 Rivers 和 Lakes① 作为障碍物集（见图 3 和图 4）. 用 R-tree Portal 提供的数据生成软件②生成不确定对象数据集，每个对象具有一个圆形不确定区域，并在该区域内服从高斯分布. 这些对象采用均匀分布和 Zipf 分布两种形式分布在实验空间中，其中 Zipf 分布的斜相关系数设为 $\alpha = 0.85$. 查询点移动路径分为直线路径和随机路径两种. 直线路径当查询点接触障碍物后按镜面反射角度继续直线前进；随机路径每前进 50 单位距离后随机变换前进方向.

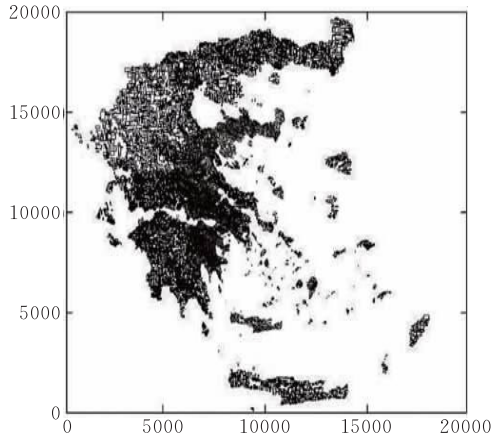


图 3　Rivers

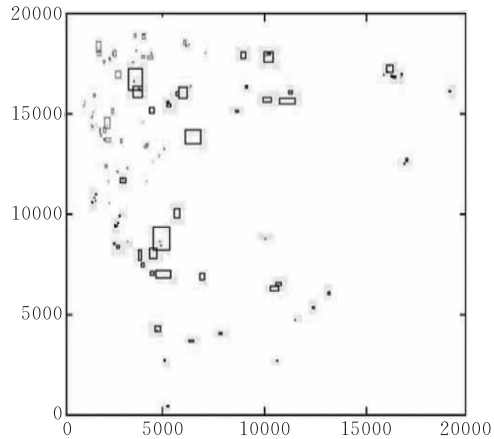


图 4　Lakes

实验通过变化障碍物个数 $|O|$（从数据集 Rivers 和 Lakes 中随机选择）、查询对象与障碍物比例 $|P|/|O|$、不确定区域大小 radius 等变量，考察了提出方法的 I/O 代价、预处理（即生成安全区域）时间、运行时间、通信代价等. 为保证实验结果的准确，每组实验令对象采用两种分布查询点和两种移动路径产生 4 种组合并取其平均值.

图 5 给出了不同障碍物个数 $|O|$ 的 I/O 代价的实验结果. I/O 代价随障碍物个数 $|O|$ 的增大逐渐增大，并且 Quad-tree 实现方式效率明显好于

① Rivers 和 Lakes 数据集可以从 R-tree Portal 网站下载 (http://www.rtreeportal.org)

② http://www.rtreeportal.org/software/SpatialDataGenerator.zip

R-tree 实现. 这是由于 Quad-tree 对点在区域内的检测方式优于 R-tree.

图 6 给出了不同障碍物个数|O|的查询时间的实验结果. 与 I/O 代价类似,查询时间随障碍物个数|O|的增大逐渐增大,并且 Quad-tree 实现方式效率明显好于 R-tree 实现.

