支持偏好调控的路网隐私保护 k 近邻查询方法

倪巍伟 陈 萧 马中希

(东南大学计算机科学与工程学院 南京 211189)

(东南大学计算机网络和信息集成教育部重点实验室 南京 211189)

摘 要 随着人们对个体隐私的日益关注,位置服务中的隐私保护问题成为数据库领域新兴的研究热点.针对面向路网的隐私保护 k 近邻查询中,保护位置隐私引发的难以兼顾查询质量问题及查询者对查询效率与准确性间偏好调控需求问题,引入 PoI(Points of Interest)概率分布概念,通过分析服务器端 PoI 邻接关系,生成 PoI 概率分布.将服务器端查找 k 近邻 PoI 过程分解为路网扩张查询阶段和迭代替换阶段,为迭代替换阶段构建基于 PoI 概率分布的可替换 PoI 概率预测机制.基于所构建概率预测机制,提出支持用户偏好调控的保护位置隐私 k 近邻查询方法 AdPriQuery(Adjustable Privacy-preserving k nearest neighbor Query),查询者通过调节筛选概率阈值,在兼顾位置隐私安全的同时,实现对查询效率与准确性的偏好调控.所提调控机制对已有的基于空间混淆的路网环境保护位置隐私近邻查询方法具有良好的兼容性.理论分析和实验结果表明,所提方法在兼顾保护位置隐私的同时,能有效提高服务器端查询效率,同时支持查询结果准确性与查询效率的偏好调控要求.

关键词 位置隐私保护;k 近邻查询;路网;PoI 概率分布;偏好调控

中图法分类号 TP311 **DOI 号** 10.3724/SP.J.1016.2015.00884

Location Privacy Preserving k Nearest Neighbor Query Method on Road Network in Presence of User's Preference

NI Wei-Wei CHEN Xiao MA Zhong-Xi

(*Department of Computer Science and Engineering, Southeast University, Nanjing* 211189)

(*Key Laboratory of Computer Network and Information Integration in Southeast University, Ministry of Education, Nanjing* 211189)

Abstract With the increasing concerning to individual privacy, privacy protection in location based services becomes a hot topic in the domain of database research. In recent years, privacy-preserving location based k nearest neighbor querying on road networks receives thriving attentions for its complexity and practical application values. Most of current work in common falls short in ignoring user's preferred adjusting requirement in relation to query accuracy and query efficiency, as well as the heavy workload at the server side. Concerning these problems, definition of adjacent PoI (Points of Interest) and probability distribution of PoI are introduced and the method of constructing probability distribution of PoI is devised by analyzing neighboring relation of those PoIs stored at the server side. Further, the server-side querying process is partitioned into two stages, namely incremental network expansion query stage and iteratively replacing stage. In iteratively replacing stage, the probability prediction scheme is elaborated to estimate the probability that there exists some replaceable PoI for current searched kth nearest neighbor PoI within the given number of search steps. Based on the aforementioned definitions and scheme, an adjustable privacy-preserving k nearest neighbor query method AdPriQuery is proposed, which can provide

收稿日期:2014-05-15;最终修改稿收到日期:2014-11-20.本课题得到国家自然科学基金(61370077,61003057)资助.**倪巍伟**,男,1979年生,博士,副教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为数据挖掘、数据隐私安全保护.E-mail:wni@seu.edu.cn.**陈 萧**,男,1990年生,硕士研究生,主要研究方向为数据隐私安全保护.**马中希**,男,1991年生,硕士研究生,主要研究方向为数据隐私安全保护.

query users the function to seek trade-off between query efficiency and query accuracy by adjusting the parameter of probability threshold. The scheme can be deployed to those existing cloaking based privacy-preserving nearest neighbor query solutions on road networks. Theoretical analysis and empirical study demonstrate our solution's performance.

Keywords　location privacy protection; k nearest neighbor query; road network; probability distribution of PoIs; adjusting in presence of user's preference

1　引　言

移动通信与空间定位技术的快速发展促进了基于位置服务(Location Based Services,LBS)的兴起,k 近邻查询是位置服务中常见的查询模式,指查找距查询者当前位置最近的 k 个目标 PoI(Points of Interest),例如查找距查询者最近的 k 个餐厅或加油站等.从查询者运动模式角度,可分为无约束环境下近邻查询和路网环境下近邻查询,前者假设查询者可以向任何方向运动,后者要求查询者的运动必须受实际路网约束.该服务需要查询者向服务提供方提供自身准确位置以获取查询结果,实时位置信息蕴含用户行为模式,随着人们对个体信息安全的日益关注,用户的位置安全日益受到重视,如何在不泄漏个体位置隐私的同时实现 k 近邻查询成为数据库领域隐私敏感位置服务研究的一个热点.

近年来,研究者提出了一系列保护位置隐私查询的方法,基本思想是通过隐匿机制,将隐藏后查询者位置及查询请求提交服务器,服务器返回候选结果,由可信第三方或查询者从中甄选出查询结果.从隐藏技术角度,这些方法可分为三类[1-5]:位置干扰(Location Obstruction)、空间变换(Space Transformation)和空间混淆(Spatial Cloaking).已有研究多数以维持查询准确性为前提,在查询效率与查询者位置隐私安全间寻求折衷.然而,查询处理中对位置隐私的保护引发了一些新问题:

(1)保护位置隐私导致难以兼顾查询质量

相对于不考虑查询者位置隐私的 k 近邻查询,保护查询者位置隐私的约束不可避免地导致查询效率降低.尽管已有一些研究致力于提高隐私保护 k 近邻查询的效率,但从隐私保护查询内在制约机理角度分析,查询效率、准确性与位置安全是对立的.在保证位置隐私安全的前提下,对提高查询效率的无限追求,最终不可避免需要牺牲查询的准确性.

(2)保护位置隐私引起查询者偏好调控需求

不考虑位置隐私的情况下,查询者只需解决查询效率和查询准确性二元关系,相对于保护位置隐私查询,其查询机理相对简单,使得查询者能够兼顾效率与准确性.保护位置隐私查询中,查询者面对位置安全性、查询准确性、查询效率三元关系,其查询机理的复杂,使得查询者需要在查询效率、准确性与位置安全间进行取舍,出现偏好调控需求.例如,对最近邻查询,准确性要求通常高于位置安全和查询效率要求;而对 k 近邻查询,其处理较最近邻查询更为复杂,在保证查询者位置安全前提下,查询者通常能够接受牺牲查询准确性以换取查询效率的提升.

如图 1 所示路网,$n_1 \sim n_{12}$ 对应公路交点,$p_1 \sim p_9$ 为分布在公路附近的 PoI(如加油站),查询者在图示位置发起距其最近 3 个加油站的查询,根据路网扩张搜索方法(具体见 3.1 节),服务器需要遍历 n_8n_5、n_8n_2、n_8n_9、n_9n_7、n_9n_{10} 这 5 条边才能确定距其最近的 3 个加油站 p_1,p_2,p_3.这一过程中,查找第 1、2 近邻,仅需遍历查询者所在边 n_8n_9,确定第 3 近邻 p_3 时,则需遍历其余 4 条边,若查询者愿意接受所查找到的第 3 近邻可能不准确(取决于选取 4 条边的哪条边遍历),即若选择 n_8n_2,p_5 将被认为是第 3 近邻反馈给查询者,从而节省 60%遍历时间.

