室内概率阈值反向最近邻查询

王 丽 秦小麟 许建秋
(南京航空航天大学计算机科学与技术学院 南京 210016)

摘 要 室内空间变得越发的庞大和复杂,随之产生了越来越多的室内空间查询需求。目前已有文献提出了针对室内空间环境的范围查询和最近邻查询,而作为常见的空间查询类型的反向最近邻查询,尚未有相关的研究。为此,提出了室内概率阈值反向最近邻查询和基于定位设备的设备可达图模型。在图模型基础上,提出了室内概率阈值反向最近邻查询处理算法,该算法由基于图模型的批量剪枝、基于室内距离的剪枝、基于概率的剪枝和概率计算4部分构成,通过剪枝策略修剪掉不可能出现在结果集中的对象,从而缩小了查询空间,提高了效率。

关键词 室内空间,反向最近邻,设备可达图模型,查询处理
中图法分类号 TP311.13 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.1.045

Probabilistic Threshold Reverse Nearest Neighbor Queries for Indoor Moving Objects

WANG Li QIN Xiao-lin XU Jian-qiu
(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract Indoor spaces are becoming increasingly large and complex, and more and more demand for initiating indoor spatial queries has emerged. Range queries and nearest neighbor queries specifically for indoor space have been proposed, while there are no studies on reverse nearest neighbor queries. Thus, this paper presented the formal definition of probabilistic threshold reverse nearest neighbor query (IPRNN) for indoor moving objects and a device reachable graph model. The query processing algorithm of IPRNN was proposed based on the graph model. The algorithm consists of four parts: model pruning, indoor distance pruning, probability pruning and probability calculation. The objects that are not likely to appear in the result set are built out through pruning strategy, thereby significantly reducing the search space and improving efficiency.

Keywords Indoor space, Reverse nearest neighbor, Device reachable graph model, Query processing

1 引言

在日常生活中,人们会花费大量的时间在办公楼、购物中心、会展中心、机场等各种室内空间[1],而现在的室内空间也变得越发的庞大和复杂,例如,北京金源时代购物中心占地18.2公顷,建筑面积68万平方米,设有大型室内停车楼,顾客可开车到达各个楼层,车位近万个,有各类商铺1020家,每日客流量可达数万人次[2]。越来越多的用于查询感兴趣的地点或者感兴趣的移动对象的室内空间查询需求随之产生。反向最近邻查询(Reverse Nearest Neighbor Query, RNN)作为常见的空间查询,在混合现实游戏、室内安全控制、决策支持、数据挖掘等方面有很多的应用。

现有的RNN查询技术并不适用于室内空间:1)因为室内定位技术的特殊性[3,4,7,8]。室外定位技术如GPS,可以连续精确地获得对象的位置信息,而室内定位技术如RFID、蓝牙等,基于邻近分析技术,无法获知移动对象的准确位置和运动方向,只有当对象进入到定位设备的覆盖范围时,才能够被监测到。因而,室内移动对象的位置信息是离散的,具有较大程度的不确定性。2)由于室内空间距离度量的特殊性。室内环境往往是由房间、走廊、楼梯、门等多种实体组成,由于诸多室内空间实体的约束,不能使用传统的欧氏距离及空间网络距离作为室内距离的度量标准[5,6,8,9]。因此,有必要提出一种针对室内空间的RNN查询,从而既能考虑到室内移动对象的不确定性,又能兼顾到室内距离度量的特殊性。

本文提出了一种针对RFID定位技术的室内空间的RNN查询——室内概率阈值反向最近邻查询(Indoor Probabilistic threshold RNN Query, IPRNN)。IPRNN采用不确定区域[10]的概念表示室内移动对象的不确定性,同时采用最短室内移动距离(Minimal Indoor Walking Distance, MIWD)[10]作为距离度量标准,符合室内空间的实际场景。本文提出了基于定位设备的设备可达图模型,该图模型不仅保持了定位设备间的拓扑关系,还能够对室内距离进行粗粒度的表示。在图模型的基础上,提出了一种室内符号空间的概率阈值反向最近邻查询算法。该算法主要由4部分构成:基于图模型

到稿日期:2014-03-03 返修日期:2014-06-04 本文受国家自然科学基金项目(61373015,61300052),国家教育部高等学校博士学科点博导基金资助项目(20103218110017),中央高校基本科研业务费专项项目(NP2013307)资助。

王 丽(1989-),女,硕士,主要研究方向为移动对象建模与查询处理,E-mail:wxl_nuaa@sina.com;秦小麟(1953-),男,教授,博士生导师,主要研究方向为分布式数据管理与安全;许建秋(1982-),男,副教授,主要研究方向为空间数据库、移动对象数据库。

的批量剪枝、基于室内距离的剪枝、基于概率的剪枝及概率计算。通过剪枝策略修剪掉那些没有可能出现在结果集中的对象，从而缩小查询空间，提高效率。最后通过实验，对所提方法的有效性和高效性进行了验证。

2 相关工作

2.1 RNN查询

现有的基于确定数据的RNN查询大多都是在欧氏空间中，也有特定的基于路网空间、高维空间、度量空间等的研究。而基于不确定数据的PRNN(Probabilistic Reverse Nearest Neighbor Query)查询，目前研究得还不是很多，而且都集中在欧氏空间中。