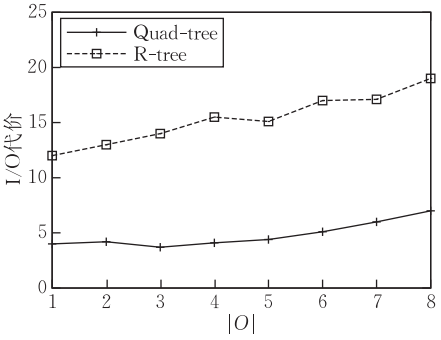


图 5 不同|O|值的 I/O 代价

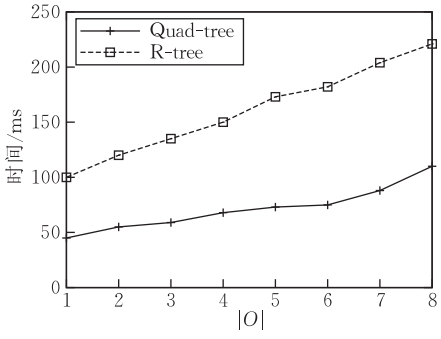


图 6 不同|O|值的时间代价

图 7 给出了不同的对象障碍物比|P|/|O|的预处理时间的实验结果. 查询时间随|P|/|O|的变化先减少后增加,这是因为当|P|相对较大时,安全区域划分就相对较小,安全区域个数相对较大,因此需要较长的计算时间. 而当|O|较大时,安全区域的边界受障碍物的影响较大,因而计算边界的代价较大,因此也需要较长的计算时间.

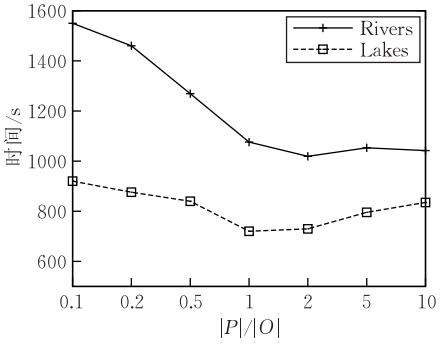


图 7 不同|P|/|O|的查询时间

图 8 给出了不同的对象障碍物比|P|/|O|的查询通信代价的实验结果. 查询通信代价随|P|/|O|的变化不明显,并且 Lakes 数据集的结果明显优于 Rivers 数据集的结果. 这是由于 Lakes 数据集中障碍物的数量少于 Rivers 数据集,从而导致 Lakes 数据集中安全区域较大,查询点移动时穿过安全区域边界(需要与服务器通信)的次数较少. 图 9 给出了不同不确定对象半径的预处理时间代价实验结果. 预处理时间随着不确定对象的半径增长,预处理时间也不断增长. 这是由于不确定对象半径增加造成对象之间可能长度的计算复杂程度增加,进而造成整体预处理时间的增长. Lakes 数据集的结果明显优于 Rivers 数据集的结果,是因为对该数据集的处理数量明显小于后者.

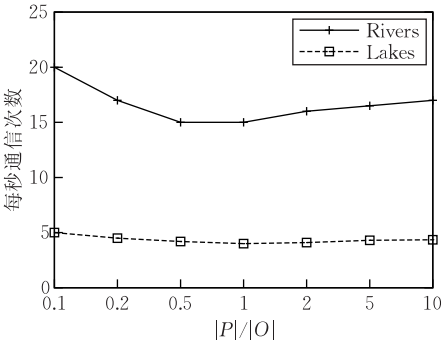


图 8 不同|P|/|O|的通信代价

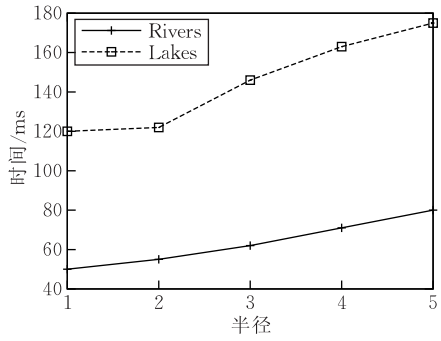


图 9 不同对象半径的时间代价

图 10 给出了不同不确定对象半径和通信代价实验结果. 从图中可以看出,由于通信代价与不确定半径无关,不同半径对通信代价几乎没有影响.