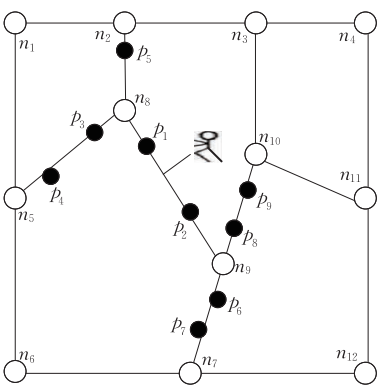

图 1　路网 3 近邻查询示意图

面向路网的隐私敏感查询中,通常采用空间混淆技术将查询者位置泛化为混淆区域(公路子网)提交服务器处理,服务器端通过扩大搜索空间返回候

选解集实现对查询者位置隐私的保护，搜索空间的扩大和复杂路网结构限制，更是加大了查询处理开销，查询准确性与效率间的调控机制更显重要.

考虑位置服务中，查询者对位置隐私安全关注度日益提高，本文针对面向路网的隐私敏感 k 近邻查询存在的上述问题，提出一种查询者可根据自身偏好要求，在不影响查询者位置安全前提下，动态调控查询准确性与查询效率的面向路网隐私保护 k 近邻查询方案. 论文主要贡献包括：

(1) 针对保护查询者位置隐私约束，片面提高查询效率存在的局限性，采用牺牲准确性换效率策略，缓解满足隐私需求约束下查询效率与准确性间的根本对立关系；

(2) 关注查询者偏好调控需求，在满足查询者保护位置隐私约束前提下，构建查询准确性与查询效率间的偏好调控机制；

(3) 提出用户可调控的路网环境保护位置隐私 k 近邻查询方法 AdPriQuery(Adjustable Privacy-preserving k nearest neighbor Query)，查询者可通过调节筛选概率阈值，在保证位置安全的同时调控查询准确性与效率，并通过理论分析和实验验证所提方法的有效性.

本文第 1 节介绍保护位置隐私近邻查询研究现状；第 2 节对相关工作进行概述；第 3 节对问题进行描述，提出解决思路并引入 PoI 概率分布概念；第 4 节提出 PoI 概率分布生成方法；第 5 节构建基于 PoI 概率分布的可替换 PoI 概率预测机制，并提出查询者可调节查询准确性与效率的保护位置隐私 k 近邻查询方法；第 6 节对所提方法的有效性进行实验分析和验证；最后，总结全文并展望后续工作.

2 相关工作

近年来，保护位置隐私近邻查询得到了研究者的持续关注，采用的主要技术包括：(1) 位置干扰(Location Obstruction). 查询者持续向 LBS 服务器提交关于特定假位置的查询请求，直到返回满足其查询准确性与隐私安全要求的结果；(2) 空间变换(Space Transformation). 设计保持数据空间关系的数据变换方法将原始位置和查询转换到新的数据空间执行以保证查询者位置隐私与查询准确性；(3) 空间混淆(Spatial Cloaking). 将查询者位置扩展为包含该位置的泛化区域并提交至 LBS 服务器，由查询者或可信第三方从返回候选解中甄别出查询结果.

研究者在面向非路网环境保护位置隐私近邻查询领域开展了大量工作[1,6-11]，文献[1]利用空间混淆技术，由可信第三方将查询者位置泛化为包含该位置且满足隐私模型约束的矩形区域，并发送至 LBS 服务器，LBS 服务器查找出泛化区域内所有可能位置的近邻 PoI 作为候选解返回可信第三方，可信第三方从候选解中筛选出目标结果返回查询者；文献[7]采用空间变换思想，首先将 LBS 服务器端存储的 PoI 位置坐标转换为 Hilbert 编码，查询者通过内置 Hilbert 编码转换器将自身位置转换为 Hilbert 编码并提交给 LBS 服务器，LBS 服务器利用 Hilbert 编码连续相邻特性，将查询者位置编码的邻近 PoI 编码返回查询者，查询者解码获取返回 Hilbert 编码对应的 PoI 信息；文献[8]提出基于位置干扰的 SpaceTwist 查询方法，查询者生成假位置坐标，并向 LBS 服务器提交关于假位置的近邻查询请求，LBS 服务器向查询者返回相应假位置的近邻 PoI，查询者根据返回 PoI 判断已搜索区域和待搜索区域内切这一临界条件是否满足，若满足则返回的 PoI 即为查询者真实位置的近邻 PoI；否则继续生成新的假位置，重复这一过程.

多数方法不支持查询准确性与查询效率的可调控性. 尽管基于位置干扰的 SpaceTwist 方法可以在不满足临界条件情况结束查询，实现牺牲查询准确性，换取查询效率提升. 然而基于假位置干扰的查询处理中，查询者对其与 LBS 服务器间迭代轮次的未知，使得查询者难以对查询准确性与查询效率进行可控的动态调节.

目前，面向路网的隐私保护近邻查询主要通过空间混淆技术实现[12-16]，即由查询者或可信第三方将查询者位置匿名为覆盖该位置且满足特定隐私约束的混淆区域，将生成的混淆区域及查询请求发送给服务器，服务器根据混淆区域完成相应查询并返回候选解，由查询者或可信第三方筛选出目标结果. 文献[12]采用空间混淆技术，由可信第三方生成星状子网，进一步通过合并星状子网生成满足查询者位置匿名需求的混淆区域提交至 LBS 服务器，LBS 服务器对混淆区域进行路网扩张查询获取候选解集并返回可信第三方，由可信第三方筛选出查询结果；文献[13]通过可信第三方对路网进行层次划分，选取不同层次的路网构成满足查询者隐私保护需求的

混淆区域并提交至 LBS 服务器完成后续路网扩张查询与反馈过程；文献[14]通过可信第三方对路网进行 Hilbert 编码，利用 Hilbert 编码生成匿名区域. 路网的复杂结构加大查询处理难度，不可避免的导致服务器处理代价激增. 已有研究多数侧重于可信第三方混淆区域生成策略及 LBS 服务器处理效率的提升，提升效果最终受制于保护位置隐私和查询准确性约束，缺少对查询准确性与查询效率间偏好调控问题的关注.

3　问题描述与相关概念

3.1　问题描述

面向路网的保护位置隐私 k 近邻查询常见模式如下：LBS 服务器存储查询者感兴趣的 PoI 对象（例如加油站、宾馆等），查询者将其当前位置及查询请求（查询内容及位置隐私要求）提交可信第三方，由可信第三方将其位置坐标泛化为满足保护位置隐私约束的混淆区域提交至 LBS 服务器，LBS 服务器分析混淆区域与路网的所有交点，并查找这些交点的 k 近邻 PoI，将所有交点的 k 近邻 PoI 及混淆区域所包含 PoI 组成候选解集返回可信第三方，可信第三方从中筛选出目标结果并发送给查询者. 文献[17]给出了证明：上述候选解集中的 PoI 必定覆盖查询者真实位置的 k 近邻 PoI. 本文要解决的问题是如何在兼顾查询者位置隐私保护需求前提下，支持查询者对查询准确性与查询效率的动态调控.

路网环境保护位置隐私查询领域，查询者位置隐私保护需求主要通过引入位置隐私模型实现. 目前常用位置隐私模型是基于空间混淆的匿名隐私模型，诸如 K-匿名位置隐私模型、(K,l) 位置隐私模型[16]等，K-匿名位置隐私模型要求混淆区域至少包含 $K-1$ 个其他同类移动用户；(K,l) 位置隐私模型要求混淆区域至少包含 $K-1$ 个其他同类移动用户，且这些用户分散在至少 l 条公路边上. 查询者向可信第三方发送自身位置及所指定位置隐私模型约束，可信第三方生成包含查询者位置且满足位置隐私模型的混淆区域代替查询者准确位置提交至 LBS 服务器进行查询处理，以实现对查询者位置隐私安全的保护. 目前，路网环境近邻查询中，查询者位置隐私保护需求主要由可信第三方实现.

前已述及，保护位置隐私查询中，查询者位置隐私安全性、查询准确性、查询效率间存在复杂的三元关系，要实现查询者对查询准确性与查询效率动态调控的同时不影响查询者位置隐私保护要求，调控机制应独立于可信第三方，不影响可信第三方根据查询者位置及查询者隐私保护约束生成混淆区域.

