F-Seeker:基于重匿名的粒度化好友搜索架构

周志刚,张宏莉,叶 麟,余翔湛
(哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 针对社交网络中好友检索服务的隐私保护问题,本文提出一种基于重匿名技术的粒度化好友搜索架构 F-Seeker. 对用户发布的位置信息采用增强的 k 匿名策略—(k,m,e)-匿名,用以防止"好奇"的搜索服务提供方对用户隐私的推测. 在处理好友搜索服务过程中,由服务提供方根据粒度化的可视策略对数据实施重匿名,实现了对用户位置信息粒度化的访问控制. 此外,文中对发布数据采用 Z 序编码并在搜索过程中通过运用剪枝策略提高搜索效率. 实验结果表明,文中提出的匿名策略在保护用户隐私的同时并没有大幅度地增加计算开销.

关键词: 重匿名; 粒度化检索; 基于位置的服务; 泰森多边形; Z 序空间填充曲线
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2016)10-2477-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.10.028

F-Seeker:Privacy-Aware Granular Moving-Object Query Framework Based on Over-Anonymity

ZHOU Zhi-gang, ZHANG Hong-li, YE Lin, YU Xiang-zhan
(*School of Computer Science and Engineering, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*)

Abstract: Aiming to the privacy-preserving problem for moving-object retrieval services in social network, we propose a granular friend retrieval framework based on over-anonymity, called F-Seeker. Before outsourcing data, we adopt an enhanced anonymity strategy – –(k,m,e)-anonymity, which preserving user privacy from the curious retrieval service provider. In the processing of providing services, the service provider employs over-anonymity strategy based on visibility requirements to realize granular data access control. In addition, we encode data using Z-order address and the retrieval efficiency can be improved by pruning. Experimental results show that the proposed strategy can protect user privacy while the computation overhead does not increase greatly.

Key words: over-anonymity; granular search; location-based service (LBS); Voronoi diagram; Z-order space filling curve

1 引言

隐私保护的"好友"搜索是社交网络领域研究的热点问题. 它可以看成是数据隐私保护、海量数据检索以及基于地理位置的服务(Location Based Service, LBS)三种技术的有机融合;即在不泄露用户隐私的前提下,由服务提供商从海量的数据中搜索符合条件的目标用户的位置信息. 近年来随着智能手机的普及,基于地理位置的社交网络应用服务层出不穷(如 Gowalla、Brightkit、陌陌等). "好友"搜索服务为生活在都市快节奏的人群构建了一座无形的沟通桥梁,拉近了人与人之间的距离,因此一经推出就自然吸引了大量用户. 然而在用户享受这些社交服务所带来便捷的同时,其对隐私信息泄露风险的顾虑已经成为阻碍这类应用服务发展的首要问题. 其次,在如何实现对用户上传的社交数据实施粒度化的访问控制以及高效的海量数据检索方面也面临极大挑战.

针对社交网络好友搜索应用中所存在的隐私保护及海量多维数据搜索问题,本文提出一个基于重匿名策略的粒度化好友搜索架构. 在用户将位置信息上传至社交网络服务器前,首先由位置匿名服务器使用(k,

收稿日期:2015-01-30;修回日期:2015-10-31;责任编辑:孙瑶
基金项目:国家 973 重点基础研究发展计划(No. 2011CB302605, No. 2013CB329602);国家自然科学基金(No. 61202457, No. 61173144, No. 61402137, No. 61402149)

m,e)-匿名算法对其进行匿名化操作,并将匿名后的数据发给 LBS 服务器. LBS 端使用 Z 序编码技术对高维用户信息进行编码,在处理好友搜索服务过程中,采用基于 Z 序编码的剪枝策略对候选集进行筛选,从而极大地提高搜索效率. 同时为实现粒度化的访问控制,LBS 服务器根据粒度化的可视策略对数据实施重匿名.