Korn[11]首次提出了RNN查询的概念，并且提出了基于预计算的RNN查询算法。基于预计算的算法的主要思想是：对于每个数据对象 p，首先找 p 的NN，然后形成一个圆心在 p 点，p 到它的最近邻的距离为半径的RNN限定圆，并通过RNN树对限定圆进行索引。对于一个查询点为 q 的RNN查询，可以通过遍历索引结构找出包含 q 的所有RNN限定圆，所有这些圆对应的数据对象都是 q 的RNN。

Stanoi[12]提出了第一个不需预计算的RNN算法，算法基于以下性质：通过3条相互成60度角并相交于查询点 q 的直线将查询空间分为6个相等的扇区 s_i ($1\leqslant i\leqslant 6$)，证明可知每个扇区 s_i 中最多只有1个反向最近邻，则整个空间最多存在6个 q 的反向最近邻。从而将RNN问题转化为NN问题，即在过滤阶段分别在6个扇区 s_i 寻找 q 的NN，即仅获得6个候选对象，再在精炼阶段检查这6个候选对象是否以 q 作为它的最近邻。

TPL算法[13]是当前性能最好的RNN算法。TPL算法的基本思想是通过查询点 q 与对象点 p 之间的线段的垂直平分线将空间分成两个半平面，则 p 所在的半平面肯定不包含 q 的反向最近邻，可以裁剪掉。通过这样不断地重复，使得查询的空间区域越来越小。

2009年Chen[14]和Cheema[15]分别首次提出了基于连续概率分布和基于离散分布的PRNN查询算法。Bernecker[16]提出了PRNN算法的通用框架，指出PRNN查询处理过程一般由对象近似、空间剪枝、概率剪枝和精炼4个步骤构成，而文献[14]和文献[15]的算法同样适用于该框架。

2.2 室内建模与查询

一些研究学者提出了三维模型，如文献[17]中提出了度量拓扑模型，该模型既能描述空间单元的几何形状，又能够表示单元间的拓扑关系，可以很好地用于路径寻找。文献[18]提出了3D Poincare Duality模型，其主要通过对偶转换，将三维空间内各个区域的连通关系转换为对偶空间，可用于多层建筑中紧急逃生路线的规划。这些三维模型更多地关注拓扑关系而非定量的室内距离，并不支持常见的空间查询。

现有的图模型[6,19-21]多是用顶点表示室内单元，移动对象在图模型中的位置定位也只能精确到单元的粒度，无法实现对室内移动对象更细粒度的定位。如文献[21]中的混合空间模型以房间作为定点，门作为边。室内移动对象的位置与室内定位设备的位置密切相关，因此本文提出了基于定位设备的设备可达图模型，该图模型以定位设备作为顶点，用定位设备的位置近似表示移动对象的位置，位置表示更为精确，粒度也更细。文献[19]中Jensen等人根据RFID阅读器在室内的部署情况，建立了一个基于定位设备的部署图，但是它只考虑了划分型设备[19]的情况，对于存在型设备[19]没有进行说明。

2009年Yang等人提出了室内移动对象的连续范围查询[3]，给定一个室内空间范围，随着时间的变化，连续范围查询能够连续不断地报告当前时刻出现在该空间范围内的移动对象。在随后的研究中，其又提出了室内移动对象的概率阈值 k 最近邻查询[10]，给定一个查询发出点，一个 k 最近邻查询返回距离该查询发出点最近的 k 个移动对象。

但反向最近邻查询作为一种常见的空间查询，在室内空间中尚无相关的研究。

3 设备可达图模型

定位设备(RFID阅读器)需要被部署在室内空间，并且每个定位设备覆盖一部分空间。本节提出了设备可达图模型，它可以将定位设备的信息统一集成表示在图模型中。为了方便讨论，一种可能的室内定位设备部署如图1所示。在介绍设备可达图之前，先介绍与该图模型相关的定义。

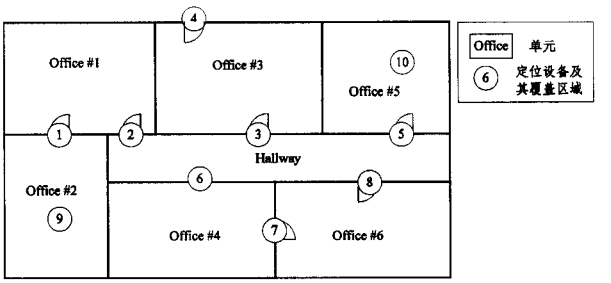

图1 设备部署图

定义1(设备可直达) 如果一个移动对象从定位设备 v_i 的覆盖范围离开后，在不被其他设备探测到的情况下可以到达另外一个定位设备 v_j 的覆盖范围，则称设备 v_i 可直达设备 v_j，表示为 $v_i \rightarrow v_j$。

例如图1中，2号设备可直达1、3、5、6、8号设备，但却无法直达9号设备，因为一个移动对象在从2号设备的覆盖范围离开后，若想到达9号设备，必然要被1号设备探测到。为使问题简化，本文暂不考虑单向门的情形，因此，若有 $v_i \rightarrow v_j$，那么有 $v_j \rightarrow v_i$。

