

一种基于概率模型的预测性时空区域查询处理^{*}

张 炜¹, 李建中^{1,2+}, 刘 禹¹

¹(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

²(黑龙江大学 计算机科学与技术学院,黑龙江 哈尔滨 150080)

A Probabilistic Model Based Predictive Spatio-Temporal Range Query Processing

ZHANG Wei¹, LI Jian-Zhong^{1,2+}, LIU Yu¹

¹(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

²(School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China)

+ Corresponding author: Phn: +86-451-86415827, E-mail: lijzh@hit.edu.cn, <http://db.hit.edu.cn>

Zhang W, Li JZ, Liu Y. A probabilistic model based predictive spatio-temporal range query processing. *Journal of Software*, 2007,18(2):279-290. <http://www.jos.org.cn/1000-9825/18/279.htm>

Abstract: A probabilistic approach is proposed, which adopts filter-refinement framework for query processing. First, all objects that possibly satisfy a query are retrieved as candidate results. Then, probabilities that the candidates will satisfy the query are evaluated based on a probability model proposed in the paper. Finally, a user defined minimum probability threshold is used to filter unqualified candidates to get a final predictive result. The future location of a moving object is defined as a random variable in the probability model. Two modes are proposed to describe object's movement status in spatio-temporal query range, and the corresponding methods are presented to compute the probability that an object will satisfy the query in the proposed modes. A trajectory analyzing algorithm is proposed to estimate the probability density functions (PDF) from the historical trajectories. An index structure is designed to efficiently support the storing and accessing of the PDFs. The experimental result shows that the proposed solution can effectively process the predictive spatio-temporal range query and improve the correctness of the predictive results. It is suitable for processing the query with small spatial range and long-term future time interval.

Key words: spatial database; query processing; index method; location-dependent and sensitive

摘 要: 提出了一种基于概率模型的预测性时空区域查询处理方法.该方法采用 Filter-Refinement 方式来处理查询.首先,从数据库中选择所有可能满足查询的候选移动对象;然后,根据概率模型中定义的方法来计算候选移动对象满足查询的概率;最后,根据查询中指定的最小概率阈值过滤候选移动对象并返回查询结果.该概率模型

* Supported by the National Natural Science Foundation of China under Grant Nos.60473075, 60533110 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.2006CB303000 (国家重点基础研究发展计划(973)); the Program of the Ministry of Education of China for New Century Excellent Talents in University (NCET) (国家教育部新世纪优秀人才支持计划); the Natural Science Foundation of Heilongjiang Province of China under Grant No.ZJG03-05 (黑龙江省自然科学基金); the Heilongjiang Province Fund for Yong Scholars of China under Grant No.QC06C033 (黑龙江省青年科技专项资金)

Received 2005-10-27; Accepted 2006-04-05

将移动对象未来可能出现的位置定义为一个随机变量,并给出了计算移动对象在两种不同的运动模式下满足查询的概率值的方法.还提出了一种通过对大量历史轨迹抽样来获得概率密度函数(probability density function, 简称 PDF)的轨迹分析算法,并设计了概率密度函数索引 STP-Index(spatio-temporal PDF-index).该索引能够有效地提高轨迹分析算法和概率计算的效率.实验结果表明,该查询处理方法能够有效地支持预测性时空区域查询的处理,提高查询结果的正确性,特别适合于具有较小的空间区域和长时间范围的预测性时空区域查询.

关键词: 空间数据库;查询处理;索引方法;位置相关和敏感

中图法分类号: TP311 文献标识码: A

随着无线网络、传感器网络、移动定位等相关技术的发展,移动对象数据库系统引起了人们的关注.移动对象数据库系统管理大量移动对象的位置信息,在基于位置的信息服务、军事等领域有着广泛的应用.预测性时空区域查询是移动对象数据库应用中常见的查询,用于预测在用户关注的未来时空范围内出现的移动对象信息.处理预测性时空区域查询的关键问题是如何预测移动对象未来可能出现的位置.预测方法的准确程度直接影响查询结果的正确性.目前,在预测性时空区域查询处理技术中采用的预测方法主要有如下 3 种:

(1) 基于匀速直线运动规律的预测方法.这种方法利用移动对象最近报告的位置和速度矢量,根据匀速直线运动规律来预测其未来可能出现的位置^[1-4].因为移动对象在实际运动中很少会长期保持速度矢量不变,所以,基于匀速直线运动的预测方法仅能用于预测移动对象在未来很短时间内的运动趋势,且预测的精度很低.

(2) 基于线性递推函数的预测方法^[5].这种方法利用移动对象最近 k 个时刻的历史位置,使用一个线性递推函数来预测移动对象未来出现的位置.该方法需要每个移动对象都具有一定的计算能力和存储资源.

(3) 基于概率模型的预测方法^[6-9].这种方法假设移动对象在未来可能出现的空间区域内服从某种概率分布,并利用这种概率分布计算移动对象满足查询的概率.这些方法都假设移动对象的位置在其未来可能出现的空间区域中服从均匀分布或正态分布,并利用相应概率密度函数来计算概率.随着预测时间的增加,移动对象可能出现的区域也会逐渐增大,均匀分布或正态分布的假设与实际概率分布的误差也将逐步增大.而且,查询区域越小,概率分布函数的误差对概率计算结果正确性的影响越大.由于无法保证这些概率分布假设的正确性,采用基于概率模型的预测方法计算的概率值往往无法准确地体现移动对象满足查询的可能性.

综上所述,目前的预测性时空区域查询处理方法存在预测精度低、预测时间短、概率计算不准确等问题.这些问题将阻碍预测性时空区域查询的广泛应用.此外,目前基于概率模型查询处理技术的研究主要关心未来时刻的预测性时空查询的处理,很少考虑未来一段时间内预测性时空区域查询的处理.

针对上述问题,本文提出了一种基于概率模型的预测性时空区域查询处理方法.本文把移动对象在未来查询空间区域内的运动模式分为 Sometimes 和 Always 两种. Sometimes 模式表示移动对象至少在未来时间段内的某个时刻出现在查询区域中. Always 模式表示移动对象在未来时间段内一直在查询区域中运动.本文考虑的预测性时空区域查询的定义是:给定未来时间范围、空间范围、运动模式和最小概率阈值等条件,获取满足这些条件的移动对象.本文采用 Filter-Refinement 方式处理预测性时空查询,与其他方法相比,它具有如下特点:(1) 对移动对象的运动方式没有作任何限制,使得该方法更适用于处理远期预测性时空区域查询;(2) 把移动对象未来可能出现的位置定义为一个随机变量,对该随机变量的概率密度函数的类型不作任何特殊假设,使得该方法更具一般性,适用范围更广.模拟实验表明,本文提出的方法克服了目前预测性时空区域查询处理技术的缺点,并且能够有效地提高查询结果的准确性,尤其适合于较为长期的未来预测性时空区域查询的处理.

1 问题的定义和查询处理方法概述

本文假设移动对象在有限二维矩形空间 $S=[L_1, L_2] \times [U_1, U_2]$ 内运动.本文的结果可以直接推广到任意维空间.移动对象定时向服务器报告其位置信息.设 t_n 表示移动对象最近一次报告位置信息的时刻, $O.P(t)=(O.P_x(t), O.P_y(t))$ 表示移动对象 O 在 t 时刻的位置.移动对象在相邻两次信息报告间隔期间的位置由线性插值函数来确定.为了便于描述,下面,本文使用大写字母来表示位置坐标,例如 X 表示位置 (x, y) .

定义 1(时变空间区域). 时变空间区域 $MR(R, V_R, t, t_s)$ 是大小和位置随时间变化的矩形空间区域, 其中, 矩形区域 $R=[x_1, x_2] \times [y_1, y_2]$ 是时变空间区域在时间点 t_s 的初始位置; t 是时间变量, $t \geq t_s$, $V_R = \{v_l, v_r, v_u, v_d\}$, 其中, v_l 和 v_r 是 R 的左、右边界水平移动速度; v_u 和 v_d 是 R 的上、下边界垂直移动的速度. $MR(R, V_R, t, t_s) = [x_1 + v_l(t - t_s), x_2 + v_r(t - t_s)] \times [y_1 + v_d(t - t_s), y_2 + v_u(t - t_s)]$.