因此，路网环境兼顾查询者位置隐私保护需求的查询准确性与查询效率动态调控机制需满足以下要求：

(1) 调控机制独立于可信第三方，仅通过对 LBS 服务器端处理流程的调控实现，以保证查询者位置隐私安全不受调控机制影响；

(2) 调控机制支持常见有基于空间混淆的路网匿名隐私模型，保证调控机制对已有的基于空间混淆的路网环境隐私保护近邻查询方法有较好的适用性.

下一节，对 LBS 服务器端查询处理过程进行分析，提出解决问题的基本思路.

3.2　基本思路

路网环境保护位置隐私查询中，LBS 服务器获取可信第三方提交的混淆区域后，处理主要由对混淆区域与路网交点的 k 近邻查询组成，对每个交点的 k 近邻查询可分为以下两阶段：

(1) 路网扩张查询阶段. 从查询起点出发，每搜索到一个新的节点，将与其连接且还未搜索过的公路边加入待搜索队列，按照公路边的中点距搜索出发点的距离排序依次遍历，找到 k 个 PoI.

(2) 迭代替换阶段. 将 k 个 PoI 按其距查询位置距离由近到远排序，寻找第 k 近邻 PoI 可能存在的替换 PoI，对替换后的 k 个 PoI 按其距查询位置距离由近到远重新排序；重复这一过程，直到找不到第 k 个近邻 PoI 的替换点为止. 最终保留的 k 个 PoI 即为交点位置的 k 近邻 PoI.

分析易知，服务器端处理开销主要集中于迭代替换阶段反复寻找第 k 近邻 PoI 的可替换点过程，特别当 k 值较大时，处理效率低下，查找最后几个 PoI 点耗费的时间甚至可能超过查找前面所有点所耗费时间. 如图 1 所示，需要查找查询者的 3 近邻 PoI，从查询位置出发，处理了查询者所在边的 PoI 点 p_1、p_2 后，只处理了 1 条边；继续查找第 3 近邻 PoI，需要检验 n_8n_5、n_8n_2、n_9n_7、n_9n_{10} 这 4 条边上的 PoI. 可见，k 越大，服务器端查找 k 近邻 PoI 的开销越大，并且大量时间消耗在确定距查询位置较远的 PoI 是否为相应近邻 PoI 的处理上. 容易证明迭代替换阶段寻找替换 PoI 的计算复杂度与 k 呈近似线

性关系(证明见附录 A).

鉴于现实世界查询者在查询准确性与查询效率间存在不同的偏好取舍,在迭代替换阶段查询处理中若能预判断当前 PoI 邻近区域存在距查询位置更近 PoI 的概率,以便对是否继续查找可替换 PoI 进行预判断,将能有效提高服务器端处理效率,同时兼顾查询者对查询准确性与查询效率的偏好要求.因此,考虑从概率论角度建立 k 近邻查询准确性与查询效率间的调控机制,让查询者能够通过适当降低 k 近邻查询准确性,换取查询效率的提升.

为叙述方便起见,表 1 列出下文出现的主要变量符号含义.

表 1 变量符号含义

符号	含义	符号	含义
G	路网	D	服务器端 PoI 集合
T	搜索生成树	p,q	平面点或 PoI
$maxnum$	最大容忍搜索边数	x	路网距离
$Prob(\ )$	概率函数	k	查询对象数
e_i	路网的边	P_u	筛选概率阈值

3.3 相关概念

尽管服务器端可以分析得到所有 PoI 对象关于某个位置点的准确位置分布信息,但对查询中的每个当前 PoI 点,实时地分析该点邻近区域是否存在距查询位置更近的 PoI,将带来额外的计算开销:若分析获知该 PoI 无可替换 PoI,则无需继续查询;否则,继续查询直至找到可替换的 PoI.即便通过构建 PoI 索引结构可以一定程度提高效率,也需要对索引表进行额外的查询处理,不管分析结果如何,分析过程本身都需要对 PoI 点集(或索引表)进行至少一轮遍历.

在现实生活中可以发现,某种类型的 PoI 有其自身分布规律,例如餐厅的分布通常比较稠密,从一家餐厅出发在较近的距离内能以较大概率找到另一家餐厅;而加油站的分布比较稀疏,从一个加油站出发找另一个加油站,在较近的距离内找到的概率比较低.因此,考虑利用 PoI 分布的潜在规律,生成服务器端 PoI 概率分布,实现对是否查询可替换 PoI 的高效预判断.

引入路网距离与 PoI 概率分布定义如下.

定义 1. 路网.采用无向图 $G=(V_G,E_G)$ 描述路网,其中 V_G 代表路网中公路交点的集合,E_G 代表路网中各个公路边组成的集合.

现实世界 PoI 位置不可能在公路上,只是邻近公路,为叙述方便起见,下文用 PoI 属于某条公路表示 PoI 距该公路最近.

定义 2. 路网距离.假设 p,q 坐标点位于路网 G 所在平面,p 与 q 的路网距离指通过 G 中的边,由 p 到达 q 的最短路径距离.

定义 3. PoI 概率分布.对任意 $p\in D$,路网距离 x 为正实数,PoI 概率分布指 p 的路网距离 x 范围内存在 PoI 的概率,PoI 概率分布用 $Prob(p,x)$ 表示.

如前所述,隐私敏感 k 近邻查询中,LBS 服务器对混淆区域处理的基本操作是对混淆区域与路网的交点进行 k 近邻查询,查询中消耗大量时间判断当前 PoI 是否为相应近邻.在此过程中,可以借助 PoI 概率分布信息,预判断存在距查询位置较当前所处理 PoI 更近 PoI 的概率,根据概率值选择继续查找可替换 PoI 或停止查找,实现对查询处理效率的调控.下一节,介绍 LBS 服务器端 PoI 概率分布的构建方法.

4 PoI 概率分布

LBS 服务器端存储了路网及所有 PoI 对象的位置信息,容易计算出给定 PoI 到其余 PoI 的路网距离,考虑通过获取某个 PoI 到其余部分 PoI 的路网距离,分析距该 PoI 特定路网距离范围内存在 PoI 的概率.引入邻接 PoI 定义如下.

定义 4. 邻接 PoI.对 $p,q\in D$,若存在 $e_1,\cdots,e_m\in E_G$,$e_1,\cdots,e_m$ 构成一条路经,满足 $p\in e_1$,$q\in e_m$,且满足任意 $r\in D$,$r\notin e_i(i=2,\cdots,m-1)$,称 p,q 为邻接 PoI.

对 $p\in D$,计算 p 到其所有邻接 PoI 的路网距离:从 p 出发进行路网扩张搜索,获取其所有邻接 PoI,记录所有邻接 PoI 与 p 的路网距离.当 PoI 分布较稀疏时,分析 p 的邻接 PoI 可能需遍历整个路网,因此考虑设置搜索终止条件以避免这种极端情况的发生.终止条件为已遍历边的数目超过阈值 a,且查找到的邻接 PoI 数目亦大于阈值 b,其中 a,b 为正整数.在此基础上,给出查找给定 PoI 邻接 PoI 过程所访问路径的搜索生成树定义.

定义 5. 搜索生成树.以给定 PoI 为生成树根节点,生成树第 2 层节点为该 PoI 所在公路边的两端节点,按照路网扩张查询方法,每搜索到一个新节点 v,将 v 及搜索 v 经过的边 (u,v) 加入生成树.

如图 2 所示路网,外围 3 条公路边长度为 3,内部 3 条边长度为 2,p_0,p_1,p_2,p_4 位于 1/3 处,p_3 位于所在道路中间(注:图 2 仅作示意,图中的边非直

线).假设搜索终止条件阈值 a 设置为 6,b 设置为 2,p_0 的搜索生成树如图 3 所示.