定义2(设备可达图) 设备可达图(Device Reachable Graph)是一个无向图，形式化定义为 $G_{devr}=(V,E)$：

1) $V=\sum devices$ 是顶点的集合，每个顶点对应一个定位设备。

2) E 是边的集合。若 $v_i \rightarrow v_j$ 且 $v_j \rightarrow v_i$，那么 v_i，v_j 间存在一条边 $\{v_i, v_j\}$。其中 $v_i, v_j \in V$，且 $v_i \neq v_j$。

基于图1的设备部署图，可得如图2所示的设备可达图。图中顶点为定位设备，连接两个顶点的边则表示移动对象可以在不被其他设备探测到的情况下，从头节点表示的设备覆盖范围移动到尾节点表示的设备覆盖范围。

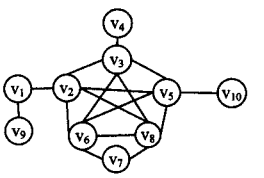

图2 设备可达图

文献[3]中根据移动对象所在的位置,将移动对象分为激活和非激活两种状态。若一个移动对象在 t 时刻被定位设备探测到,那么称该对象在 t 时刻处于激活状态;反之,若该移动对象在 t 时刻没有被任何定位设备探测到,则称该移动对象在 t 时刻处于非激活状态。与之相对应,本文对室内定位设备进行了状态划分。

定义 3(设备空闲) 若在 t 时刻,设备 v_i 没有探测到任何一个移动对象,则称 v_i 在 t 时刻处于空闲状态。

定义 4(设备忙碌) 若在 t 时刻,设备 v_i 至少探测到一个移动对象,则称 v_i 在 t 时刻处于忙碌状态。

在设备可达图和设备状态的基础上,提出了步长及忙碌步长的概念,用于表示粗粒度的室内距离。

定义 5(步长) 在设备可达图中,从顶点 v_i 到达顶点 v_j 所经过的最少顶点数目称为 v_i 到 v_j 的步长。

定义 6(忙碌步长) 在设备可达图中,从顶点 v_i 到达顶点 v_j 所经过的最少忙碌顶点数目称为 v_i 到 v_j 的忙碌步长。

例如在图 3 中,v_5,v_9 均为空闲设备,其余设备为忙碌设备,从 v_1 到达 v_{10} 需要经过的最少顶点数目为 3,因而 v_1 到 v_{10} 的步长为 3,但 v_1 到 v_{10} 的忙碌步长为 2,因为从 v_1 到 v_{10},最少要经过 v_1、v_2 这两个忙碌设备。本文中符号的相关说明如表 1 所列。

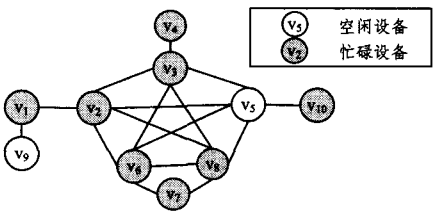

图 3 带有设备状态的设备可达图

表 1 符号说明

Notation	Meaning
v,v_i	定位设备
p,q	室内位置
loc(v_i)	设备 v_i 的部署位置
r(v_i)	设备 v_i 覆盖范围的半径
o,o_i	室内移动对象
$v_i \rightarrow v_j$	设备 v_i 可直达设备 v_j
V(busystep$\rightarrow v_i$=1)	到 v_i 的忙碌步长等于 1 的设备集合
dist(p,q)	p q 两点间的欧氏距离
d_{MIW}(p,q)	p 到 q 的最短室内移动距离
UR(o_i)	当前时刻对象 o_i 的不确定区域[10]

4 室内概率阈值反向最近邻查询

本节首先给出室内概率阈值反向最近邻查询(IPRNN)的定义,然后给出 IPRNN 的查询处理算法,该算法主要由基于设备可达图的批量剪枝、基于室内距离的剪枝、基于概率的剪枝和概率计算 4 个阶段构成。

定义 7(室内概率阈值反向最近邻查询) 给定一个室内移动对象的集合 $O=\{o_1,o_2,\cdots,o_n\}$ 和一个实数阈值 $T\in(0,1]$,一个在时刻 t 位置 q 发起的室内概率阈值反向最近邻查询返回结果集 $R=\{o_i \mid o_i\in O\wedge P_{IPRNN}(q,o_i)\geqslant T\}$,其中 o_i 是一个室内移动对象,并且 o_i 以 q 为最近邻的概率 $P_{IPRNN}(q,o_i)$ 大于给定阈值 T。式(1)给出了 $P_{IPRNN}(q,o_i)$ 的计算公式。

$$P_{IPRNN}(q,o_i)=E(\Pr\{\forall o'\in O\backslash\{o_i\},d_{MIW}(q,o_i)\leqslant d_{MIW}(o_i,o')\})$$

$$=\int_{r_1}^{r_2}(\Pr\{d_{MIW}(q,o_i)=r\}\cdot\prod_{\forall o'\in O\backslash\{o_i\}}\Pr\{d_{MIW}(o_i,o')\geqslant r\})\mathrm{d}r \quad (1)$$