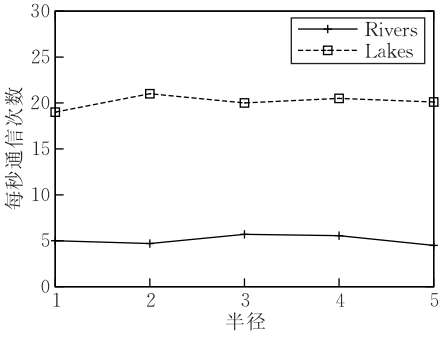


图 10 不同对象半径的通信代价

图 11 给出了本文方法与传统的非基于安全区域的最近邻查询方法的执行时间对比. 本文方法在执行时间上优于传统方法，并且在障碍物增多时远远优于传统方法. 这是因为本文提出的方法节省了很多次查询操作，当查询点没有移出安全区域时，本文方法可以直接返回结果，而不需要与服务器通信. 只有当查询点越过安全区域边界时，本文方法才需要向服务器请求新的查询. 图 12 给出了本文方法与传统的非基于安全区域的最近邻查询方法的通信代价对比. 本文方法在通信代价上远远优于传统方法，原因同上.

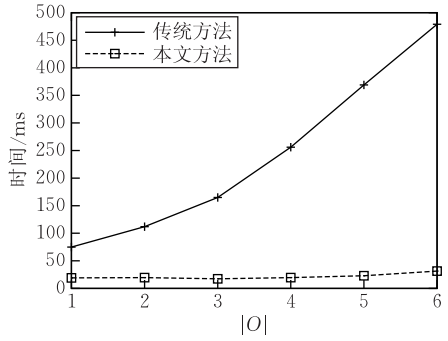


图 11　与传统算法对比时间代价

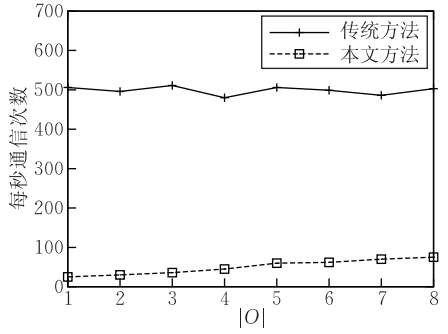


图 12　与传统算法对比通信代价

5　结　论

本文研究障碍空间中不确定对象连续最近邻查询的处理方法. 基于空间障碍物及不确定对象的数据模型，形式化地提出障碍空间中不确定对象最近邻查询问题，并设计了一种高效的基于障碍空间距离的算法来进行查询处理，其中运用了一种剪枝技术来提高性能. 本文提出了不确定对象分割区域的概念，并以之为基础设计一种有效的安全区域生成方法. 实验结果表明，本文所提出的方法具有良好的效率和可扩展性.

参　考　文　献

[1] Tao Y, Zhang J, Papadias D, Mamoulis N. An efficient cost model for optimization of nearest neighbor search in low and medium dimensional spaces. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(10): 1169-1184

[2] Berchtold Stefan, Ertl Bernhard, Keim Daniel A, Kriegel Hans-Peter, Seidl Thomas. Fast nearest neighbor search in high-dimensional space//Proceedings of the ICDE. Bombay, India, 1998: 215-226

[3] Benetis R, Jensen C S, Kariauskas G, Altenis S. Nearest and reverse nearest neighbor queries for moving objects. VLDB Journal, 2006, 15(3): 229-249

[4] Roussopoulos Nick, Kelley Stephen, Vincent Frederic. Nearest neighbor queries//Proceedings of the SIGMOD. Minneapolis, Minnesota, 1995: 71-79

[5] Okabe Atsuyuki, Boots Barry, Sugihara Kokichi, Chiu Sung Nok. Spatial Tessellations. Hoboken, USA: John Wiley & Sons, Inc, 2000

[6] Nutanong Sarana, Zhang Rui, Tanin Egemen, Kulik Lars. The V* Diagram: A query dependent approach to moving *k*NN queries//Proceedings of the VLDB. Auckland, New Zealand, 2008: 1095-1106