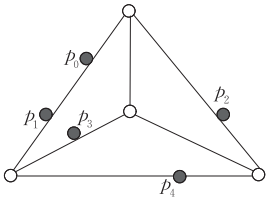

图 2　路网示意图

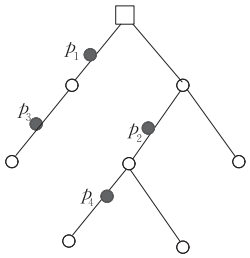

图 3　搜索生成树示意图

由于设置了终止条件,需要记录没有找到邻接 PoI 的路径数量以及找到邻接 PoI 的路径数量.这些路径被选中的概率是不一样的.因此定义边的权重以便计算某条边被选中的概率,一条边被选中的概率与这条边的权重成正比.

定义 6.　生成树边的权重.给定生成树 T,e 为 T 中一条边,e' 为 e 的父亲边,e' 具有 c_p 条子女边,假设生成树第 1 层边的权重设置为 1,e' 的权重为 $w(e')$,则 e 的权重 $w(e)=w(e')/c_p$.

叶子边代表一条从生成树根节点到叶节点的唯一路径,因此叶子边被选中的概率也能够代表这条路径被选中的概率.

性质 1.　对 $p\in D$,p 的路网距离 x 范围内存在其邻接 PoI 的概率为

$$Prob(p,x)=\frac{dist\_sum(p,x)\times\sum_{i=1}^{|D|}sum(p)}{2|D|\times dist\_sum(p)},$$

其中:$dist\_sum(p,x)$ 表示距离 p 路网距离不大于 x 的邻接 PoI 数量;$dist\_sum(p)$ 表示 p 的所有邻接 PoI 数量;$sum(p)$ 表示当前生成树中各邻接 PoI 所在路径对应叶子边的权重之和.

证明.　考虑最基本情况,所有方向都能找到邻接 PoI,从 p 点出发在路网距离 x 范围内存在其邻接 PoI 的概率为 $dist\_sum(p,x)/dist\_sum(p)$;一般情况下,并非所有方向都能找到邻接 PoI,因此上述概率为 $Prob(A)\times dist\_sum(p,x)/dist\_sum(p)$,事件 A 表示在某个路径上能够找到邻接 PoI.

首先查询 p 所在边,从 p 向所在边两端进行搜索,p 的生成树第 1 层具有两条边,设置这两条边的权重为 1,由生成树边的权重定义可知,父亲边的权重为其所有子边权重之和,因此生成树所有叶节点所连接边的总权重等于第 1 层边的总权重 2.

对任意一条边 e,其父亲边为 e',e' 有 c_p 条子女边,$Prob(\text{选取}e)=Prob(\text{选取}e')/c_p$,$Prob(\text{选取}e)/Prob(\text{选取}e')=1/c_p=w(e)/w(e')$.选择到第 1 层的任意一条边的概率为 1/2,即第 1 层边的权重/生成树叶子边总权重.递推易得,$Prob(\text{选取}e)=w(e)/2$.每个叶子边代表一条路径,事件 B 表示在当前生成树中沿某条路径能找到邻接 PoI,因此 $Prob(B)=sum(p)/2$.在生成树中 p 的路网距离 x 范围内存在其邻接 PoI 的概率为 $Prob(B)\times dist\_sum(p,x)/dist\_sum(p)$.

服务器端有 $|D|$ 个 PoI,对应 $|D|$ 棵生成树,

$$Prob(A)=\frac{1}{|D|}\sum_{i=1}^{|D|}P(B_i)=\frac{\sum_{i=1}^{|D|}sum(p)}{2|D|},$$

其中 B_i 表示在每个 PoI 对应生成树中沿某条路径能找到邻接 PoI.故路网距离 x 范围存在 p 的邻接 PoI 的概率为

$$Prob(p,x)=Prob(A)\times dist\_sum(p,x)/dist\_sum(p)$$
$$=\frac{dist\_sum(p,x)\times\sum_{i=1}^{|D|}sum(p)}{2|D|\times dist\_sum(p)}.$$　证毕.

设置递增 x 值序列(增幅结合路网度量规模设置),根据性质 1,容易生成服务器端距任意 PoI 路网距离 x 范围内存在其邻接 PoI 的概率,进一步结合所生成概率序列,汇总形成 PoI 概率分布.

仍以图 2 所描述路网为例,由图 3 可得 p_0 距其两个邻接 PoI 的距离分别为 1、3,对其余 PoI 进行类似处理,各 PoI 的邻接 PoI、距邻接 PoI 距离、权重等信息见表 2.若设置递增 x 序列为 1,2,3,4,根据性质 1,容易得到相应(路网距离、概率)序列为(1,0.2),(2,0.4),(3,0.65),(4,0.775).各 PoI 的邻接 PoI 点及其路网距离生成算法见算法 1.

表 2　邻接 PoI 信息

PoI	邻接 PoI	距离	PoI 所在边权重	不含 PoI 边权重
p_0	p_1,p_2	1,3	1,1/2	1/2
p_1	p_0,p_3,p_4	1,2,3	1,1/2,1/2	/
p_2	p_4,p_0,p_3	2,3,4	1,1/2,1/4	1/4
p_3	p_1,p_4	2,3	1/2,1/2	1/2,1/2
p_4	p_2,p_1,p_3	2,3,3	1/2,1/2,1/2	1/2

算法 1.　NeighborPoI(a,b).

输入:终止条件参数 a,b

输出:各 PoI 的邻接 PoI 距离

```
1.  Qset=NULL; //Qset 记录各 PoI 距其邻接点距离
2.  FOR each p in D //服务器端每个 PoI 点
3.    Lset=NULL; //初始化待搜索边集合
4.    Num=0; //已查找到的邻接 PoI 数目
5.    M=0; //已搜索边数
6.    lp=the edge containing p
7.    IF (lp contains two neighbors of p)
8.      Calculate their distances l1,l2 to p;
9.      Qset.add(l1,l2); //把距离 l1,l2 加入集合
10.   BREAK;
11.   END IF;
12.   Lset.add(lp.link_edges);//把 lp 的邻接边加入队列
13.   FOR each le in Lset;
14.     IF le has PoI
15.       Qset.add(le.PoI.distance);
          //将距离 p 最近的 PoI 距离加入 Qset
16.       Num+=1;
17.     ELSE
18.       Lset.add(le.link_edges);
          //将 le 的邻接边加入待搜索队列
19.     END IF;
20.     M+=1;
21.     IF (M≥a and Num≥b) BERAK;
          //满足终止条件,终止搜索
22.   END FOR;
23. RETURN Qset;
```

对服务器端 PoI 集合,可以设置适当的距离增幅,离线计算一组 PoI 概率分布序列并存储为 PoI 概率分布表,或基于所生成 PoI 概率分布序列,拟合 PoI 概率分布函数,生成服务器端所存储 PoI 的概率分布,为后续查询处理提供预测支持.

PoI 概率分布机制具有如下优点:

(1)采用离线处理方式产生,不占用服务器端在线处理时间,实现对路网中给定位置任意路网距离存在邻接 PoI 概率的计算;

(2)生成 PoI 点集的全局概率分布序列或概率分布函数,无需存储具体 PoI 索引信息.较之构建 PoI 索引结构,不仅节省存储空间,也能够更大程度的提升查询和判断效率.