式中,r_1、r_2 分别表示 q 到 $UR(o_i)$ 的室内距离的最小值和最大值,$d_{MIW}(\cdot)$是室内距离函数。

对于确定数据的 RNN 查询,一个直接的解决方案就是预计算所有移动对象的 NN,然后比较该对象到查询点 q 的距离与到其 NN 的距离哪一个较小,但是这种方法需要遍历全部的移动对象,导致计算代价较高。为了减小计算代价,研究者们提出了很多的剪枝策略[12,13],但这些剪枝策略都是针对确定数据的。

类似地,对于 IPRNN 查询,遍历所有移动对象的方法效率极低,并且由于数据对象本身的不确定性使得现有的 RNN 技术不适用于 PRNN,而由于室内空间的复杂性和距离度量的特殊性,现有的 PRNN 技术[14-16]对于 IPRNN 同样也是不适用的。因此,本文设计了针对 IPRNN 的特定的剪枝策略来减小概率计算阶段需要计算的移动对象数目。

4.1 基于设备可达图的批量剪枝

在基于 RFID 的室内符号定位中,移动对象 o_i 的位置与 RFID 阅读器是密切相关的,在任意时刻,每一个定位设备都与一个移动对象的集合相关联。根据以上观察,本节提出了基于设备可达图模型的批量剪枝策略,即根据 RFID 阅读器设备之间的空间拓扑关系及设备状态,通过设备剪枝,进而批量剪枝掉与该设备相关的移动对象。

q 为查询点,设 q 恰在 v_q 设备的覆盖范围内。

批量剪枝策略 1 若 $v_i\in V(busystep\rightarrow v_q\geqslant 2)$,那么 v_i 内所有激活状态的移动对象可以被剪枝,最后出现在 v_i 内的所有非激活状态的移动对象也可以被剪枝。

例如在图 3 中,假设查询点 q 在 v_9 设备的覆盖范围内,那么距离 v_9 的忙碌步长等于 1、2、3、4 的设备集合,分别为 $V(busystep\rightarrow v_9=1)=\{v_1\}$,$V(busystep\rightarrow v_9=2)=\{v_2,v_5\}$,$V(busystep\rightarrow v_9=3)=\{v_3,v_6,v_8,v_{10}\}$,$V(busystep\rightarrow v_9=4)=\{v_4,v_7\}$,即 $V(busystep\rightarrow v_9\geqslant 2)=V(busystep\rightarrow v_9=2)\cup V(busystep\rightarrow v_9=3)\cup V(busystep\rightarrow v_9=4)=\{v_2,v_3,v_4,v_5,v_6,v_7,v_8,v_{10}\}$。

对于 v_2 来说,v_1 是 v_2 到 v_9 的路径上必须经过的忙碌顶点,因为 v_1、v_2 均处于忙碌状态,那么 v_1、v_2 的覆盖范围内至少分别含有一个移动对象 o_1、o_2,显然 $d_{MIW}(o_2,o_1)<d_{MIW}(o_2,q)$,所以 o_2 必然不是 q 的 RNN,可以被剪枝。同理,处于 $\{v_2,v_3,v_4,v_6,v_7,v_8,v_{10}\}$设备范围内的所有激活对象均可以被剪枝。

假设存在有一个非激活态的移动对象 o_{2i},o_{2i} 最后出现在 v_2 的覆盖范围内,那么对于 o_{2i} 来讲,$d_{MIW}(o_{2i},o_1)<d_{MIW}(o_{2i},q)$,所以 o_{2i} 必然不是 q 的 RNN,可以被剪枝。最后出现在 $\{v_2,v_3,v_4,v_5,v_6,v_7,v_8,v_{10}\}$ 中的所有非激活态的对象同理可以被剪枝。

批量剪枝策略 2 若 $v_i\in V(busystep\rightarrow v_q=1)$且 v_i 在 t 时刻处于空闲状态,那么最后出现在 v_i 中的非激活状态的移动对象可以被剪枝。

在图 3 中,假设查询点 q 在 v_1 设备的覆盖范围内,

$V(busystep\rightarrow v_1=1)=\{v_2,v_5\}$，即 $v_5\in V(busystep\rightarrow v_1=1)$ 且 v_5 在 t 时刻处于空闲状态，v_2 为 v_5 到达 v_1 路径上经过的忙碌顶点，那么 v_2 设备的覆盖范围中必然存在有至少一个移动对象 o_2。假设存在有一个非激活态的移动对象 o_5，o_5 最后出现在 v_5 的覆盖范围内，那么对于 o_5 来讲，$d_{MIW}(o_5,o_2)<d_{MIW}(o_5,q)$，所以 o_5 必然不是 q 的 RNN，可以被剪枝。最后出现在 v_5 覆盖范围中的其他非激活态的移动对象同理可以被剪枝。

批量剪枝策略 3　$v_i\in V(busystep\rightarrow v_q=1)$，$v_i$ 在 t 时刻处于忙碌状态且 v_i 内的激活态的移动对象数目≥2，若 v_i 到 v_q 的最近距离大于 v_q 覆盖范围的直径，那么 v_i 内所有激活态的移动对象可以被剪枝。