移动对象 O 在未来时间范围 $T=[t_l, t_h]$ 内在时变空间区域 $MR(R, V_R, t, t_l)$ 中的运动模式可分为两种: 持续模式(always) 和非持续模式(sometimes). 持续模式表示 O 在 T 期间保持在 $MR(R, V_R, t, t_l)$ 中运动, 即 $\{\forall t \in T, O.P(t) \in MR(R, V_R, t, t_l)\}$, 而非持续模式表示 O 至少在 T 期间的某一时刻 t 出现于 $MR(R, V_R, t, t_l)$ 中, 即 $\{\exists t \in T, O.P(t) \in MR(R, V_R, t, t_l)\}$. 若 T 是某个未来的时刻 t_f , 则两种运动模式等价. 本文将移动对象未来出现在查询区域看作是一个随机事件, 并基于概率模型来处理预测性时空区域查询.

定义 2. 令 $Mode$ 是 *Always* 或 *Sometimes*, p_i 是 O_i 在 $T=[t_l, t_h]$ 期间以 $Mode$ 模式运动于 $MR(R, V_R, t, t_l)$ 内的概率, 基于概率模型的预测性时空区域查询为 $PRQ(T, MR(R, V_R, t, t_l), Mode, MinC) = \{(O_i, p_i) | p_i \geq MinC\}$, 其中, $MinC$ 是用户给定的最小概率阈值.

本文研究的问题是如何有效地处理基于概率模型的预测性时空区域查询, 并获得正确性较高的查询结果. 关键问题是如何获得可能满足查询的移动对象以及如何表示和计算移动对象满足查询的概率. 本文把移动对象在未来时刻出现的位置定义为一个随机变量, 称为时空随机变量. 移动对象在未来时刻满足预测性时空区域查询的概率可以直接根据时空随机变量的概率密度函数来计算. 本文在处理关于未来时间范围 T 的预测性时空区域查询时, 首先将 T 划分为多个较短的时间段, 然后分别计算移动对象在每个划分点时刻出现在查询区域中的概率, 最后用已计算的概率值来估计移动对象以不同的运动模式在 T 期间出现在查询区域内的概率. 需要说明的是, 本文的预测方法对移动对象在 $[t_u, t_l]$ 期间的运动方式没有任何限制(如匀速直线、加速运动), 比其他方法更适于处理较长未来时间范围的时空查询.

本文使用 Filter-Refinement 的方式来处理预测性时空区域查询. 查询处理过程包括 3 个步骤: (1) 获取候选移动对象: 根据移动对象最近报告的位置信息从移动对象数据库中选出所有可能满足查询条件的移动对象作为候选对象; (2) 概率计算: 根据移动对象最近报告的位置信息和查询关注的预测时刻选择相应的时空随机变量的概率密度函数, 针对查询指定的运动模式计算移动对象满足查询条件的概率; (3) 产生查询结果: 根据用户给定的最小概率阈值 $MinC$, 确定满足查询条件概率大于 $MinC$ 的移动对象的集合, 得到查询结果. 为了提高查询处理的效率, 本文使用了两种索引结构: TPMR-Tree(time-parameterized max speed rectangle tree)^[10] 存储移动对象最近报告的位置及其最大速度, STP-Index(spatio-temporal PDF-index) 记录了轨迹分析方法获得的概率密度函数. 前者用于提高候选移动对象的获取效率, 后者用于提高概率计算的效率. 轨迹分析算法通过对大量历史轨迹的随机抽样来获得移动对象在不同时间后出现的位置, 并利用这些位置来估计时空随机变量的概率密度函数. 在概率计算过程中使用的概率密度函数分为两类: 一类是由轨迹分析方法获得的概率密度函数; 另一类是通用概率密度函数(如均匀分布). 前者是进行概率计算的主要依据, 只有在 STP-Index 中找不到计算所需的概率密度函数时才使用通用概率密度函数.

2 预测性时空区域查询的处理

2.1 候选结果移动对象的获取

给定查询 $PRQ(T, MR(R, V_R, t, t_l), Mode, MinC)$, 所有在 T 期间可能出现在 $MR(R, V_R, t, t_l)$ 内的移动对象称为候选结果移动对象集, 简称候选对象集. 虽然移动对象未来可能出现的位置难以被准确预测, 但是, 人们可以根据移动对象的最大运动速度来估计它在未来时刻可能出现的空间区域. 设移动对象 O 最近 t_u 时刻报告的位置为 $O.P(t_u)$, 且其最大速度为 $O.V_{max}$, 则 O 在未来时刻 t_f 可能出现的空间范围就是以 $O.P(t_u)$ 为圆心和以 $O.V_{max}(t_f - t_u)$ 为半径的圆形区域. 下面, 本文用 $PLS(O, t)$ 表示移动对象 O 在未来时刻 t 可能出现的空间范围. 当查询中 $Mode = Always$ 时, 如果在 T 期间的任何时刻 t , $PLS(O, t) \cap MR(R, V_R, t, t_l) \neq \emptyset$, 则 O 是候选对象; 当 $Mode = Sometimes$ 时, 如果在 T 期间存在某一时刻 t_f , 使 $PLS(O, t_f) \cap MR(R, V_R, t_f, t_l) \neq \emptyset$, 则 O 是候选对象.

为了有效地获得候选对象集,本文使用 TPMR-Tree 索引结构^[10]来存储移动对象最近报告的位置信息.利用 TPMR-Tree 获取查询 $PRQ(T, MR(R, V_R, t, t_i), Mode, MinC)$ 候选结果集的算法从根节点开始逐层遍历与查询空间范围 $MR(R, V_R, t, t_i)$ 在未来时间范围 T 期间有交迭区域的节点.根据非叶节点存储的最大速度可以计算它的 MBR(minimum bounding rectangle)在未来时刻覆盖的空间范围.若当前节点为非叶节点,则继续遍历以该节点为根的子树;否则,逐个检验该叶节点中存储的移动对象在未来时间范围 T 期间是否可能满足查询.若叶节点中的移动对象未来可能出现的空间区域与查询区域 $MR(R, V_R, t, t_i)$ 的交迭情况满足 $Mode$ 模式的要求,则将相应移动对象作为候选结果返回.下面是基于 TPMR-Tree 的预测性时空区域查询候选对象集的获取算法,其中 $RefineCheck(T, MR(R, V_R, t, t_i), O)$ 是文献[10]中的算法,它用于判断 T 期间 $PLS(O, t)$ 是否会与 $MR(R, V_R, t, t_i)$ 相交.

PRQuery($Tr, PRQ(T, MR(R, V_R, t, t_i), Mode, MinC)$).

输入: Tr :TPMR-Tree; $PRQ(T, MR(R, V_R, t, t_i), Mode, MinC)$:预测性时空区域查询.

输出:查询 $PRQ(T, MR(R, V_R, t, t_i), Mode, MinC)$ 的候选对象集 C .

1. If (Tr 为叶节点) $C=C \cup Check\ Leaf(Tr, PRQ(T, MR(R, V_R, t, t_i), Mode, MinC))$;
2. Else { // Tr 为非叶节点
3. For all $Node_i$ in Tr
4. If ($Node_i$ is not leaf node) $C=C \cup Check\ Nonleaf(Node_i, PRQ(T, MR(R, V_R, t, t_i), Mode, MinC))$;
5. Else $C=C \cup Check\ Leaf(Tr, PRQ(T, MR(R, V_R, t, t_i), Mode, MinC))$;
6. Return C ;

Check Leaf($Node, PRQ(T, MR(R, V_R, t, t_i), Mode, MinC)$).

输入:叶节点 $Node$, 预测性时空区域查询 $PRQ(T, MR(R, V_R, t, t_i), Mode, MinC)$.

输出: $Node$ 中的候选对象集 C .