2 相关工作

针对社交数据的隐私保护问题,相关研究人员提出了许多用户位置隐私保护方案. 根据匿名原理的不同,这些方案大体可分为两类:基于位置模糊化的匿名策略[1~3]和基于 k-匿名的空间位置匿名策略[4~8]. 基于位置模糊化的匿名模型是通过发布虚假位置信息或位置坐标区域化等策略对用户位置隐私实施保护. 文献[1,2]提出基于"插入假人"的位置隐私保护策略,即用户在提交自身真实位置信息的同时,还发送若干伪造的虚假位置给 LBS 服务器,使得攻击者无法辨别出用户的真实位置. 但由于用户的精确位置会混在虚假的信息中,所以该策略依然会造成一定的安全隐患. 文献[3]提出位置相似度的概念:将服务区域细化为网格,位置服务器根据每个网格的兴趣点查询结果的相似度,计算出服务区域的相似地图,并传递给移动设备;用户再依据相似地图构造出一个服务轮廓,用于替代自身位置向 LBS 服务器请求服务,从而实现位置保护. Gruteser M 等人[4]首次将 k-匿名方法应用到用户位置隐私保护中. 该策略为每个用户构造一个匿名化的空间区域(Anonymous Spatial Region,ASR),其中 ASR 内至少含有 k 个不可区分的用户(含目标用户),使得攻击者无法以高于 $1/k$ 的概率识别出目标用户. Gedik B 等人[5]提出了一种带有区域约束的匿名策略,该策略增加了一个最小匿名区域 Amin 的隐私保护约束,从而防止因这 k 个用户的匿名区域过小而造成的位置泄露问题. 然而由这类策略生成的匿名区域可能落在某个敏感位置(Point Of Interest,POI)附近,攻击者能够据此推测出目标用户身份、职业、健康状况等隐私信息. 一个自然的解决思路是在满足 k-匿名的基础上实现对 ASR 位置多样化的需求[6],即在生成的 ASR 中至少包含 m 个 POIs,同时为了防止匿名退化攻击①,要保证 ASR 中用户在各个 POIs 间分布的均衡性;从服务效用看,用户希望构造的 ASR 是最小的. 然而,构造符合以上需求的 ASR 是一个 NP-complete 问题.

在海量数据的高效检索方面,目前大多数研究集中在通过构造高效的索引结构检索数据,如基于 R-tree[9]的索引(RUM-tree[10],TPR-tree[11]),基于 B^+-tree 的索引(B^x-tree[12]),基于 quad-tree 的索引(STRIPES[13]),然而这些技术仅仅根据移动用户空间地理位置的邻近关系存储数据,难以满足现实的社交网络应用中用户对好友多维度立体式的搜索需求,且在海量多维数据下,这些技术在处理范围查询(如经典的最近邻查询)时,需要扫描待搜索区域内所有的实体,产生极大的计算及存储开销,无法适用于对实时性要求较高的社交网络应用场景.

3 系统设计架构

在现行的"用户 - LBS 服务器"二元社交网络模型中,所有用户的社交数据(如 Check-In、兴趣、社交关系)都置于 LBS 服务器上,LBS 服务器宛然成为与用户群构成星形拓扑中的核节点. 一旦 LBS 服务器被外部攻击者入侵或内部员工恶意泄露其上数据,则用户个人隐私及其社交关系隐私都将泄露. 为此,本文提出一种基于"用户 - 位置匿名器 - LBS 服务器"的三元社交网络架构如图 1 所示,其中,位置匿名器作为可信第三方被引入系统,采用(k,m,e)-匿名算法对用户的位置信息进行匿名,并将匿名后的区域 ASR 发送给 LBS 服务器. 为防止位置匿名器成为第二个超级节点,这里位置匿名器仅用来处理用户的位置信息,而无法获知用户间的社交关系;同理,LBS 服务器仅处理好友搜索业务逻辑而无法获知用户精确的位置信息.