查询点 q 不在某定位设备的覆盖范围时，可以在设备可达图中添加一个虚拟顶点 v_q，批量剪枝策略同上。

4.2　基于室内距离的剪枝

对于候选对象 o_i，若存在一个对象 o_j $(o_j\neq o_i\wedge o_j\in O)$ 使得 $\mathrm{Max}d_{MIW}(o_i,o_j)<\mathrm{Min}d_{MIW}(o_i,q)$，则 o_i 可以被剪枝掉。

其中 $\mathrm{Min}d_{MIW}(o_i,q)=\mathrm{Min}(\forall p\in UR(o_i),d_{MIW}(p,q))$，$\mathrm{Max}d_{MIW}(o_i,o_j)=\mathrm{Max}(\forall p\in UR(o_i)\ \forall p'\in UR(o_j),d_{MIW}(p,p'))$。

如图 4 所示，q 是走廊中的一点，候选对象 o_i 位于设备 v_5 的覆盖范围中，设备 v_{10} 的覆盖范围内存在一个对象 o_j，那么 $\mathrm{Min}d_{MIW}(o_i,q)=d_1-r(v_5)=dist(q,loc(v_5))-r(v_5)$，$\mathrm{Max}d_{MIW}(o_i,o_j)=d_2+r(v_5)+r(v_{10})=dist(loc(v_5),loc(v_{10}))+r(v_5)+r(v_{10})$，若 $\mathrm{Max}d_{MIW}(o_i,o_j)<\mathrm{Min}d_{MIW}(o_i,q)$，则候选对象 o_i 必然不属于结果集，可以将其修剪掉。

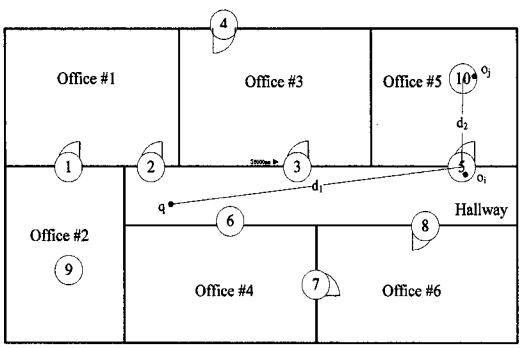

图 4　基于距离的剪枝示例

4.3　基于概率的剪枝

文献[14]中提出了一个 post-processing 步骤，用于减少概率计算阶段所需遍历的移动对象数目。该步骤为候选集合中的每一个候选对象 o_i 定义了一个后处理区域 PR(Post-processing Region)，只有那些不确定区域[10]与 o_i 的 PR 相交的对象才会对 $P_{IPRNN}(q,o_i)$ 产生影响。

类似地，本文提出了 IPRNN 的概率限制区域 PBR (Probability Bounding Region)的概念。q 为查询点，o_i 是一个候选对象，$\odot(o_i)$ 是 o_i 不确定区域的最小外接圆，o_i 对应的 PBR 定义为一个以 $\odot(o_i)$ 的圆心为圆心、以 $\mathrm{Max}d_{MIW}(q,o_i)+r(\odot(o_i))$ 为半径的圆，如图 5 所示。显然，PBR 具有与文献[14]中的 PR 相同的性质，即只有那些不确定区域与 PBR 相交的对象才有可能是 o_i 的 NN，才会对 $P_{IPRNN}(q,o_i)$ 产生影响。因此，在最后通过式(1)计算候选对象 o_i 的概率 $P_{IPRNN}(q,o_i)$ 的时候，只需要考虑那些与 o_i 的 PBR 相交的对象。

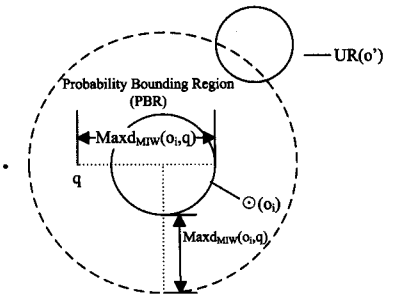

图 5　概率限制区域

对于候选对象 o_i，假设存在有移动对象集合 $O'=\{o_1',o_2',\cdots,o_m'\}\subseteq O$，$\forall o\in O'$ 使得 $UR(o')\cap PBR(o_i)\neq\emptyset$，那么 o'、q 相比，o' 距离 o_i 更近的概率 $P_{Nearer}(o_i,o'<q)$ 满足如下条件：

$$P_{Nearer}(o_i,o'<q)<P_{PBR}(o_i,o')$$

其中，$P_{PBR}(o_i,o')=\dfrac{PBR(o_i)\cap UR(o')}{UR(o')}$。

因为

$$\begin{aligned}P_{Nearer}(o_i,o'<q)&=P(d_{MIW}(o',o_i)\leqslant d_{MIW}(q,o_i))\\&<P(d_{MIW}(o',o_i)\leqslant \mathrm{Max}d_{MIW}(q,o_i))\\&<P(\mathrm{Min}d_{MIW}(o',o_i)\leqslant \mathrm{Max}d_{MIW}(q,o_i))\\&=P_{PBR}(o_i,o')\end{aligned}$$