1. $C=\emptyset$;
2. For all $O_i \in Node$
3. If ($Mode=Always$)
4. If ($(PLS(O_i, t_l) \cap MR(R, V_R, t_l, t_i) \neq \emptyset)$ and ($PLS(O_i, t_h) \cap MR(R, V_R, t_h, t_h) \neq \emptyset$)) {
5. If ($PLS(O_i, t_l) \cap MR(R, V_R, t_l, t_i) = PLS(O_i, t_l)$) and ($PLS(O_i, t_h) \cap MR(R, V_R, t_h, t_h) = PLS(O_i, t_h)$)
6. $O_i.p=1$; // O_i 在 $[t_l, t_h]$ 期间必然满足查询
7. Else $O_i.p=-1$; // 初始化 O_i 满足查询的概率
8. $C=C \cup O_i$;
9. Else // $Mode=Sometimes$
10. If ($RefineCheck(T, MR(R, V_R, t, t_i), O_i)$) {
11. If ($\exists t_f \in [t_l, t_h], PLS(O_i, t_f) \cap MR(R, V_R, t_f, t_i) = PLS(O_i, t)$) $O_i.p=1$;
12. Else $O_i.p=-1$; // 初始化 O_i 满足查询的概率
13. $C=C \cup O_i$;
14. Return C ;

$Check\ Nonleaf(Node, PRQ(T, MR(R, V_R, t, t_i), Mode, MinC))$ 与 $Check\ Leaf$ 类似,略.

2.2 概率模型和概率计算方法

定义 3(时空随机变量). 设移动对象 O 在 t_u 时刻位于 $X \in S$. 对于任意未来时刻 t 和 $\forall X \in S$, 设 $(\Omega_x(X, t), F, P)$ 是一个概率空间, 样本空间 $\Omega_x(X, t) = \{\omega | \forall O \in MOD, O.P(t_u) = X, \text{ 且 } O.P_x(t) \in [L_1, L_2]\}$, F 是 $\Omega_x(X, t)$ 的一个 σ -代数, 实值函数 P 为 $(\Omega_x(X, t), F)$ 上的概率. 对于 $\forall \omega \in \Omega_x(X, t)$, 时空随机变量 $\zeta_x(\omega, t)$ 是定义在 $\Omega_x(X, t)$ 上的单值实函数, 且对于 $\forall x \in [L_1, L_2]$, 总有 $\{\omega | \zeta_x(\omega, t) \leq x\} \in F$. $\zeta_x(\omega, t)$ 可简记为 $\zeta_x(t)$, 其中, X 称为初始位置; t 称为预测时刻.

根据定义 3, 我们可以定义二维时空随机变量. 设二维时空随机变量 $\xi_x(t) = (\zeta_x(t), \eta_x(t))$. 其中: $\zeta_x(t)$ 的样本空间是 $\Omega_x(X, t) = \{\omega | \forall O \in MOD, O.P(t_u) = X, \text{ 且 } O.P_x(t) \in [L_1, L_2]\}$; 而 $\eta_x(t)$ 的样本空间是 $\Omega_x(X, t) = \{\omega | \forall O \in MOD, O.P(t_u) = X, \text{ 且}$

$O, P_y(t) \in [U_1, U_2]$ }. 下文中的 $\xi_X(t)$ 均是二维时空随机变量.

时空随机变量及其概率密度函数可以用于表示和计算移动对象满足预测性时空区域查询的概率. 令随机事件 $MT_X(T, MR(R, V_R, t_s), Mode)$ 表示位于 X 的移动对象在未来时间范围 T 期间以 $Mode$ 模式运动于 $MR(R, V_R, t_s)$ 中, 且若 T 仅为未来时刻 t_f , 则简记为 $MT_X(t_f, MR(R, V_R, t_f, t_s))$. 设 $f_X(x(t), y(t))$ 是二维时空随机变量 $\xi_X(t)$ 的概率密度函数, 其中 $x(t)$ 和 $y(t)$ 表示 t_u 时刻位于 X 的移动对象在未来 t 时刻的 x, y 坐标, 则 t_u 时刻位于 X 的移动对象在 t_f 时刻运动到区域 R 内的概率为 $P\{MT_X(t_f, R) = P\{\xi_X(t_f) \in R\} = \int_{Rf} f_X(x(t_f), y(t_f)) dx(t_f) dy(t_f)$. 本文的概率模型是通用模型, 它未对概率密度函数的类型作任何限制, 仅需 $P\{\xi_X(t) \in S\} = \int_S f_X(x(t), y(t)) dx(t) dy(t) = 1$, 且 $P\{\xi_X(t) \notin S\} = 0$. 本文假设移动对象在较短时间范围 ΔT 期间保持匀速直线运动. 计算移动对象在未来时间范围 $T = [t_l, t_h]$ 期间满足查询 $PRQ(T, MR(R, V_R, t_l), Mode, MinC)$ 的概率的过程包括 3 步: 首先, 选定包括 t_l 和 t_h 在内的 n 个时刻, 将 T 划分为 $n-1$ 个长度不超过 ΔT 的时间段; 然后, 分别计算移动对象在每个划分时刻 t_k 出现在 $MR(R, V_R, t_k, t_l)$ 中的概率; 最后, 利用上一步获得的概率值来计算移动对象满足查询的概率.

引理 1. 给定未来时间范围 $T = [t_l, t_h] (|T| \leq \Delta T)$ 和时变空间范围 $MR(R, V_R, t_l)$, 对于 $\forall X \in S, MT_X(T, MR(R, V_R, t_l), Always)$ 当且仅当 $\{\xi_X(t_l) \in R, \text{且 } \forall t \in (t_l, t_h), \xi_X(t) \in S\} \cap \{\xi_X(t_h) \in MR(R, V_R, t_h, t_l), \text{且 } \forall t \in [t_l, t_h], \xi_X(t) \in S\}$.

引理 1 说明, 若移动对象在 t_l 和 t_h 时刻都出现在 $MR(R, V_R, t_l)$ 中, 则事件 $MT_X(T, MR(R, V_R, t_l), Always)$ 必然发生. 虽然移动对象在相邻时刻(或时间范围)可能出现的位置具有一定的相关性, 但若不限制移动对象在其他时刻的位置(即移动对象在其他时刻的位置可为 S 中的任意位置), 可证明如下结论成立:

引理 2. 给定未来时间范围 $T = [t_l, t_h]$ 和 $I = [t_m, t_n] \subset T$, 时变空间范围 $MR(R, V_R, t_l)$, 对 $\forall X \in S$, 有 $P\{MT_X(I, MR(R, V_R, t_l), Mode)\} = P\{MT_X(I, MR(R, V_R, t_m), Mode)\}$, 且 $\forall t \in (T-I), \xi_X(t) \in S$.

引理 3. 给定未来时间范围 $T = [t_l, t_h] (|T| \leq \Delta T)$ 和时变空间范围 $MR(R, V_R, t_l)$, 对 $\forall X \in S$, 设 $p_1 = P\{MT_X(t_l, R)\}, p_2 = P\{MT_X(t_h, MR(R, V_R, t_h, t_l))\}$, 则 $\text{Max}\{0, (p_1 + p_2 - 1)\} \leq P\{MT_X(T, MR(R, V_R, t_l), Always)\} \leq \text{Min}\{p_1, p_2\}$.

定理 1. 给定未来时间范围 $T = [t_l, t_h]$ 和时变空间范围 $MR(R, V_R, t_l)$, 设至少有 n 个时刻将 T 分为 $n-1$ 个长度不大于 ΔT 的时间段. 对于 $\forall X \in S$, 令 $p_1 = P\{MT_X(t_l, R)\}, p_2 = P\{MT_X(t_1, MR(R, V_R, t_1, t_l))\}, \dots, p_{n-1} = P\{MT_X(t_{n-2}, MR(R, V_R, t_{n-2}, t_l))\}, p_n = P\{MT_X(t_n, MR(R, V_R, t_h, t_l))\}$. 设

$$L(k) = \begin{cases} p_1, & k=1 \\ \text{Max}\{0, (L(k-1) + p_k - 1)\}, & 2 \leq k \leq n \end{cases}, \quad H(k) = \begin{cases} p_1, & k=1 \\ \text{Min}\{H(k-1), p_k\}, & 2 \leq k \leq n \end{cases}$$

则 $L(n) \leq P\{MT_X(T, MR(R, V_R, t_l), Always)\} \leq H(n)$.