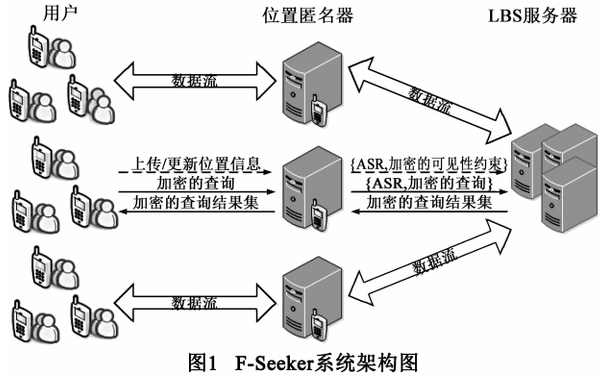

图1 F-Seeker系统架构图

4 位置匿名算法

k-匿名算法要求每一个实体所对应的记录在数据集中存在其它 $k-1$ 条与其不可区分的记录,由于其符合人们对数据安全的共识,且易于实现,已经成为数据安全评价指标并广泛应用于和数据隐私保护相关的各个领域. 针对移动用户的位置发布场景,位置匿名器通过构造一个 ASR 使得其中包含至少 k 个移动用户(含目标用户),从而保护用户的位置隐私. 然而使用传统 k 匿名算法构造的 ASR 并没有对其中包含的 POI 的数量

① 匿名退化攻击:攻击者可以根据用户在 POIs 间分布的不均衡性,在概率上缩小 ASR 的范围,使得能以高于 $1/k$ 的概率推测出目标用户的精确位置.

进行限制,当 ASR 中仅包含一个敏感 POI(如肿瘤医院)时,用户的个人隐私将可能因此而被泄露. 本文提出一个 k-匿名的增强策略(k,m,e)-匿名并将其作为安全评价指标. (k,m,e)-匿名策略在 k-匿名算法的基础上还限定了匿名区域中 POI 的数量以及用户在期间的分布.

定义 1((k,m,e)-匿名)　给定一个 ASR $S(P,Q)$,其中 P 是 S 中覆盖的 POI 集,Q 是 S 中包含的用户集,我们称 S 是(k,m,e)-匿名的,当且仅当以下三个条件同时满足:(1) S 中至少包含 k 个不可区分的用户,即 $|Q|\geqslant k$;(2) S 中至少包含 m 个 POIs,即 $|P|\geqslant m$;(3) S 中用户在 POIs 间的分布熵 $e_S=\sum_{j=1}^{|P|}-\frac{|Q_{P_j}|}{|Q|}\log_2\frac{|Q_{P_j}|}{|Q|}\geqslant e$,其中 $|Q_{P_j}|$ 表示 POI(P_j)附近的用户数量.

位置匿名器对用户位置的匿名化处理包含两个阶段:(1)预处理阶段,位置匿名器离线地对其管理范围的 POI 集合进行划分;(2)匿名阶段,采用(k, m, e)-匿名策略生成 ASR.

4.1　预处理阶段

在预处理阶段,根据 Voronoi 规则[14,15],位置匿名器使用其管理范围内所覆盖的 POI 集合对区域进行划分,使得

(1)每一个划分区域 Voronoi cell(Vcell)中包含且仅包含一个 POI.

(2)给定一个 Vcell_i 及其覆盖的 POI(P_i),则其上任意一点 q 到 P_i 的距离小于等于其到其它 POI 的距离,当且仅当 q 位于区域边界点时,等号成立.

基于 Voronoi 规则的区域划分建立了子区域与 POI 集合元素的一一映射关系,其目的在于当给定目标用户的位置信息能够快速检索到其邻近的 POI,从而提高位置匿名的效率. 图 2 展示了一个区域划分实例,其中四个 POI(P_1,P_2,P_3,P_4)基于 Voronoi 规则分别被划分到 V_1,V_2,V_3,V_4 四个 Vcells 中. 给定目标用户 q 位于 Vcell(V_1),能够得出 P_1 为与之最近的 POI.