[7] Zhou Ao-Ying, Jin Che-Qing, Wang Guo-Ren, Li Jian-Zhong. A survey on the management of uncertain data. Chinese Journal of Computers, 2009, 32(1): 1-16(in Chinese)
(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)

[8] Li Jian-Zhong, Yu Ge, Zhou Ao-Ying. Requirements and challenges of the management of uncertain data. Communications of the China Computer Federation, 2009, 5(4): 6-14 (in Chinese)
(李建中，于戈，周傲英. 不确定性数据管理的要求与挑战. 中国计算机学会通讯，2009，5(4)：6-14)

[9] Cheng Reynold, Xie Xike, Yiu Man Lung, Chen Jinchuan, Sun Liwen. UV-Diagram: A voronoi diagram for uncertain data//Proceedings of the ICDE. Long Beach, California, USA, 2010: 796-807

[10] Yuen Sze Man, Tao Yufei, Xiao Xiaokui, Pei Jian, Zhang Donghui. Superseding nearest neighbor search on uncertain spatial databases. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(7): 1041-1055

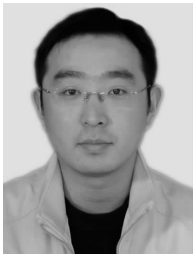
[11] Cheng Reynold, Chen Jinchuan, Mokbel Mohamed, Chow Chi-Yin. Probabilistic verifiers: Evaluating constrained nearest-neighbor queries over uncertain data//Proceedings of the ICDE. Cancun, Mexico, 2008: 973-982

[12] Gao Yunjun, Zheng Baihua. Continuous obstructed nearest neighbor queries in spatial databases//Proceedings of the SIGMOD. Providence, Rhode Island, USA, 2009: 577-590

[13] Gao Yunjun, Zheng Baihua, Chen Gencai, Lee Wang-Chien, Lee Ken C K, Li Qing. Visible reverse k-nearest neighbor query processing in spatial databases. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1314-1327

[14] Gao Yunjun, Zheng Baihua, Chen Gencai, Lee Wang-Chien, Lee Ken C K, Li Qing. Visible reverse k-nearest neighbor queries//Proceedings of the ICDE. Shanghai, China, 2009: 1203-1206

[15] Li Chuanwen, Gu Yu, Li Fangfang, Chen Mo. Moving k-nearest neighbor query over obstructed regions//Proceedings of the APWeb. Buscan, Korea, 2010: 29-35



LI Chuan-Wen, born in 1982, Ph.D.. His major research interests include spatio-temporal data management and complex event processing.

GU Yu, born in 1981, Ph. D., associate professor. His major research interests include spatio-temporal data management, RFID data management and data stream.

LI Fang-Fang, born in 1977, Ph. D., lecturer. Her major research interests focus on sensor spatio-temporal data management.

YU Ge, born in 1962, Ph. D., professor, Ph. D. supervisor. His major research interests include database theory and technology.

Background

This paper studies the continuous nearest neighbor query by the existence of obstacles and the uncertainty of data. In recent years, there has been a growing need for location-based services(LBS), ranging from resource tracking to personal life assistance. The Moving k Nearest Neighbor (MkNN) query is a main problem in the LBS area, which retrieves the top k nearest neighbors while the query consumer moving. Considerable attention is attracted to this query type and a large number of studies are proposed recently. However, most of these studies focus on ideal Euclidean plane where any two points are visible. In other words, these works rarely consider obstacles in the space.

In this paper, the authors consider the continuous kNN query over fuzzy objects in obstructed spaces. The authors also offer the users freedom to choose the confidence level on which the kNN set is required.

The moving k-nearest neighbor query over obstructed regions for certain objects was proposed by Li (2010). In this paper we extend the k-NN query of our previous work to uncertain objects.

This work is supported by the National Natural Science Foundation of China (60773220, 60933001).