证明: 当 $n=1$ 时, 命题显然成立. 当 $n=2$ 时, $T = [t_l, t_h] (|T| \leq \Delta T)$. 由引理 3 可知, 此时命题成立. 设 $n=k$ 时, $T = [t_l, t_h] (|T| \leq (k-1) \times \Delta T)$, 命题成立. 当 $n=k+1$ 时, $T = [t_l, t_h] (|T| \leq k \times \Delta T)$. 设事件 $E_k = \{MT_X([t_l, t_k], MR(R, V_R, t_l), Always)\}$, 且 $\forall t \in (t_k, t_h), \xi_X(t) \in S, E_{k+1} = \{MT_X(t_h, MR(R, V_R, t_h, t_l))\}$, 且 $\forall t \in [t_l, t_h], \xi_X(t) \in S$. 由 E_{k+1} 可知, $\{\xi_X(t_{k+1}) \in MR(R, V_R, t_{k+1}, t_l)\}$. 根据引理 1 易证 $MT_X([t_l, t_h], MR(R, V_R, t_l)) = E_k \cap E_{k+1}$. 由假设可知, $L(k) \leq P\{MT_X([t_l, t_{k+1}], MR(R, V_R, t_l), Always)\} \leq H(k)$. 由引理 2 得 $P\{E_{k+1}\} = P\{MT_X(t_h, MR(R, V_R, t_h, t_l))\} = p_n$. 因此, $P\{MT_X([t_l, t_h], MR(R, V_R, t_l), Always)\} \leq \text{Min}\{P\{E_k\}, P\{E_{k+1}\}\} \leq \text{Min}\{H(k), p_n\} = H(k+1)$. 因为 $P\{E_k \cap E_{k+1}\} = 1 - P\{\overline{E_k} \cup \overline{E_{k+1}}\} = P\{E_k\} + P\{E_{k+1}\} - P\{\overline{E_k} \cap \overline{E_{k+1}}\} - 1 \geq \text{Max}\{0, P\{E_k\} + P\{E_{k+1}\} - 1\} \geq \text{Max}\{0, L(k) + p_{k+1} - 1\} = L(k+1)$. 因此, $L(k+1) \leq P\{MT_X([t_l, t_h], MR(R, V_R, t_l), Always)\} \leq H(k+1)$ 成立.

引理 4. 给定未来时间范围 $T = [t_l, t_h] (|T| \leq \Delta T)$ 和时变空间范围 $MR(R, V_R, t_l)$, 对 $\forall X \in S$, 设 $p_1 = P\{MT_X(t_l, R)\}, p_2 = P\{MT_X(t_h, MR(R, V_R, t_h, t_l))\}$, 则 $\text{Max}\{p_1, p_2\} \leq P\{MT_X(T, MR(R, V_R, t_l), Sometimes)\} \leq 1$.

定理 2. 给定未来时间范围 $T = [t_l, t_h]$ 和时变空间范围 $MR(R, V_R, t_l)$, 设至少有 n 个时刻将 T 分为 $n-1$ 个长度不大于 ΔT 的时间段. 对于 $\forall X \in S$, 令 $p_1 = P\{MT_X(t_l, R)\}, p_2 = P\{MT_X(t_1, MR(R, V_R, t_1, t_l))\}, \dots, p_{n-1} = P\{MT_X(t_{n-2}, MR(R, V_R, t_{n-2}, t_l))\}, p_n = P\{MT_X(t_h, MR(R, V_R, t_h, t_l))\}$. 设 $G(k) = \begin{cases} p_1, & k=1 \\ \text{Max}\{G(k-1), p_k\}, & 2 \leq k \leq n \end{cases}$, 则 $G(n) \leq P\{MT_X(T, MR(R, V_R, t_l), Sometimes)\} \leq 1$.

证明: 当 $n=1$ 时, 命题显然成立. 当 $n=2$ 时, $T = [t_l, t_h] (|T| \leq \Delta T)$. 由引理 3 可知, 此时命题成立.

设 $n=k$ 时, $T = [t_l, t_h] (|T| \leq (k-1) \times \Delta T)$, 命题成立. 当 $n=k+1$ 时, $T = [t_l, t_h] (|T| \leq k \times \Delta T)$. 显然, $P\{MT_X(T, MR(R, V_R, t_l),$

$Sometimes\} \leq 1$ 成立. 设事件 $E_k = \{MT_X([t_k, t_{k+1}], MR(R, V_R, t_k), Sometimes), \forall t \in (t_k, t_{k+1}), \xi_X(t) \in S\}$, $E_{k+1} = \{MT_X(t_h, MR(R, V_R, t_h, t_i))\}$, 且对于 $\forall t \in [t_i, t_h], \xi_X(t) \in S$. 显然, $E_k \subseteq MT_X(T, MR(R, V_R, t, t_i), Sometimes)$, $E_{k+1} \subseteq MT_X(T, MR(R, V_R, t, t_i), Sometimes)$. 因此, $\text{Max}\{G(k), p_{k+1}\} = G(k+1) \leq P\{MT_X(T, MR(R, V_R, t, t_i), Sometimes)\} \leq 1$ 成立. 由数学归纳法可知, 命题成立.

上述两个定理给出了计算候选对象满足预测性时空区域查询 $PRQ(T, MR(R, V_R, t, t_i), Mode, MinC)$ 的概率值范围的方法. 在此基础上, 本文使用以下两个递推公式来计算移动对象满足查询的概率值. 设有 n 个时刻将时间范围 T 划分为 $n-1$ 个长度不超过 ΔT 的时间段, 则 $P\{MT_X(T, MR(R, V_R, t, t_i), Always)\} = PA_\lambda(n)$, $P\{MT_X(T, MR(R, V_R, t, t_i), Sometimes)\} = PS_\lambda(n)$, 其中, $PA_\lambda(k)$ 和 $PS_\lambda(k)$ 分别定义为

$$PA_\lambda(k) = \begin{cases} p_1 = P\{MT_X(t_i, R)\}, & k=1 \\ \text{Min}\{p_k, PA_\lambda(k-1)\} - \lambda_k \times (\text{Min}\{p_k, PA_\lambda(k-1)\} - \text{Max}\{0, (p_k + PA_\lambda(k-1) - 1)\}), & 2 \leq k \leq n \end{cases} \quad (1)$$

$$PS_\lambda(k) = \begin{cases} p_1 = P\{MT_X(t_i, R)\}, & k=1 \\ \text{Max}\{1, (\text{Max}\{p_k, PS_\lambda(k-1)\} + \lambda_k \times (1 - \text{Max}\{p_k, PS_\lambda(k-1)\}))\}, & 2 \leq k \leq n \end{cases} \quad (2)$$

在上面两个公式中, 系数 $\lambda_k \in [0, 1]$, 对于不同的 k , λ_k 可取不同的值, 也可以通过实验来确定. 它与 ΔT 的长度、时变空间范围 $MR(R, V_R, t, t_i)$ 和移动对象的运动速度有关. 因为在很短的时间段内, 移动对象只能移动很短的距离, 所以, 若 $(t_{k+1} - t_k) \rightarrow 0$, 则 $P\{MT_X([t_k, t_{k+1}], MR(R, V_R, t, t_k), Always)\} \rightarrow P\{MT_X(t_k, R)\}$, 且 $\lambda_k \rightarrow 0$. 若 $\lambda_k = 0$, 则由定理 2 可知, $PS_\lambda(n)$ 就是移动对象在划分时刻满足查询的最大概率值 $\text{Max}\{p_1, p_2, \dots, p_n\}$. 另外, 从递推公式中还可以发现: 若在计算过程中, 某个 $p_k = 0$, 则 $PS_\lambda(n) = 0$; 类似地, 若在计算过程中, 某个 $p_k = 1$, 则 $PS_\lambda(n) = 1$.

当计算移动对象在未来时刻满足查询的概率时, 我们还利用某些信息来调整概率值, 以提高预测的准确性. 设加权函数 $W(QR, p, M)$ 可根据信息 M 对利用 $f_X(x(t), y(t))$ 计算的位于 X 的移动对象 O 在未来时刻 t_f 出现在查询区域 QR 内的概率值 p 进行调整. 如果已知位于 X 的移动对象的速度矢量 V , 本文采用加权函数 $W(QR, p, V) = \text{Min}\{(p + p \times \cos(\theta)), 1\}$ 来调整已获得的概率值 p , 其中, θ 是 X 和查询区域 QR 的中心点 Q 之间的连线 XQ 和速度矢量 V 之间的夹角. 该加权函数的基本假设是移动对象将一直朝向其目的地运动. 因此, 若移动对象速度方向和查询区域中心点方向的夹角小于 90° , 则相应的概率值均有所提高, 反之则会降低.