5 查询处理方法 AdPriQuery

本节介绍路网环境支持查询者偏好调控的隐私保护 k 近邻查询方法 AdPriQuery.算法思路如下:查询者将自身位置提交至可信第三方生成满足隐私安全要求的混淆区域,可信第三方将混淆区域及查询请求提交给 LBS 服务器,LBS 服务器计算混淆区域与服务器端路网的所有交点,并利用基于 PoI 概率分布的可调控筛选替换机制,查找这些交点的 k 近邻 PoI,最终向可信第三方返回所有交点的 k 近邻 PoI 集合及将混淆区域内所包含 PoI 作为候选解集,可信第三方从中筛选出目标结果并反馈给查询者.

5.1 客户端处理

AdPriQuery 方法客户端处理流程与已有的多数面向路网隐私保护 k 近邻查询方法相同.采用查询客户端、可信第三方与 LBS 服务器三方架构,主要采用空间混淆技术实现查询者位置隐藏处理.具体包括以下步骤:

(1)查询者将自身位置、查询请求(包括查询对象数 k、隐私安全要求①)及概率阈值 P_u 提交给可信第三方;

(2)可信第三方对接收的查询请求进行处理,将查询者提交的当前位置点泛化为覆盖该点且满足查询者隐私安全要求的混淆区域;

(3)可信第三方将查询对象数 k、概率阈值 P_u 及混淆区域提交给 LBS 服务器,并等待 LBS 服务器返回候选查询结果集;

(4)可信第三方结合查询者真实位置点信息,从返回的候选解集中筛选出查询者的 k 近邻 PoI 并将结果反馈给查询者.

5.2 服务器端查询处理

LBS 服务器接收可信第三方提交的混淆区域后,首先计算出混淆区域与服务器端路网的所有交点,随后对这些交点逐个进行路网扩张查询,生成各交点的 k 近邻 PoI.

设计最小堆 H_r 和最大堆 H_p,最小堆 H_r 记录查询过程中遇到的新边,将所记录的边按其与查询位置点(交点)的距离排序,每次从最小堆 H_r 顶部寻找距查询位置最近的边进行搜索,搜索完成后将该边的邻接边加入堆并将该边从堆中删除;同时,将每次搜索到的 PoI 点添加到最大堆 H_p 中,这样搜索到 k 个 PoI 后,只需要对最大堆 H_p 顶部的 PoI 点进行替换分析处理.最小堆结构如图 4 所示,堆中每个节点存储本条边起始点到查询位置的距离 d_1 以及上一个 PoI 点距该边起始点的距离 d_2,以便计算该边存在 PoI 的概率.

算法主要思想是在隐私保护 k 近邻查询服务器端处理过程中根据用户需求,通过牺牲所获取较远

① 隐私安全约束并非本文主要关注点,基于空间混淆技术的路网位置隐匿模型均可应用本文所提调控方法实现兼顾位置隐私安全的查询准确性与效率可调控 k 近邻查询.

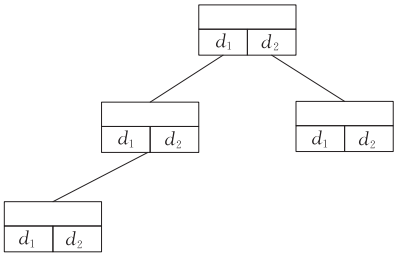

图 4　H_r堆结构示意图

k 近邻对象的准确性换取服务器端处理效率的提升.因此引入服务器端 k 近邻查询过程各 PoI 重要程度系数定义：

定义 7.　PoI 重要程度系数. 给定参数 $s(0<s<1)$，LBS 服务器查找位置点 p 的 k 近邻 PoI 过程中，其第 k 近邻 PoI 的重要程度系数定义为$I_k=s^{k-1}$.

重要程度系数体现了 p 的 k 个近邻 PoI 中第 k 近邻 PoI 的重要程度，重要程度值域为$(0,1)$，s 为指数系数，k 值越大，重要程度越低.

定义 8.　最大容忍搜索边数. 给定重要程度系数 I_k和查询前$k-1$ 个 PoI 已搜索边数 m，最大容忍搜索边数 $maxnum=(I_k\times m)/(1-I_k)$.

假设查找位置 p 的 k 近邻过程时间耗费为 O，按照近邻点搜索时间与其重要程度一致原则，第 k 近邻 PoI 的搜索容忍时间 T_k应满足 $T_k\leq O\times I_k$，查询过程主要时间消耗是对边的遍历. 查询第 k 近邻 PoI 过程，最大容忍搜索边数 $maxnum$ 应满足 $maxnum/(maxnum+m)\leq I_k$，由定义 7 知$maxnum\leq(I_k\times m)/(1-I_k)$. 如果继续查找 $maxnum$ 条边仍未能结束查询，则认为查找第 k 近邻 PoI 耗费的查询处理时间不可接受.

重要程度系数中参数 s 需根据服务器端路网信息，结合路网 PoI 密度 ρ 和具体需求设置. 由 $m=(k-1)/\rho$ 有 $maxnum=\dfrac{s^{k-1}(k-1)}{\rho(1-s^{k-1})}$，推导可得 $s=\sqrt[k-1]{maxnum\times\rho/(maxnum\times\rho+k-1)}$.

调控思想. 采用路网扩张搜索方法获取查询位置的 k 个近邻 PoI 后，按照距离查询位置远近由近到远排序，如果存在距离查询位置更近的 PoI 点，第一个需要替代的点必然是已获取的 k 个 PoI 中排在第 k 位的 PoI 点. 若分析得出在第 k 近邻 PoI 的容忍搜索边数约束下存在替代 PoI 的概率大于所设置概率阈值 P_u，继续查找；否则，认为查找到替代 PoI 点的期望太低、时耗过高，停止查找，并返回当前的 k 个 PoI 作为查询结果. 具体流程如下：

(1) 计算当前第 k 近邻 PoI 的最大容忍搜索边数 $maxnum$；

(2) 分析最小堆 H_r的前 $maxnum$ 条边中出现当前第 k 近邻 PoI 替代点的概率 P_k，若 P_k小于设定的概率阈值 P_u，终止搜索，返回当前 k 个 PoI 作为结果；

(3) 否则，继续搜索当前第 k 近邻 PoI 点的替代点，若未能搜索到替代 PoI 点，终止整个搜索过程；

(4) 若搜索到替代点，则用替代点替换当前第 k 近邻 PoI，调整最大堆 H_p，重复这一过程.

性质 2 给出在 PoI 最大容忍搜索边数约束下，计算当前第 k 近邻存在替换 PoI 概率的方法.

性质 2.　假设最小堆 H_r中第 i 条边上存在当前第 k 近邻 PoI 的替代 PoI 的概率为 P_i，在只能遍历 $maxnum$ 条边的约束下，找到当前第 k 近邻 PoI 替代 PoI 的概率为 $P_k=1-\prod_{i=1}^{maxnum}(1-P_i)$.

证明.　在第 i 条边查找到替代 PoI 的概率为 P_i，则前 $maxnum$ 条边搜索失败的概率为 $\prod_{i=1}^{maxnum}(1-P_i)$，存在替代点的概率为 $1-\prod_{i=1}^{maxnum}(1-P_i)$.　证毕.

由性质 1、2 及生成的 PoI 概率分布，结合最小生成堆 H_r中记录的边信息，容易计算出 P_i以及最大查找边数约束下存在的当前第 k 近邻 PoI 替代点的概率 P_k. 用户设定的概率阈值 P_u在 0～1 之间，设置为 0 表示查询者要求准确查询结果，即不进行查询优化；设置为 1 表示每次都省略最后替换查询过程，即完全省略服务器端迭代替换阶段查询处理，以换取服务器端的高处理效率，代价是承受可能存在的最大查询准确性损失.

5.3　算法描述

概率阈值 P_u可以调节查询者对查询准确性和查询效率的偏好要求，阈值 P_u设置的越小，查询准确性越高，服务器端查询性能的提升越小，对应较之查询效率，查询者更关注查询结果的准确性；反之，阈值 P_u设置的越大，查询准确性越低，LBS 服务器端查询效率的提升越大，对应查询者更关注查询效率情况. 筛选替换方法如算法 2 所示.