虽然基于 Voronoi 规则的区域划分能够精确地表述 POI 间及目标用户与 POI 间的近邻位置关系,然而仅使

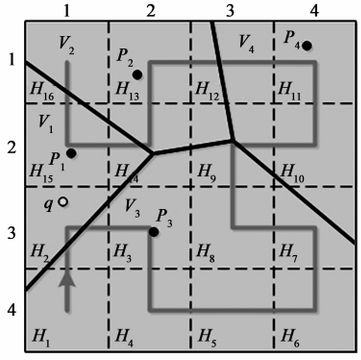

图2　基于Voronoi规则的区域划分实例

用 Voronoi 规则划分的区域存在以下几方面问题:(1)划分的子区域形状不规则,难以坐标化表示;(2)子区域的大小不均一,难以量化评估据此生成 ASR 中用户分布与 POI 分布的映射关系;(3)子区域的划分个数等于 POI 集合元素的个数,无法进一步对其进行加细划分,以致难以细粒度地构造 ASR. 为此,在以 Voronoi 规则对区域进行划分的基础上,本文使用 Hilbert 空间填充曲线[16,17]对空间区域进行迭代划分. 给定待划分区域$\mathbb{G}$:(1)首先将$\mathbb{G}$划分为 $N\times N$ 个大小均一的格状子区域(Grid cell,Gcell),其中 $N\in 2^n(n=1,2,\cdots)$;(2)使用 Hilbert 线序对本层的 Gcells 进行编号排序;(3)返回到步骤(1)对每一个 Gcell 迭代地进行加细划分. 显然,使用 Hilbert 空间填充曲线划分的区域解决了上文提出的三方面问题,此外,Hilbert 线序相近的区域其地理位置也具有近邻性. 结合 Voronoi 规则及 Hilbert 空间填充曲线的对区域划分的优点,本文构造基于网格的 B^+-tree(grid-based B^+-tree, gB^+-tree)存储各个子区域 Gcells 中 POI 和用户的位置及分布信息.

定义 2(gB^+-tree)　给定待划分区域$\mathbb{G}$,并将其设定为 gB^+-tree 的根节点(root),根据 Hilbert 空间填充曲线迭代地对其进行加细划分,其中每一个节点被划分为 $2^n\times 2^n(n=1,2,\cdots)$ 个 Gcells,称最细层划分的 Gcells 为叶节点,其它节点为中间节点. 对于每一个中间节点,使用二元组($|P|$,$|Q|$)统计其覆盖区域中 POI 及用户的数量;对于叶节点,使用三元组(P,Q,H)记录其内 POI 及用户的信息,其中 P 记录其上的 POI 集合,Q 记录其上用户集合,H 是属性 P 的扩展属性,当 $P=\varnothing$时,H 记录在 Voronoi 规则空间划分下该叶节点所属 Vcells 中 POI 所在的 Gcell 节点的 Hilbert 序 ID;否则 H 记录在 Voronoi 规则空间划分下 P 属性值所对应 Vcell 覆盖的所有格状区域的 Hilbert 序 ID.

例如,图 3 展示了针对图 2 区域实例的两层 gB^+-tree 存储结构. 所有的中间节点被划分为 2×2 子格区域,并使用 Hilbert 线序将其标记为 $H_1\sim H_{16}$. 其中 root 节点 $H_{1\sim 16}$:(4,7)表示整个区域内包含 4 个 POIs 及 7 名移动用户,类似地,$H_{1\sim 4}$:(1,2) 表示子区域$\{H_1,H_2,H_3,H_4\}$内包含 1 个 POIs 及 2 名移动用户;叶节点 H_7:$\{\varnothing,\{q_3,q_4\},\{H_3,H_{11}\}\}$表示区域 H_7 内不含 POI($P=\varnothing$),仅包含 2 名移动用户($Q=\{q_3,q_4\}$),且根据 Voronoi 规则对空间区域的划分,H_7 分属于 Vcell$\{V_3,V_4\}$,因此与其对应的 POI(p_3,p_4)是距 H_7 最近的 POI 集合,如图 2 所示,p_3,p_4 分别位于 H_3,H_{11};H_{11}:$\{\{p_4\},\{q_5\},\{H_7,H_9,H_{10},H_{11},H_{12}\}\}$表示区域 H_{11} 内覆盖 POI($P=\{p_4\}$),仅包含 1 名移动用户($Q=\{q_5\}$),且根据 Voronoi 规则对空间区域的划分,p_4 属于 Vcell(V_4),且 V_4 覆盖$\{H_7,H_9,H_{10},H_{11},H_{12}\}$.