2.3 历史轨迹分析方法和STP-Index索引结构

设移动对象的轨迹是一个空间位置序列, 记为 $Traj = tid.\{P(t_i), P(t_{i+1}), \dots, P(t_h)\}$, 其中 tid 是轨迹的标识符; $P(t_i) \in S$ 是产生轨迹 $Traj$ 的移动对象在时刻 t_i 的位置; $E_i = t_h - t_i$ 称为轨迹的持续时间.

定义 4(定位函数). 已知轨迹 $Traj = tid.\{P(t_i), P(t_{i+1}), \dots, P(t_h)\}$. 对于 $\forall t \in [t_i, t_h]$, 定位函数 $Pos(Traj, t)$ 的函数值是产生轨迹 $Traj$ 的移动对象在 t 时刻的位置, 即

$$Pos(Traj, t) = \begin{cases} g(P(t_i), P(t_{i+1}), t) \in S, & t_i \leq t \leq t_{i+1} \\ \emptyset, & t < t_i \text{ or } t > t_h \end{cases}$$

其中: $g(P(t_i), P(t_{i+1}), t)$ 为插值函数, 本文使用线性插值函数.

定义 5(空间的划分). 设 S_A 是由某种划分方法将空间 S 划分后形成的区域集合, 若 $S_A = \{a_k | a_k \subseteq S, a_k \neq \emptyset, k=1, 2, \dots, M; \forall i \neq j, a_i \cap a_j = \emptyset; \cup a_k = S\}$, 则称 S_A 为空间 S 的划分, 其中, a_k 称为划分 S_A 中的区域.

本文使用均匀网格将空间 S 划分为 m 个互不相交的区域. 令 $\xi_a(t)$ 表示所有初始位置在区域 a 且预测时刻为 t 的时空随机变量, 其中区域 a 称为初始区域. 对于未来时间范围 $T = [t_i, t_h]$, 轨迹分析算法选定包括 t_i 和 t_h 在内的 $n+1$ 个时刻, 将 T 划分为 n 个长度不超过 ΔT 的时间段 T_k . 假设初始位置在同一初始区域且预测时刻属于同一时间段 T_k 的时空随机变量具有相同的概率密度函数, 即对 $\forall a \in S_A$, 若 $X, Y \in a$, 且 $t_1, t_2 \in T_k$, 则 $f_X(x(t_1), y(t_1)) = f_Y(x(t_2), y(t_2))$. 因此, 至多需要 $(n+1) \times m$ 个时空随机变量的概率密度函数来计算移动对象满足关于未来时间范围 T 期间各种预测性时空区域查询的概率. 不同的时间和空间划分粒度决定了不同的分析精度. 在具体实现中需要综合考虑磁盘存储空间和分析精度的平衡.

定义 6(位置样本). 给定空间 S 的划分 S_A 和轨迹 $Traj = tid.\{P(t_i), P(t_{i+1}), \dots, P(t_h)\}$. 对于 $\forall a \in S_A$, 若 $\exists st \in [t_i, t_h]$ 使

$Pos(Traj, st) \in a$, 且 $\exists t > 0$ 使 $Y = Pos(Traj, t + st) \neq \emptyset$, 则三元组 $r = \langle a, t, Y \rangle$ 称为位置样本, 其中, 区域 a 称为初始区域; t 称为预测时刻; Y 称为采样位置.

轨迹分析算法通过随机抽样从大量历史轨迹中提取位置样本, 并用这些样本来估计时空随机变量的概率密度函数. 它包括 3 步: 1) 根据空间划分方法生成与每条轨迹相对应的形如 $\{\langle a_1, [t_1, t_1] \rangle, \langle a_2, [t_1, t_2] \rangle, \dots, \langle a_k, [t_{k-1}, t_h] \rangle\}$ 的区域化轨迹, 其中, $\langle a_k, [t_{k-1}, t_k] \rangle$ 表示轨迹在 $[t_{k-1}, t_k]$ 期间一直处于区域 a_k 中; 2) 参照区域化轨迹从原始轨迹中抽取位置样本; 3) 利用抽取的位置样本构造估计概率密度函数的直方图. 轨迹分析算法逐条分析历史轨迹, 结合原始轨迹及其对应的区域化轨迹来抽取位置样本. 它可以用于估计 $\xi_a(t)$ 的概率密度函数. 对于不同的未来时刻, 算法通过类似的抽样方法来获得相应的位置样本. 在抽样过程中, 轨迹分析算法同时为所有时空随机变量抽取位置样本. 下面是详细的轨迹分析算法.

Trajectory Analysis(TS, T, Δt , Δst).

输入: 轨迹数据集 $TS = \{Traj_1, Traj_2, \dots, Traj_n\}$; 预测时间范围 $T = [t_l, t_h]$; 预测时间范围切分步长 Δt ; 轨迹抽样时间间隔 Δst .

输出: 用于估计概率密度函数的直方图集合 *EPDF*.

1. $EPDF = \emptyset$; $N = \lceil (t_h - t_l) / \Delta t \rceil$;
2. For each $Traj_i$ in TS do {
3. 根据 $Traj_i$ 生成区域化轨迹 $AT_j = \{\langle a_1, [t_1, t_1] \rangle, \langle a_2, [t_1, t_2] \rangle, \dots, \langle a_k, [t_{k-1}, t_h] \rangle\}$;
4. For each $\langle a_j, [t_{j-1}, t_j] \rangle$ in AT_j do {
5. $M = \lceil (t_j - t_{j-1}) / \Delta st \rceil$;
6. For $m = 0$ to $M - 1$ do {
7. For $n = 0$ to $N - 1$ do
8. $t_k = t_l + (n \times \Delta t)$ or t_h ;
9. $st_m =$ 随机选择 $[t_{j-1} + m \times \Delta st, t_{j-1} + (m + 1) \times \Delta st]$ 期间的某一时刻;
10. $X_k = Pos(Traj_i, st_m + t_k)$; // 获取采样位置
11. If $(X_k == \emptyset)$ then break; // $(st_m + t_k) \geq t_h$, 对于任意更大的 n 都不会获得采样位置
12. If $pdf(a_j, t_k) \notin EPDF$, $EPDF = EPDF \cup new\ created\ pdf(a_j, t_k)$;
13. Else $update(pdf(a_j, t_k), X_k)$; // 根据 X_k 来更新直方图 $pdf(a_j, t_k)$
14. Return $EPDF$;

STP-Index 由一个索引时空随机变量初始区域的空间索引和多个索引时空随机变量概率密度函数的 PDF-Tree 构成. 本文在 STP-Index 的实现中采用 Grid-file 来索引时空随机变量的初始区域. Grid-file 中的每个网格单元都与一个 PDF-Tree 相对应. 网格单元中记录的信息形如 $\langle \langle C_i, I_i, A_i, ptr_i \rangle, \dots, \langle C_i, I_i, A_i, ptr_i \rangle \rangle$, 其中, 网格区域 C_i 是树指针 ptr_i 指向的 PDF-Tree 中所有时空随机变量的初始区域; 时间范围 I_i 和空间范围 A_i 分别是 PDF-Tree 中的时空随机变量的预测时间范围和取值范围的并集. PDF-Tree 具有类似于 B^+ -Tree 的结构. 它用预测时刻 t 作为索引键, 叶节点形如 $\langle \langle t_1, pdf_1 \rangle, \dots, \langle t_m, pdf_m \rangle, ptr_{next} \rangle$, 其中: t_i 是预测时刻且 $t_1 < t_2 < \dots < t_m$; pdf_i 是概率密度函数; ptr_{next} 是指向与之相邻的下一个叶结点磁盘块的指针. 每个叶结点至多存储 M 个概率密度函数. 如图 1 所示, 在初始区域 r_1 对应的描述信息中, $T = [1, 4]$ 表示在树指针指向的 PDF-Tree 中包含了预测时刻在 1~4 分钟的时空随机变量的概率密度函数, A_1 是 PDF-Tree 中所有时空随机变量的取值范围的并集.

STP-Tree 的插入和删除算法需要同时维护 Grid-file 和 PDF-Tree. 当轨迹分析算法产生一个新的概率密度函数时, 若相应的时空随机变量的初始区域对应的 PDF-Tree 并不存在, 则需要在 Grid-file 中增加该初始区域的记录, 并创建与之相关联的 PDF-Tree; 若 PDF-Tree 已经存在, 则在更新 Grid-file 中的预测时间范围 T 和样本空间并集 A 之后, 直接将概率密度函数插入该 PDF-Tree 中. 在插入过程中的结点分裂算法以及概率密度函数的删除算法均与 B^+ -Tree 类似. 当 PDF-Tree 中的所有概率密度函数都被删除以后, PDF-Tree 以及在 Grid-file 中与之关联的网格中的信息都将被删除. STP-Index 的创建过程由轨迹分析算法调用索引的插入操作来实现.