算法 2.　Filter(P_u, $max\_distance$, $maxnum$).

输入：概率阈值 P_u，最远 PoI 的距离 $max\_distance$，最大容忍搜索边数 $maxnum$

输出：True/False，是否应继续查找替换 PoI

1. $result=1$; //初始概率设置为 1
2. FOR $i=1$ to $maxnum$:
3. $cur\_distance=H_r.top(i).last\_distance$;

//$top(i)$表示 H_r中第 i 条边的 d_2值

4. $cur\_distance+=max\_distance-H_r.top(i).distance$

5. $result=result\times(1-Prob(cur\_distance))$;

//查询概率分布序列或函数获得概率值

6. END FOR

7. IF ($1-result\leq P_u$) RETURN false; //停止搜索

8. RETURN true; //继续搜索

LBS 服务器端对匿名区域与路网各个交点的 k 近邻 PoI 查询算法如算法 3 所示.

算法 3. PoIQuery(k,j,P_u).

输入:k,混淆区域和路网的交点 j,概率阈值 P_u

输出:k 个近邻 PoI

1. H_p=NULL; //初始化最大堆 H_p

2. H_r=NULL; //初始化最小堆 H_r

3. $L\_num=0$; //初始化已搜索边数 L_num

4. $H_r.add(e(j))$; //把包含 j 点的边 $e(j)$加入 H_r

5. FOR each l in H_r//遍历堆中的待搜索边

6. L_num++;

7. FOR each p in l //遍历边上的每一个 PoI

8. $H_p.add(p)$; //把 PoI 加入到最大堆 H_p中

9. IF ($H_p.num>k$);

//H_p中 PoI 数超过 k,删除堆顶

10. $H_p.pop()$;

11. END IF;

12. IF ($H_p.num=k$) $max\_distance=p.distance$;

//最远 PoI 到交点距离

13. $maxnum=(I_k\times L\_num)/(1-I_k)$;

14. IF(Filter(P_u,$max\_distance$,$maxnum$))

15. $testend$=true;

16. END IF;

17. BREAK;

18. END IF;

19. END FOR;

20. IF ($testend$) BREAK; //停止搜索

21. IF (($H_p.num=k$ and $l.distance<max\_distance$) or $H_p.num<k$) $H_r.add(l.link\_edges)$;

//l 的邻接边入队

22. ENDIF;

23. END FOR;

24. RETURN H_p中前 k 个 PoI;

LBS 服务器完成路网与混淆区域所有交点的 k 近邻查询后,合并 PoI 点集,进一步查找混淆区域包含的 PoI 点(简单范围查询即可获取).将所得 PoI 点集作为候选解集返回可信第三方,可信第三方根据查询者位置从候选解集中选出最近的 k 个 PoI 返回查询者.LBS 服务器端处理描述如算法 4 所示.

算法 4. AdPriQuery(R,k,P_u).

输入:混淆区域 R,查询对象数 k,概率阈值 P_u

输出:候选 PoI 集合

1. S=intersect(R); //生成R与服务器端路网交点集合

2. $CandSet=cover(R)$;

//候选 PoI 集合初始化为 R 区域内包含的 PoI

3. FOR each p in S

4. $CandSet=CandSet\cup PoIQuery(k,p,P_u)$;

5. END FOR;

6. RETURN $CandSet$;

分析算法调控机理及流程可知,AdPriQuery 算法能够实现 3.1 节的目标:

(1)查询者能够通过调整概率阈值 P_u对 LBS 服务器端基于可信第三方所提交混淆区域的路网扩张查询过程进行调控,实现对查询准确性和查询效率的偏好调控目的;

(2)由算法流程可知,调控处理发生在可信第三方生成混淆区域并提交 LBS 服务器后,调控过程亦未与可信第三方或查询者进行通信(除了调控过程结束将候选解集返回可信第三方).因此,AdPriQuery 算法满足预期的调控机制独立于可信第三方要求,能够保证查询者位置隐私安全不受调控机制影响;

(3)AdPriQuery 调控算法布署于 LBS 服务器端,算法的处理对象是混淆区域,本算法适用于服务器端满足以下处理模式的路网环境保护位置隐私近邻查询方法:LBS 服务器接收包含查询者位置的混淆区域进行路网扩张查询以获取候选解集,并将候选解集返回查询者或可信第三方.该模式是目前基于空间混淆的路网环境保护位置隐私近邻查询服务器端的基本处理模式,因此,所提调控机制对已有的多数基于空间混淆的路网环境保护位置隐私近邻查询方法具有良好的兼容性.

6 实验结果

本节对 AdPriQuery 算法的有效性进行实验分析.首先验证参数对算法效率和结果的影响,随后将 AdPriQuery 算法与已有的基于空间混淆技术的隐私敏感 k 近邻查询算法进行分析比较.

所用路网数据来源于美国奥尔登堡市真实路网系统(http://www.datatang.com/data/13531),该路网包括 6104 个公路交汇点及 7034 条公路,7000 个 PoI 点随机产生并分布在道路网络中.实验平台硬件配置如下:Intel(R) Core(TM)2 CPU 1.80 GHz,

2GB 内存. 根据第 4 节所设计的 PoI 概率分布生成方法，生成实验数据 PoI 概率分布如图 5 所示.

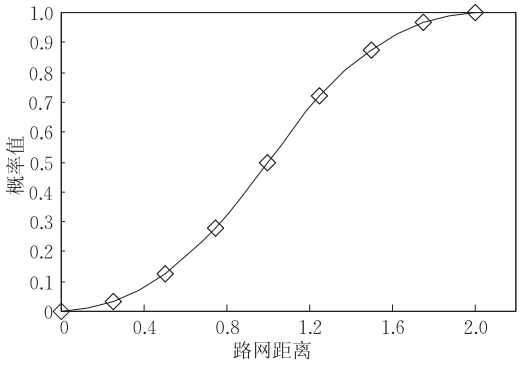

图 5　实验数据 PoI 概率分布图

6.1　AdPriQuery 算法参数实验分析

本节测试查询对象数 k 与概率阈值 P_u 对算法效率与查询准确性的影响. 参数 s 设置为 0.975(实验数据集 ρ 为 1，k 的上限取 100，最大搜索边数 $maxnum$ 设置为 10，结合定义 8 及其推论可得)，阈值 a 与 b 分别设置为 2 和 6. 图 6 对应 k 值取 10 时，随着概率阈值由 0.1 递增至 0.9，算法 AdPriQuery 服务器端处理时间及查询准确性的变化趋势. 如图 6 所示，用户设置不同概率阈值 P_u，服务器端查询处理时间随着 P_u 的增加呈减小趋势，这是由于 P_u 取值越大，服务器端执行 AdPriQuery 算法时，需要进行的查找第 k 近邻潜在替换点的处理越少，图 6 的实验结果与 5.2 节的理论分析吻合；相应的 P_u 取值越大，查找第 k 近邻替代点的操作被省略的也越多，而这当中出现的当前第 k 近邻 PoI 并非真实第 k 近邻 PoI 的可能性越大，其查询的准确率亦越低，图 6的实验结果印证了这一结论. 图 7 对应随着 k 的增加，P_u 取递增阈值(0.3，0.5，0.7)时，AdPriQuery 算法服务器端时间消耗变化趋势，由实验结果可得出如下结论：查询对象数 k 越大，服务器端查询时耗越大；但随着 P_u 的增加，时耗呈减小趋势.