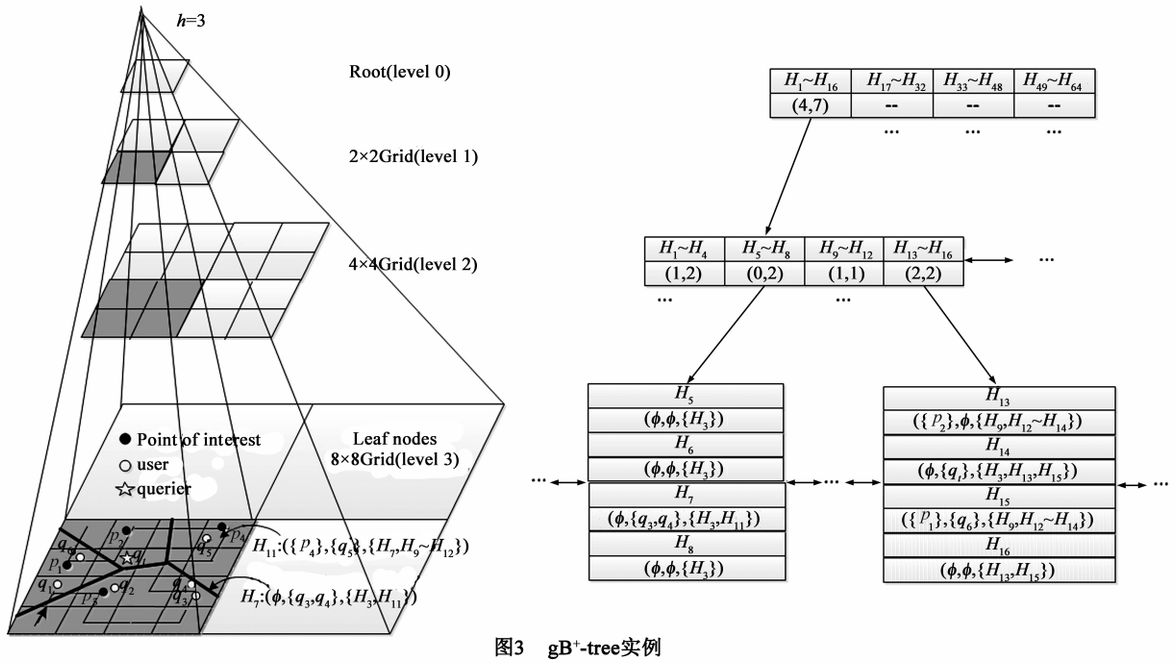

图3 gB⁺-tree实例

4.2 基于(k,m,e)-匿名的位置匿名策略

在现实的应用需求中,用户要求 ASR 在满足隐私需求(即 ASR 对隐私保护的安全度越高越好)的同时满足可用性(utility)需求(即 ASR 应该尽可能的小).一个自然的解决方案是设定一个隐私泄露的上限而求最小化的 ASR. 这里首先定义一个效用函数 $U(\mathrm{ASR})$ 用于描述 ASR 所覆盖的区域面积,

$$U(\mathrm{ASR}) = \left(\max_{g_i\in \mathrm{ASR}} o^{g_i}_{TR\to x} - \min_{g_i\in \mathrm{ASR}} o^{g_i}_{LB\to x}\right) \times \left(\max_{g_i\in \mathrm{ASR}} o^{g_i}_{TR\to y} - \min_{g_i\in \mathrm{ASR}} o^{g_i}_{LB\to y}\right) \tag{1}$$

其中,g_i 表示构造的 ASR 中所覆盖的格状单位区域,$o^{g_i}_{TR}$ 表示 g_i 的左上角点,$o^{g_i}_{LB}$ 表示 g_i 的右下角点,而(x,y)表示二维坐标的 x 轴及 y 轴分量. 据此,构造最小的满足匿名要求的 ASR 问题可以转化为下面的优化问题,

$$\begin{aligned} \text{Maximize}\quad & \frac{1}{U(\mathrm{ASR})} \\ \text{subject to}\quad & \sum_{j=1}^{|P|} -\frac{|Q_{p_j}|}{|Q_{\mathrm{ASR}}|}\log_2\frac{|Q_{p_j}|}{|Q_{\mathrm{ASR}}|} \geqslant e, \\ & \sum_{g_i\in \mathrm{ASR}} |Q_{g_i}| \geqslant k, \\ & \sum_{g_i\in \mathrm{ASR}} |P_{g_i}| \geqslant m, \\ & U(\mathrm{ASR}) \leqslant A \end{aligned} \tag{2}$$