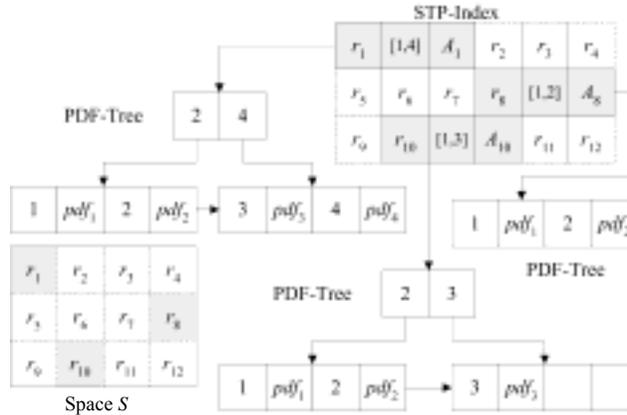


Fig.1 An example of STP-Index

图 1 STP-Index 示例

从 STP-Index 中查询概率密度函数的过程包括两步:查询时空随机变量初始区域的空间索引和查询概率密度函数索引 PDF-Tree. 查询算法首先计算在 Grid-file 中包含初始位置 $O.P(t_u)$ 的网格区域 C , 并获得 C 中的描述信息 (C, I, A, ptr) . 在访问树指针 ptr 指向的 PDF-Tree 之前, 算法首先根据 C 中的描述信息检查是否需要访问 PDF-Tree. 若 $T \cap I = \emptyset$, 则表示 PDF-Tree 中没有存储任何处理所需的概率密度函数, 查询算法利用通用概率密度函数计算概率; 若 $T \subseteq I$ 且 $\{\forall t \in T, MR(R, V_R, t, t_i) \cap A = \emptyset\}$, 则表示移动对象在 T 期间不会出现在 $MR(R, V_R, t, t_i)$ 内, 则 O 满足查询的概率为 0. 在上述两种情况下, 查询算法可以直接计算 O 满足查询的概率而无须访问 PDF-Tree; 否则, 查询算法访问 ptr 指向的 PDF-Tree 以获得概率密度函数, 并根据第 2.2 节的递推公式进行概率计算. 在计算过程中, 算法利用递推公式的性质来减少无效的概率计算, 即当 $p_k=0$ 时, 终止 $PA_{\lambda}(n)$ 的计算; 当 $p_k=1$ 时, 终止 $PS_{\lambda}(n)$ 的计算.

3 实验结果

本文利用与文献[3]相同的数据产生程序在 Illinois 的交通网络中生成模拟数据. 地图标准化为 1000×1000 的二维环境. 模拟程序的输出是一系列在交通网络上运动的移动对象的运动轨迹. 在数据产生时, 100 000 个移动对象的初始位置均匀地分布在交通网络中. 它们的运动速度在 $60 \sim 120 \text{ km/h}$ 范围内均匀分布, 且最大速度均被设定为 120 km/h . 移动对象每隔 1 分钟报告一次位置和速度信息. 数据生成程序的模拟时间为 100 分钟. 本文利用不同的随机种子基于相同参数生成两组轨迹数据, 第 1 组用于历史轨迹分析算法构造 STP-Index, 第 2 组数据用于查询处理实验和验证查询结果. 每组实验结果都是对 200 个预测性时空区域查询负载处理结果的平均值. 查询的空间区域大小分别为 100×100 和 50×50 , 覆盖了地图的 1% 和 0.25%. 查询区域的中心位置在地图内服从均匀分布. 实验运行在 P4 2.4GHz 处理器, 512MB 内存, 装有 Linux 系统的计算环境下.

3.1 查询处理的性能

历史轨迹分析算法用第 1 组数据的 100 000 条轨迹来构造 STP-Index. 在创建索引时, 划分空间的网格分辨率为 500×500 , 估计概率密度函数的直方图分辨率为 4×4 , 索引中磁盘页大小为 1KB, 算法的设置为 $T=[1, 50]$, $\Delta t=1$ 分钟. 表 1 说明查询处理方法仅需很小的 I/O 代价即可计算候选移动对象满足查询的概率.

Table 1 The I/O cost of computing probability by STP-Index

表 1 STP-Index 支持概率计算的 I/O 代价

Description	Experiment result	
	1M cache	No cache
Population of spatio-temporal random variable in STP-Index	1 419 113	
I/O cost of retrieving a PDF in STP-Index	0.3	3
I/O cost of probability calculation while $ T =5$ minutes	0.86	3.6
I/O cost of probability calculation while $ T =10$ minutes	1.7	4.8

图2比较了查询中空间区域大小对查询处理过程中概率计算 I/O 代价的影响.在实验中,所有查询的预测时间范围长度为 10 分钟,查询的空间区域大小分别为 100×100 和 50×50,Mode 分别为 Always 和 Sometimes.在图2中,(L)表示处理空间区域为 100×100 的查询时概率计算的 I/O 代价;而(S)表示处理空间区域为 50×50 的查询时概率计算的 I/O 代价.实验结果表明,在查询中的 Mode 和预测时刻相同的条件下,概率计算的 I/O 代价与查询区域的大小成正比.图3比较了预测时间范围长度对查询处理过程中概率计算 I/O 代价的影响.实验中,所有查询的空间区域大小均为 100×100,预测时间范围长度分别为 10 分钟和 5 分钟.在图3中,(10)表示处理预测时间范围长度为 10 分钟的查询时概率计算的 I/O 代价,(5)表示处理预测时间范围长度为 5 分钟的查询时概率计算的 I/O 代价.实验结果表明,概率计算的 I/O 代价与查询中预测时间范围长度成正比.

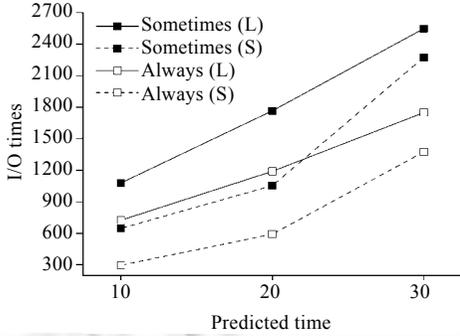


Fig.2 Effect of query's spatial range on I/O cost of probability computing

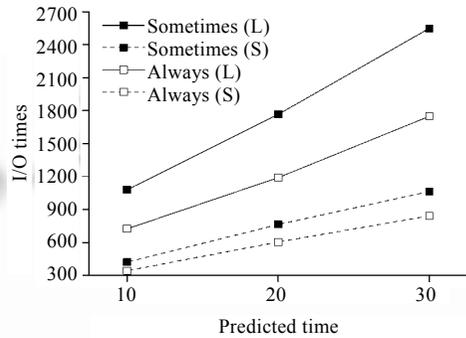


Fig.3 Effect of query's temporal range on I/O cost of probability computing

图2 查询区域大小对概率计算 I/O 代价的影响

图3 查询时间范围长度对概率计算 I/O 代价的影响

3.2 查询结果的正确性

本文使用准确率(accuracy)和召回率(recall)来评价预测性时空区域查询处理方法返回的查询结果的正确性.对于查询 $PRQ(T, MR(R, V_R, t, t_i), Mode, MinC)$, 设 CR 表示候选移动对象集合, PR_{MinC} 表示利用最小概率阈值 $MinC$ 过滤 CR 后返回的查询结果, TR 表示真正满足查询的移动对象集合.任何基于概率模型的预测性时空区域查询处理方法在计算某个移动对象 O 满足查询的概率时都可能会产生两种错误:(1) 预测 O 满足查询的概率 $p \geq MinC$, 并将其作为查询结果返回, 而实际上, O 在未来不会出现在于查询关注的时空区域中;(2) 预测 O 满足查询的概率 $p < MinC$, 于是没有将其作为查询结果返回, 而实际上, O 是满足查询的移动对象.因此, 查询结果 PR_{MinC} 中可能既包括了真正满足查询的移动对象, 又包括了一些最终在未来不会出现在查询区域中的移动对象.本文用准确率来表示 PR_{MinC} 中真正满足查询的移动对象所占的比例, 它体现了查询处理方法避免第(1)种错误的能力; 用召回率来表示 PR_{MinC} 中的正确查询结果在所有真正满足查询的移动对象中所占的比例, 它体现了查询处理方法避免第(2)种错误的能力.这两个指标的定义如下:

$$Accuracy = \begin{cases} \frac{|PR_{MinC} \cap TR|}{|PR_{MinC}|}, & PR_{MinC} \neq \emptyset \\ 0\%, & PR_{MinC} = \emptyset \text{ and } TR \neq \emptyset \\ 100\%, & PR_{MinC} = \emptyset \text{ and } TR = \emptyset \end{cases}, \quad Reall = \begin{cases} \frac{|PR_{MinC} \cap TR|}{|TR|}, & TR \neq \emptyset \\ 100\%, & TR = \emptyset \end{cases}$$