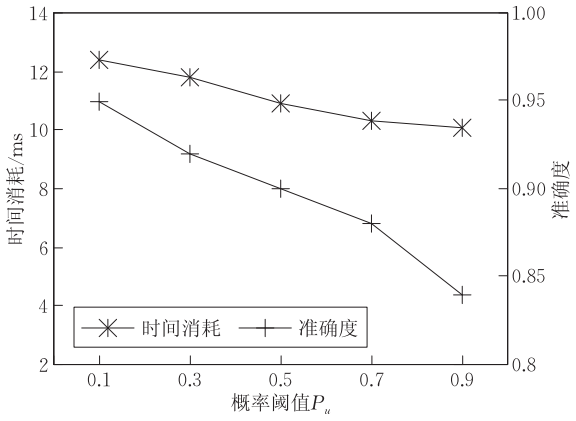

图 6　处理时间、查询准确性与概率阈值 P_u

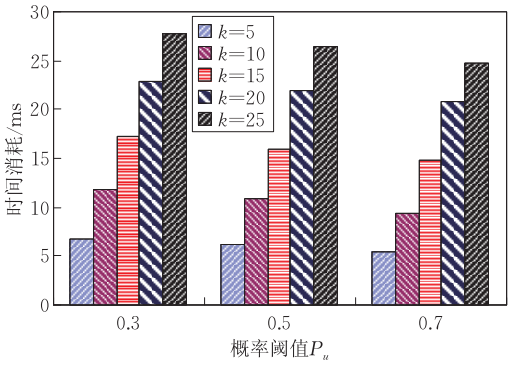

图 7　处理时间与查询规模 k 与概率阈值 P_u

可见，对不同查询对象数 k，AdPriQuery 算法通过调节概率阈值参数 P_u，能够有效的实现查询者关于查询准确性与查询时耗的偏好调控需求.

图 8 对应概率阈值 P_u 及查询对象数 k 对算法查询准确性的影响，如图所示，随着 P_u 的增加，AdPriQuery 算法查询准确性呈降低趋势，但随着 k 的增加，这种降低趋势显著变缓慢，这是由于查询对象数越多，AdPriQuery 算法筛选判断错误的近邻 PoI 占查询 PoI 数的比值下降.

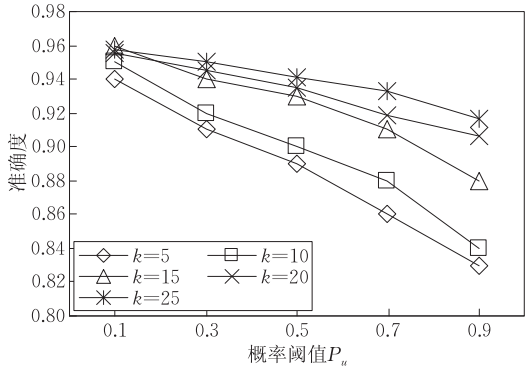

图 8　查询准确性与查询规模 k 与概率阈值 P_u

6.2　AdPriQuery 算法与传统算法对比实验

本节将 AdPriQuery 算法与传统不具备查询准确性与效率调节功能的路网环境隐私保护 k 近邻查询方法(叙述方便起见，将该类算法命名为 GenPriQuery，以文献[16]算法为代表)进行实验对比. AdPriQuery 算法与 GenPriQuery 算法客户端处理流程相同，区别在于服务器收到关于查询位置的混淆区域后，GenPriQuery 算法进行常规精确 k 近邻 PoI 查询，查询者缺少对处理过程的调控能力. 因此，只比较两算法的服务器端处理.

图 9 为查询对象数 k 为 10 时，两算法服务器端

时间消耗对比图，当 P_u 取 0 时，AdPriQuery 的时间消耗与 GenPriQuery 持平，对应概率阈值的筛选调节功能未执行情况. 随着 P_u 增加，AdPriQuery 算法的调节功效日趋显著，其服务器端时间消耗也越来越小；相应的查询准确性对比见图 10，P_u 取 0 时，概率阈值的筛选调节功能未执行，其查询准确性与 GenPriQuery 相同，查询结果均为准确 k 近邻，随着 P_u 的增加，AdPriQuery 算法的准确性呈逐渐降低趋势.

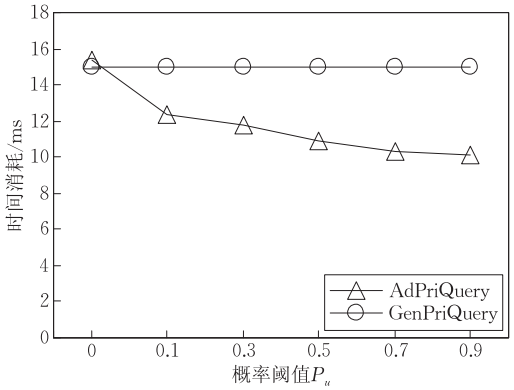

图 9 AdPriQuery 与 GenPriQuery 时间消耗对比($k=10$)

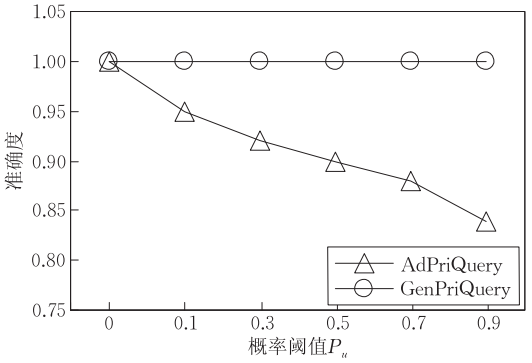

图 10 AdPriQuery 与 GenPriQuery 准确性对比($k=10$)

图 11 描述了 AdPriQuery 算法在概率阈值固定($P_u=0.5$)时，目标查询对象数 k 对 AdPriQuery 算法和 GenPriQuery 算法查询性能的影响. 如图所示，算法 GenPriQuery 服务器端执行时间明显长于算法 AdPriQuery 服务器端执行时间，随着 k 的增加，两算法运行时间均呈增长趋势，但 AdPriQuery 时间消耗的增长幅度较 GenPriQuery 算法明显趋于缓慢，这与前文指出的已有诸如 GenPriQuery 方法存在服务器端计算开销大，特别是 k 较大时，查询搜索规模激增的分析吻合. 就查询准确性而言，如图 11 所示，随着 k 的增加，AdPriQuery 算法查询准确性逐渐提高.

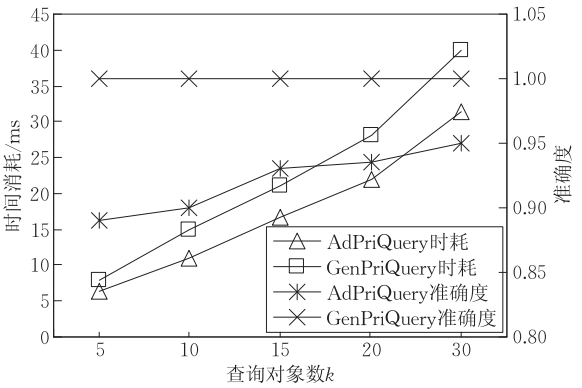

图 11 查询对象数 k 对算法查询效果的影响($P_u=0.5$)

7 总结与展望

针对面向路网的隐私保护 k 近邻查询中，保护位置隐私需求引起服务器端处理代价激增，导致保护位置隐私前提下查询效率与查询准确性绝对对立问题，提出支持查询者对查询准确性与查询效率进行偏好调控思想，实现位置查询服务的安全化和个性化. 通过引入 PoI 概率分布概念，对服务器端 PoI 邻接关系进行分析，提出 PoI 概率分布生成方法及查找替换 PoI 的概率预测机制. 在此基础上，提出查询者可调控的隐私敏感 k 近邻查询方法 AdPriQuery，查询者可通过调节筛选概率阈值，有效调控查询准确性与查询效率. 理论和实验分析验证了所提方法的有效性.