那么 o'、q 相比，q 距离 o_i 更近的概率为

$$\begin{aligned}P_{Nearer}(o_i,q<o')&=P(d_{MIW}(q,o_i)\leqslant d_{MIW}(o',o_i))\\&=1-P(d_{MIW}(o',o_i)\leqslant d_{MIW}(q,o_i))\\&=1-P_{Nearer}(o_i,o'<q)\\&>1-P_{PBR}(o_i,o')\end{aligned}$$

所以，q 与 $O'=\{o_1',o_2',\cdots,o_m'\}$ 相比，q 均更近的概率为

$$\begin{aligned}P_{NN}(o_i,q)&=P_{IPRNN}(q,o_i)\\&=\prod_{k=1}^{m}P_{Nearer}(o_i,q<o_k')\\&>\prod_{k=1}^{m}(1-P_{PBR}(o_i,o_k'))\end{aligned}$$

根据上述不等式，我们可以得到 $P_{IPRNN}(q,o_i)$ 的上下界：

$$P_{IPRNN}(q,o_i).lower=\prod_{k=1}^{m}(1-P_{PBR}(o_i,o_k'))$$

$$P_{IPRNN}(q,o_i).upper=1-P_{IPRNN}(q,o_i).lower$$

因而对于候选对象 o_i，若 $P_{IPRNN}(q,o_i).lower\geqslant T$，那么 o_i 必然是属于 IPRNN 的结果集。而若 $P_{IPRNN}(q,o_i).upper<T$，那么 o_i 必然不属于 IPRNN 结果集，可以被剪枝。

室内概率阈值反向最近邻查询处理算法 MDP 如算法 1 所示，算法根据计算代价和剪枝效率对 3 种剪枝策略进行了有效集成，首先通过基于图模型的批量剪枝策略修剪掉绝大部分的对象，并将剩余对象加入候选对象集 S_{cnd}(第 2 行)，其次进行基于室内距离的剪枝策略(第 3－6 行)和基于概率的剪枝策略(第 7－12 行)，逐步缩减候选对象集的规模，提高查询效率。最后根据式(1)对候选对象集合 S_{cnd} 中每一个候选对象进行概率计算，从而获得最终的结果集 R。

算法 1(室内概率阈值反向最近邻 IPRNN 查询处理算法 MDP)

输入：查询点 q，查询时刻 t，概率阈值 T，移动对象集合 O

输出：室内概率阈值反向最近邻查询 IPRNN 结果集 R

1. R←∅ and candidate set S_{cnd}←∅
2. apply the Model Pruning rule and add all the object left to S_{cnd}//基于图模型的剪枝
3. for each $o_i \in S_{cnd}$ do
4. for each $o_j \in O/\{o_i\}$ do
5. if $Maxd_{MIW}(o_i,o_j) < Mind_{MIW}(o_i,q)$ Then
6. delete o_i from S_{cnd}//基于室内距离的剪枝
7. for each $o_i \in S_{cnd}$ do
8. calculate the $P_{IPRNN}(q,o_i).lower$ and $P_{IPRNN}(q,o_i).upper$
9. if $P_{IPRNN}(q,o_i).lower \geqslant T$ Then
10. add o_i to R
11. if $P_{IPRNN}(q,o_i).upper < T$ Then
12. delete o_i from S_{cnd}// 基于概率的剪枝
13. for each $o_i \in S_{cnd}$ do
14. compute $P_{IPRNN}(q,o_i)$
15. if $P_{IPRNN}(q,o_i) \geqslant T$ Then
16. add o_i to R// 概率计算
17. Return R

5 实验结果与分析

为了验证本文所提方法的有效性，我们对其进行了实验分析。实验中的数据集由数据生成器模拟生成，模拟的室内环境共 4 层、186 个室内单元（包括办公室、走廊、楼梯、厕所），共部署有 345 个 RFID 阅读器。每个房间通过门与走廊相通，两个楼梯直接与走廊相连，移动对象可以在单元内部移动，也可以移动到与之相连的其它单元。

本文的实验环境为：3.30GHz CPU，4.0GB 内存，Windows 7 操作系统。实验中用到的相关参数如表 2 所列。实验随机选取 100 个室内位置作为查询出发点，结果为 100 次查询的平均值。

表 2 模拟数据集生成参数

参数类型	参数设置
移动对象个数	2k,4k,6k,8k,10k,15k,20k,25k,…,60k
概率阈值 T	0.1,0.3,0.5,0.7,0.9
查询次数	100

为了检验基于图模型的批量剪枝策略的效果，记录了候选对象的减少比例，即$|O'|/|O|$，结果如图 6 所示。其中$|O|$表示室内移动对象的数目，$|O'|$表示经过基于图模型的批量剪枝策略修剪后的候选对象数目。当$|O|=2k$时，即室内空间一共有 2000 个移动对象，90%左右的移动对象通过基于图模型的批量剪枝策略保留到了$|O'|$内。当室内移动对象的数目增多时，$|O'|/|O|$呈下降趋势。