为了与其他概率方法比较, 如果根据基于匀速直线运动规律预测移动对象未来将会出现在查询区域, 则将其满足查询的概率设为 1, 反之设为 0. 为了便于描述, 在下文中使用 UN 表示利用均匀分布假设进行预测的查询处理方法; PF 表示利用本文提出的概率模型进行预测的查询处理方法; LN 表示利用匀速直线运动规律进行预测的查询处理方法; PW 表示利用 PF 方法并结合第 2.2 节提出的速度矢量加权的查询处理方法.其中, LN 方法是文献[1-4]中使用的预测方法, UN 方法是文献[6-9]中使用的预测方法.

3.2.1 查询中空间区域大小对查询结果正确性的影响

本节实验中所有的查询都是关于未来 10~20 分钟的预测性时空区域查询, 查询的空间区域大小分别为

100×100 和 50×50.空间区域越小,对查询处理过程中使用的预测方法的准确性要求越高.在图 4~图 7 中,(L)表示关于空间区域为 100×100 的查询结果的正确性,(S)表示关于空间区域为 50×50 的查询结果的正确性.

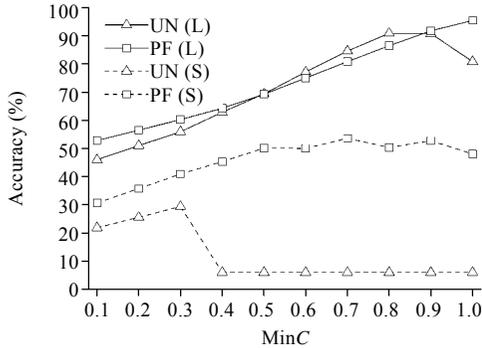


Fig.4 Accuracy of result returned by UN and PF Methods, *Mode=Always*, $T=[10,20]$

图 4 UN 和 PF 方法获得的查询结果的准确率, *Mode=Always*, $T=[10,20]$

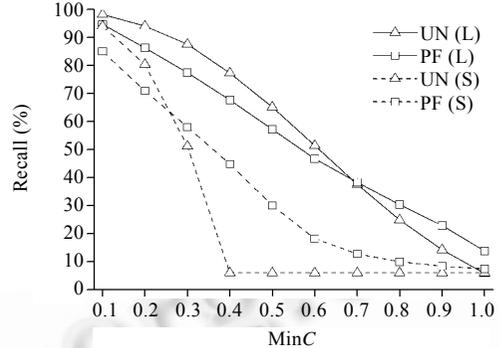


Fig.5 Recall of result returned by UN and PF Methods, *Mode=Always*, $T=[10,20]$

图 5 UN 和 PF 方法获得的查询结果的召回率, *Mode=Always*, $T=[10,20]$

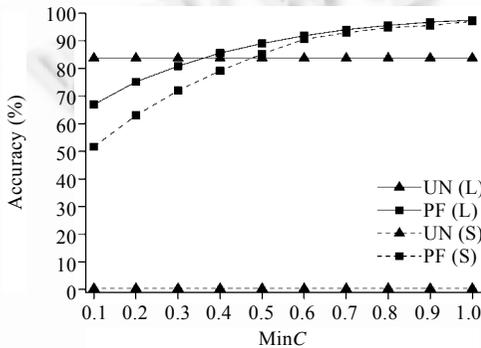


Fig.6 Accuracy of result returned by LN and PW Methods, *Mode=Sometimes*, $T=[10,20]$

图 6 LN 和 PW 方法获得的查询结果的准确率, *Mode=Sometimes*, $T=[10,20]$

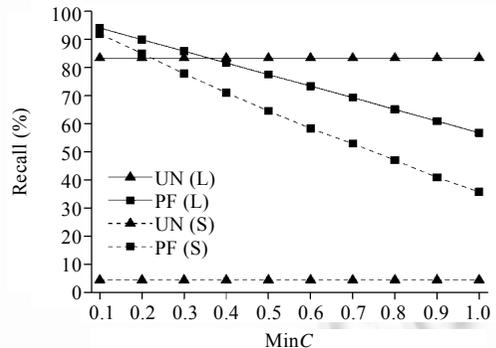


Fig.7 Recall of result returned by LN and PW Methods, *Mode=Sometimes*, $T=[10,20]$

图 7 LN 和 PW 方法获得的查询结果的召回率, *Mode=Sometimes*, $T=[10,20]$

图 4 和图 5 给出了当移动对象速度矢量信息不可用时,利用 UN 和 PF 方法处理 *Mode=Always* 的两种查询负载所返回结果的准确率和召回率.图 6 和图 7 给出了当移动对象速度矢量信息可用时,利用 LN 和 PW 方法处理 *Mode=Sometimes* 的两种查询负载所返回结果的准确率和召回率.实验结果表明,随着查询中空间区域的减小,在不同 MinC 和 *Mode* 下,各种方法获得的查询结果的正确性指标都有所降低,但 PF 和 PW 方法均好于 UN 和 LN 方法.当 *Mode=Always* 时,UN 和 PF 方法在处理关于较大空间区域的查询时返回查询结果的正确性很接近.但当查询中的空间区域减小时,UN 方法的有效性降低较大.尤其是 MinC 大于 0.4 以后,基于均匀分布假设设计的概率值已难以区分不同移动对象满足查询的可能性.此时,UN 方法返回的查询结果的准确率和召回率都降到了 5%左右.这表明,用 UN 方法返回的查询结果中只包括 5%的正确结果,而且这些正确结果仅占有所有正确查询结果的 5%.因为当查询中空间区域减小后,查询区域与移动对象未来可能的运动范围之间相交区域也相对减小.而利用均匀分布假设处理查询时预测误差则相对增大.当 *Mode=Sometimes* 时,虽然基于 LN 方法能够很好地支持关于较大空间区域的查询处理,但当查询中的空间区域缩小后,即使是处理关于相同的预测时间范围的查询时,该方法也完全失效.本文提出的 PW 方法受空间区域变化的影响较小.而且当 *Mode=Sometimes* 时,各种方法返回结果的准确率受空间区域变化的影响比 *Mode=Always* 时要小.

3.2.2 查询中预测时刻大小对查询结果正确性的影响

在实验中,所有查询的空间区域大小均为 100×100,预测时间范围分别为未来 10~15 分钟和 40~45 分钟.预测时刻越大,表示查询关注的未来时间范围越“远”,对预测方法的准确性要求越高.在图 8~图 11 中,(10)表示关于未来 10~15 分钟的查询结果的正确性,而(40)表示关于未来 40~45 分钟的查询结果的正确性.