AdPriQuery 方法依赖可信第三方，然而现实应用中，可信第三方容易成为系统安全及性能的瓶颈；另一方面，现实世界路网存在单行线等路段约束. 下一阶段将进一步开展以下研究：(1) 不依赖可信第三方的路网环境支持隐私偏好的 k 近邻查询方法；(2) 结合实际路网存在的单行道、限行道等问题，研究带约束有向图模拟路网环境隐私保护 k 近邻查询中的偏好调控问题.

参 考 文 献

[1] Mokbel M F, Chow Chi-Yin, Aref W G. The new casper: Query processing for location services without compromising privacy//Proceedings of the 32nd International Conference on Very Large Data Bases. Seoul, Korea, 2006: 763-774

[2] Kalnis P, Ghinita G, Mouratidis K, Papadias D. Preventing location-based identity inference in anonymous spatial queries. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(12): 1719-1733

[3] Xue Jiao, Liu Xiang-Yu, Yang Xiao-Chun, Wang Bin. A location privacy preserving approach on road network. Chinese Journal of Computers, 2011, 34(5): 865-878(in Chinese)
(薛娇，刘向宇，杨晓春，王斌. 一种面向公路网位置隐私保护方法. 计算机学报，2011，34(5)：865-878)

[4] Pan Xiao, Hao Xing, Meng Xiao-Feng. Privacy preserving towards continuous query in location based services. Journal of Computer Research and Development, 2011, 47(1): 121-129(in Chinese)
(潘晓，郝兴，孟小峰. 基于位置服务中的连续查询隐私保护研究. 计算机研究与发展，2011，47(1)：121-129)

[5] Ni Wei-Wei, Zheng Jin-Wang, Chong Zhi-Hong. HilAnchor: Location privacy protection in the presence of users' preferences. Journal of Computer Science and Technology, 2012, 27(2): 413-427

[6] Gedik B, Liu Ling. Protecting location privacy with personalized k-anonymity: Architecture and algorithms. IEEE Transactions on Mobile Computing, 2008, 7(1): 1-18

[7] Khoshgozaran A, Shahabi C. Blind evaluation of nearest neighbor queries using space transformation to preserve location privacy//Proceedings of the 10th International Symposium on Advances in Spatial and Temporal Databases (SSTD 2007). Boston, USA, 2007: 239-257

[8] Yiu Man-Lung, Jensen C S, Huang Xue-Gang, Lu Hua. SpaceTwist: Managing the trade-offs among location privacy, query performance, and query accuracy in mobile services//Proceedings of the 24th International Conference on Data Engineering (ICDE 2008). Cancún, México, 2008: 366-375

[9] Papadopoulos S, Bakiras S, Papadias D. Nearest neighbor search with strong location privacy. Proceedings of the VLDB Endowment, 2010, 3(1): 619-629

[10] Paulet R, Kaosar M G, Yi Xun, Bertino E. Privacy-preserving and content-protecting location based queries//Proceedings of the 28th International Conference on Data Engineering (ICDE 2012). Washington, USA, 2012: 44-53

[11] Lin Dan, Jensen C S, Zhang Rui, et al. A moving object index for efficient query processing with peer-wise location privacy. Proceedings of the VLDB Endowment, 2011, 5(1): 37-48

[12] Wang Ting, Liu Ling. Privacy-aware mobile services over road networks. Proceedings of the VLDB Endowment, 2009, 2(1): 1042-1053

[13] Li Po-Yi, Peng Wen-Chih, Wang Tsung-Wei, et al. A cloaking algorithm based on spatial networks for location privacy//Proceedings of the 24th International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, Taichung, China, 2008: 90-97

[14] Kim Yong-Ki, Hossain A, Hossain Al-Amin, Chang Jae-Woo. Hilbert-order based spatial cloaking algorithm in road network. Concurrency and Computation: Practice and Experience, 2013, 25(1): 81-88

[15] Mouratidis K, Yiu Man-Lung. Anonymous query processing in road networks. Knowledge and Data Engineering, 2010, 22(1): 2-15

[16] Ku Wei-Shinn, Chen Yu, Zimmermann R. Privacy protected spatial query processing for advanced location based services. Wireless Personal Communications, 2009, 51(1): 53-65

[17] Ku Wei-Shinn, Zimmermann R, Peng Wen-Chih, Shroff S. Privacy protected query processing on spatial networks//Proceedings of the 23th International Conference on Data Engineering Workshops. Istanbul, Turkey, 2007: 215-220

附录 A.

假设路网每条边长度为 1，路网 PoI 密度为 ρ(平均每条边分布 PoI 数量)，搜索生成树节点的平均子女节点数为 c，迭代替换阶段寻找替换 PoI 过程的计算复杂度为 $2c(c^n-1)/(c-1)+1-k/\rho$，其中 $n=\left\lceil \log_c\left(1+\left(\frac{k}{\rho}-1\right)(c-1)/2c\right)\right\rceil$.

证明. 假设搜索生成树为 n 层，搜索到 k 个 PoI 所需要遍历边数为 k/ρ，根据路网扩张搜索算法，搜索生成树达到第 n 层时，一共遍历的边数为 $2c(c^n-1)/(c-1)+1$，由 $2c(c^n-1)/(c-1)+1>k/\rho$ 有 $n>\log_c\left(1+\left(\frac{k}{\rho}-1\right)(c-1)/2c\right)$，由于需要对搜索生成树的第 n 层边搜索完才能最终确定目标 k 近邻 PoI，故而 $n=\left\lceil \log_c\left(1+\left(\frac{k}{\rho}-1\right)(c-1)/2c\right)\right\rceil$，路网扩张查询阶段搜索到前 k 个近邻的计算代价为 k/ρ，故迭代替换阶段的计算代价为 $2c(c^n-1)/(c-1)+1-k/\rho$.

NI Wei-Wei, born in 1979, Ph. D, associate professor, Ph. D. supervisor. His current research interest includes data mining and data privacy protection.

CHEN Xiao, born in 1990, M. S. candidate. His main research interest is privacy preserving data application.

MA Zhong-Xi, born in 1991, M. S. candidate. His main research interest is privacy preserving data application.

Background

In recent years, privacy protection in location based services becomes a hot topic in the domain of database technology. Most of current work in privacy preserving location based query on road networks in common falls short in ignoring user's individual adjusting requirement in relation to query accuracy and query efficiency, as well as the heavy workload at the server side. Concerning these problems, definition of adjacent PoI and PoI probability distribution are introduced and the method of constructing PoI probability distribution is proposed by analyzing neighboring relation of those PoIs stored at the server side. Further, the important factor and the maximum number of search edges are devised for the kth nearest neighbor PoI of a given point. The probability prediction scheme is elaborated to estimate the probability that there exists some replaceable PoI for current searched kth nearest neighbor PoI within the given number of search steps. Based on the aforementioned definitions and scheme, an adjustable privacy-preserving k nearest neighbor querying method AdPriQuery is proposed, which provides query users the function to seek tradeoff between query efficiency and query accuracy by adjusting the parameter of probability threshold. Our work is supported by the National Natural Science Foundation of China (No. 61370077 and No. 61003057). The projects with No. 61370077 just focus on the problem of privacy preference supporting in privacy preserving location based query. The other project with 61003057 mainly solves the problem of clustering oriented privacy preserving data publishing, which is the foundation of privacy protection in location based services. Our group has obtained a series of research achievement in the domain of privacy preserving data publishing and proposed some effective data obfuscation methods to realize maintaining data cluster utility and protecting data privacy simultaneously. As for privacy protection in location based services, we have made some efforts in addressing privacy preference supporting problem in relation to query efficiency and strength of location privacy protection in k nearest neighbor query, which has been published in Journal of Computer Science and Technology (reference [5] of this paper) and so on.