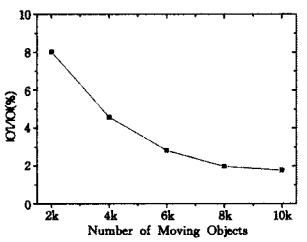

图 6 基于图模型的批量剪枝策略

实验接下来对基于室内距离和基于概率的剪枝策略的效果进行衡量。为了验证基于室内距离的剪枝策略的修剪效果，将经过了基于室内距离的剪枝策略修剪后的候选对象数目$|O'|$报告在图 7 中，分析发现随着室内移动对象数目的增多，$|O'|$也增多。为了验证基于概率的剪枝策略的修剪效果，图 8 中报告了概率阈值 $T=0.5$ 和 $T=0.9$ 时，经过 3 步剪枝策略后候选对象数目$|O'|$的对比图。实验结果表明，当阈值 T 较大时，剪枝效果更好。

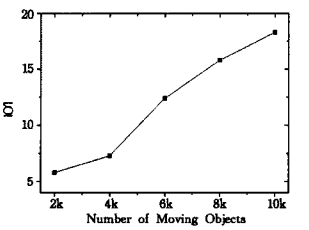

图 7 基于室内距离的剪枝策略

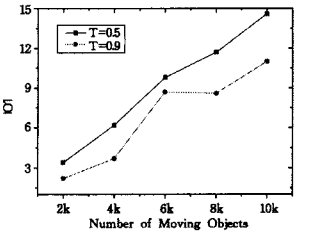

图 8 基于概率的剪枝策略

图 9 和图 10 分别报告了 $T=0.1$ 时，依次经过基于图模型的剪枝、基于室内距离的剪枝、基于概率的剪枝后候选对象的数目和概率计算阶段确认的对象数目以及 4 步骤对应的时间代价。可以看出，基于图模型的剪枝策略修剪效果最为显著，且具有最小的时间代价；同时本文提出的 3 种剪枝策略是有效的。

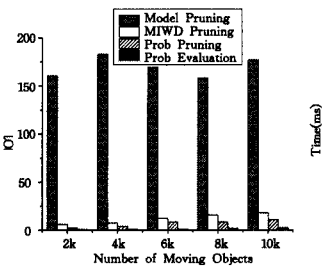

图 9 4 步骤结果图

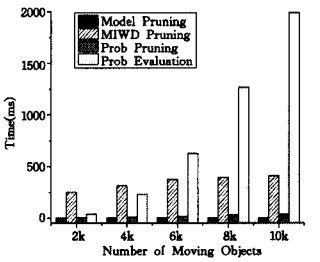

图 10 4 步骤时间对比图

为了验证查询算法的优越性，将 MDP 算法与遍历所有移动对象的 Naive 方法进行实验对比，图 11 显示了两种算法的时间代价。从实验结果图可以看出，MDP 算法优于遍历的方法，且基本上维持了 20 倍以上的加速比。为了验证查询算法的可扩展性，我们将数据规模从 10k 扩展到 60k，观察时间的变化趋势，如图 12 所示。实验结果表明，随着移动对象数目的增多，3 步剪枝策略处理时间增长趋于平缓，概率计算阶段时间增长较快，同时表明在总的查询处理时间中，概率计算阶段时间所占比重逐步增大。

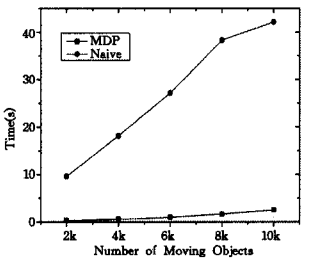

图 11 MDP 算法与 Naive 算法的对比

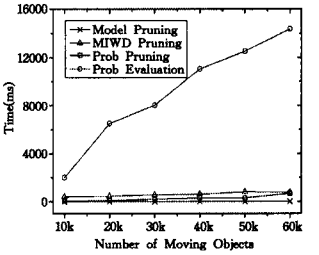

图 12 查询处理扩展图

结束语 本文针对室内环境，研究了室内概率阈值反向最近邻查询 IPRNN，提出了可以有效支持 IPRNN 查询的设备可达图模型，并在此基础上提出了 3 种有效的剪枝策略：基于图模型的批量剪枝策略、基于室内距离的剪枝策略和基于概率的剪枝策略。实验结果表明了 IPRNN 通过 3 步剪枝可缩减查询空间，提高查询效率。

（下转第 214 页）

LP. Barcelona, Spain, 2004: 412-418

[4] Dang Yan, Zhang Yu-lei, Chen Hsin-chun. A lexicon enhanced method for sentiment classification: An experiment on online product reviews[J]. IEEE Intelligent Systems, 2010, 25(4): 45-53

[5] Liu Tie-yan, Yang Yi-ming, Wan Hao, et al, Support Vector Machines Classification with Very Large Scale Taxonomy [J]. SIGKDD Explorations, Special Issue on Text Mining and Natural Language Processing, 2005, 7(1): 36-43

[6] Cortes C, Vapnik V. Support-Vector Networks [J]. Machine Learning, 1995, 20(3): 273-297

[7] Boser B E, Guyon I M, Vapnik V N. A training algorithm for optimal margin classifiers[C]// Haussler D. ed. 5th Annual ACM Workshop on COLT. Pittsburgh, PA, ACM Press, 1992: 144-152

[8] 张雯,张化祥. 属性加权的朴素贝叶斯分类器[J]. 计算机工程与应用,2010,46(29)