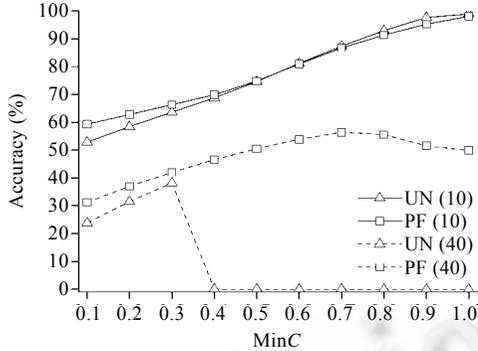


Fig.8 Accuracy of result returned by UN and PF methods, Mode=Always, T=[10,15],[40,45]

图 8 UN 和 PF 方法获得的查询结果的准确率, Mode=Always, T=[10,15],[40,45]

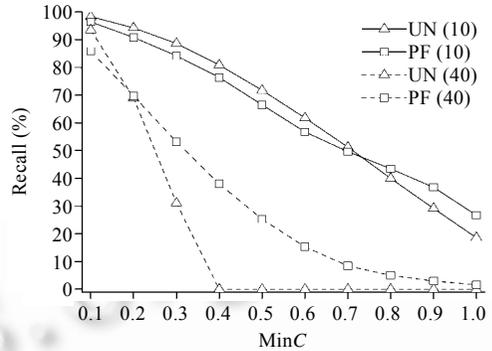


Fig.9 Recall of result returned by UN and PF methods, Mode=Always, T=[10,15],[40,45]

图 9 UN 和 PF 方法获得的查询结果的召回率, Mode=Always, T=[10,15],[40,45]

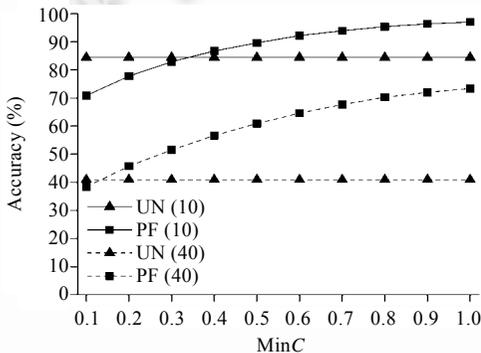


Fig.10 Accuracy of result returned by LN and PW methods, Mode=Always, T=[10,20]

图 10 LN 和 PW 方法获得的查询结果的准确率, Mode=Always, T=[10,20]

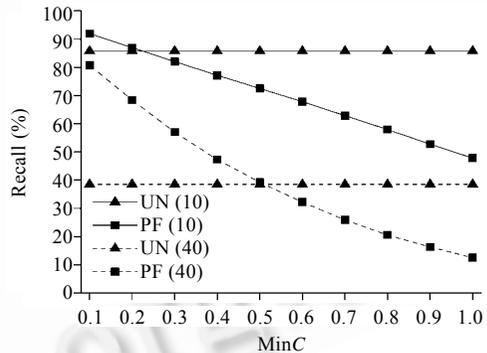


Fig.11 Recall of result returned by LN and PW methods, Mode=Always, T=[10,20]

图 11 LN 和 PW 方法获得的查询结果的召回率, Mode=Always, T=[10,20]

图 8 和图 9 给出了当移动对象速度矢量信息不可用时,利用 UN 和 PF 方法处理两种查询负载所返回结果的准确率和召回率.图 10 和图 11 给出了当移动对象速度矢量信息可用时,利用 LN 和 PW 方法处理两种查询负载所返回结果的准确率和召回率.实验结果表明,随着查询中预测时刻的增加,在不同 MinC 下,4 种方法获得的查询结果的正确性指标都随之降低,但本文提出的 PW 和 PF 方法返回结果正确性指标比 LN 和 UN 方法的正确性指标降低的幅度要小.在预测时间范围为[40,45]时,虽然 MinC 大于 0.5 以后 PW 方法返回结果的召回率,低于 LN 方法返回结果的召回率,但 PW 方法返回结果的准确率几乎一直高于 LN 方法返回结果的相应指标.而且,PF 方法返回结果的正确性指标几乎一直高于 UN 方法返回结果的相应指标.这表明,随着预测时刻的增加,基于均匀分布假设的预测方法的有效性在逐步减小.因此,本文提出的查询处理方法能够更有效地支持关于较长的未来情况的预测性时空区域查询的处理.

3.2.3 查询中最小概率阈值对查询结果正确性的影响

根据准确率的定义可知,如果查询处理过程中使用了正确的概率模型,则查询结果的准确率应当随着 MinC

的增加而增大.从以上实验结果可以发现,本文提出的查询处理方法所返回的查询结果更加符合这一要求.同时我们发现,虽然可以通过提高概率阈值来提高查询结果的准确率,但查询结果的召回率却随之降低.这说明,虽然可以通过提高 MinC 来避免查询处理过程中产生第(1)种错误(即引入错误的结果)的几率,但同时也增大了产生第(2)种错误(即过滤正确的结果)的几率.因此,如果一个基于概率模型的查询处理方法所返回的结果在获得相同的正确率时具有更高的召回率,则该方法的有效性更高.

4 结 论

预测性时空区域查询是移动对象数据库应用中最常见的查询之一.针对目前的查询处理方法存在的问题,本文提出了一种基于概率模型的预测性时空区域查询处理方法.基于本文提出的概率模型和概率计算方法,移动对象在不同的运动模式下满足查询的可能性可以被清楚地表示和计算出来.通过对大量历史轨迹分析获得的概率密度函数,提高了概率计算的准确性和预测查询结果的正确性.实验结果表明,利用文中提出的索引结构,本文提出的查询处理方法能够有效地支持预测性时空区域查询的处理,特别是在处理关于较小的空间区域和较大的预测时刻的查询时,它能够有效地提高查询结果的正确性.

References:

- [1] Saltinis S, Jensen CS, Leutenegger ST, Lopez MA. Indexing the positions of continuously moving objects. In: Chen W, Naughton JF, Bernstein PA, eds. Proc. of the Int'l Conf. on Management of Data (SIGMOD). ACM Press, 2000. 331-342.
- [2] Hadjieleftheriou M, Kollios G, Tsotras VJ, Gunopulos D. Efficient indexing of spatiotemporal objects. In: Jensen CS, Jeffery KG, Pokorn J, Saltinis S, Bertino E, Böhm K, Jarke M, eds. Proc. of the 8th Int'l Conf. on Extending Database Technology (EDBT). Springer-Verlag, 2002. 251-268.
- [3] Hadjieleftheriou M, Kollios G, Gunopulos D, Tsotras VJ. On-Line discovery of dense areas in spatio-temporal databases. In: Hadzilacos T, Manolopoulos Y, Roddick JF, Theodoridis Y, eds. Proc. of the 8th Int'l Symp. Springer-Verlag, 2003. 306-324.
- [4] Tao Y, Sun J, Papadias D. Analysis of predictive spatio-temporal queries. ACM Trans. on Database Systems, 2003,28(4):295-336.
- [5] Tao Y, Faloutsos C, Papadias D, Liu B. Prediction and indexing of moving objects with unknown motion patterns. In: Weikum G, König AC, Debloch S, eds. Proc. of the Int'l Conf. on Management of Data (SIGMOD). ACM Press, 2004. 611-622.
- [6] Wolfson O, Chamberlain S, Dao S, Jiang L, Mendez G. Cost and imprecision in modeling the position of moving objects. In: Proc. of the 14th Int'l Conf. on Data Engineering (ICDE). IEEE Computer Society, 1998. 588-596.
- [7] Pfoer D, Tryfona N. Capturing fuzziness and uncertainty of spatiotemporal objects. In: Caplinskas A, Eder J, eds. Proc. of the Advances in Databases and Information Systems, the 5th East European Conf. (ADBIS). Springer-Verlag, 2001. 112-126.
- [8] Cheng R, Kalashnikov DV, Prabhakar S. Querying imprecise data in moving object environments. IEEE Trans. on Knowledge and Data Engineering, 2004,16(9):1112-1127.
- [9] Cheng R, Xia Y, Prabhakar S, Shah R, Vitter JS. Efficient indexing methods for probabilistic threshold queries over uncertain data. In: Nascimento MA, Özsu MT, Kossmann D, Miller RJ, Blakeley JA, Schiefer KB, eds. Proc. of the 13th Int'l Conf. on Very Large Data Bases (VLDB). Morgan Kaufmann Publishers, 2004. 876-887.
- [10] Zhang W, Li JZ, Pan HW. Probabilistic predictive window query in moving objects database. Journal of Computer Research and Development, 2004,41(Supplement):50-55 (in Chinese with English abstract).

附中文参考文献:

- [10] 张炜,李建中,潘海为.基于概率模型的移动对象数据库预测性区域查询处理.计算机研究与发展,2004,41(增刊):50-55.



张炜(1975 -),男,贵州贵阳人,博士,讲师,主要研究领域为移动对象数据管理,传感器网络.



刘禹(1980 -),男,博士生,主要研究领域为传感器网络.



李建中(1950 -),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,传感器网络.