[9] Zhang H. The Optimality of Naive Bayes[C]// FLAIRS 2004 Conference. 2004: 562-567

[10] 缪凯,赵志刚. RBF 神经网络的研究与应用[D]. 青岛:青岛大学,2007

[11] 张义超,卢英,李炜. RBF 网络隐含层结点的优化[J]. 计算机技术与发展,2009,19(1)

[12] Breiman L. Random Forests [J]. Machine Learning, 2001, 45(1): 5-32

[13] 张洪强,刘光远,赖祥伟. 随机森林算法在肌电的重要特征选择中的应用[J]. 计算机科学,2013,40(1): 200-202

[14] 孙秋秋. "好"在语义上的模糊性与确定性[J]. 辽宁大学学报:哲学社会科学版,1982(1): 70-76

[15] 刘康,王素格,廖祥文,等. 第四届中文倾向性分析评测总体报告[R]. 第四届中文倾向性分析评测论文集,2012: 1-32

[16] Wu Y F, Jin P. SemEval 2010 Task 18: Disambiguating Sentiment Ambiguous Adjectives[C]// Proceedings of the 2010 Evaluation Exercises on Semantic Evaluation. 2010: 81-85

(上接第 205 页)

参 考 文 献

[1] Jensen C S, Li K J, Winter S. The other 87%: a report on The Second International Workshop on Indoor Spatial Awareness (San Jose, California-November 2, 2010) [J]. SIGSPATIAL Special, 2011, 3(1): 10-12

[2] http://finance.sina.com.cn/money/consume9/20041025/10361104770.shtml

[3] Yang B, Lu H, Jensen C S. Scalable continuous range monitoring of moving objects in symbolic indoor space[C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management. ACM, 2009: 671-680

[4] Lu H, Yang B, Jensen C S. Spatio-temporal joins on symbolic indoor tracking data[C]// 2011 IEEE 27th International Conference on Data Engineering (ICDE). IEEE, 2011: 816-827

[5] Yuan W, Schneider M. Supporting continuous range queries in indoor space[C]// 2010 Eleventh International Conference on Mobile Data Management (MDM). IEEE, 2010: 209-214

[6] Lu H, Cao X, Jensen C S. A foundation for efficient indoor distance-aware query processing[C]// 2012 IEEE 28th International Conference on Data Engineering (ICDE). IEEE, 2012: 438-449

[7] Yu J, Ku W S, Sun M T, et al. An RFID and particle filter-based indoor spatial query evaluation system[C]// Proceedings of the 16th International Conference on Extending Database Technology. ACM, 2013: 263-274

[8] Xie X, Lu H, Pedersen T B. Efficient distance-aware query evaluation on indoor moving objects[C]// ICDE. 2013

[9] 甘早斌,袁永光,赵贻竹,等. 基于 DR-tree 的室内移动对象索引研究 [J]. 计算机科学,2012,39(10): 177-181

[10] Yang B, Lu H, Jensen C S. Probabilistic threshold k nearest neighbor queries over moving objects in symbolic indoor space [C]// Proceedings of the 13th International Conference on Extending Database Technology. ACM, 2010: 335-346

[11] Korn F, Muthukrishnan S. Influence sets based on reverse nearest neighbor queries[J]. ACM SIGMOD Record, 2000, 29(2): 201-212

[12] Stanoi I, Agrawal D, El Abbadi A. Reverse Nearest Neighbor Queries for Dynamic Databases[C]// ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. 2000: 44-53

[13] Tao Y, Papadias D, Lian X. Reverse kNN search in arbitrary dimensionality[C]// Proceedings of the Thirtieth international conference on Very large data bases-Volume 30. VLDB Endowment, 2004: 744-755

[14] Lian X, Chen L. Efficient processing of probabilistic reverse nearest neighbor queries over uncertain data[J]. The VLDB Journal—The International Journal on Very Large Data Bases, 2009, 18(3): 787-808

[15] Cheema M A, Lin X, Wang W, et al. Probabilistic reverse nearest neighbor queries on uncertain data[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(4): 550-564

[16] Bernecker T, Emrich T, Kriegel H P, et al. Efficient probabilistic reverse nearest neighbor query processing on uncertain data[J]. Proceedings of the VLDB Endowment, 2011, 4(10): 669-680

[17] Whiting E, Battat J, Teller S. Topology of urban environments [M]// Computer-Aided Architectural Design Futures (CAAD-Futures) 2007. Springer Netherlands, 2007: 114-128

[18] Lee J. 3D GIS for geo-coding human activity in micro-scale urban environments [M] // Geographic Information Science. Springer Berlin Heidelberg, 2004: 162-178

[19] Jensen C S, Lu H, Yang B. Graph model based indoor tracking [C]// Tenth International Conference on Mobile Data Management: Systems, Services and Middleware, 2009 (MDM'09). IEEE, 2009: 122-131

[20] Stoffel E P, Schoder K, Ohlbach H J. Applying hierarchical graphs to pedestrian indoor navigation[C]// Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2008: 54

[21] Lorenz B, Ohlbach H J, Stoffel E P. A hybrid spatial model for representing indoor environments[M]// Web and Wireless Geographical Information Systems. Springer Berlin Heidelberg, 2006: 102-112