其中,$|Q_{g_i}|$ 表示区域 g_i 中移动用户的数量,$|P_{g_i}|$ 表示 g_i 中 POI 的数量,参数(k,m,e)是预设的隐私泄露上限,阈值 A 是用户设定的可接受的最大 ASR,$|Q_{p_j}|$ 表示 POI p_j“附近”的用户数量并且 $Q_{p_j} = \sum_{g_i\in \mathrm{ASR} \text{ and } p_j \text{ in } g_i\cdot H} \frac{|Q_{g_i}|}{|H|}$.

对式(2)中目标函数的求解能够得到满足需求的 ASR,然而该目标函数没有体现对用户位置隐私的保护程度,即满足隐私泄露上限的具有相同效用的不同解决方案对隐私保护的程度也会有很大差异. 为此,下面给出式(2)优化问题所对应的拉格朗日松弛解:

$$\begin{aligned} \text{Maximize}\quad & \frac{1}{U(\mathrm{ASR})} + \lambda(E(\mathrm{ASR}) - e) \\ \text{subject to}\quad & E(\mathrm{ASR}) \geqslant e, \\ & \sum_{g_i\in \mathrm{ASR}} |Q_{g_i}| \geqslant k, \\ & \sum_{g_i\in \mathrm{ASR}} |P_{g_i}| \geqslant m, \\ & U(\mathrm{ASR}) \leqslant A \end{aligned} \tag{3}$$

其中,λ 是权重参数,

$$E(\mathrm{ASR}) = \sum_{j=1}^{|P|} -\frac{|Q_{p_j}|}{|Q_{\mathrm{ASR}}|}\log_2\frac{|Q_{p_j}|}{|Q_{\mathrm{ASR}}|}.$$

然而对式(2)及式(3)的求解是 NP-complete. 本文提出一个基于贪心策略的启发式算法(算法 1)近似地解决式(2)及式(3)中的优化问题. 算法 1 的基本思想是迭代地扫描 gB⁺-tree:

(1)逐层加细地定位目标用户所在的区域,直到其子节点覆盖的 POI 数量少于用户设定的阈值 m.

(2)检测目标用户所在子区域内用户的数量及用户分布熵:(a)若用户的数量小于 k,则根据 Hilbert 序对子区域进行扩张;(b)若用户分布熵小于 e,则根据 Voronoi 规则对用户分布熵贡献最小的格状区域进行扩张. 直到满足用户的隐私需求或打破可用性需求,算法终止.

算法 1　基于贪心策略的 ASR 构造算法(GenASR)

输入:匿名参数 k,m,e;
　　目标用户的精确位置 q_{loc};
　　gB$^+$-tree $*\,b$
输出:匿名域$\mathbb{G}$
GenASR(k,m,e,q_{loc},b)
1: for each entry$_i$ in b do
2: 　if 2D to L(q_{loc})$\leqslant$entry$_i$. ID then
　　//2D to L( ) transforms 2D region to Hilbert ID
3: 　　if entry$_i$. $|P|\geqslant m$ then
4: 　　GenASR(k,m,e,q_{loc},entry$_i\rightarrow$child);
5: 　　else
6: 　　$\mathbb{G}=\mathbb{G}\cup$FindPOI(entry$_i$,entry$_i$. $|P|$);
7: 　　$k=k-$entry$_i$. $|Q|$;
8: 　　$m=m-|\mathbb{G}|$;
9: 　　if $m>0$ then
10: 　　GenASR(k,m,e,q_{loc},entry$_i\rightarrow$next);
11: 　if $k>0$ then
12: 　　GenASR(k,m,e,q_{loc},entry$_i\rightarrow$next);
13: 　if $e-\left(\sum_{j=1}^{|P|}-\frac{|Q_{p_j}|}{|Q_{\text{ASR}}|}\log_2\frac{|Q_{p_j}|}{|Q_{\text{ASR}}|}\right)>0$ then
14: 　　$p_x=\min_{p_i\in P}|Q_{p_j}|$;
15: 　　for each G_i in $\mathbb{G}$ do
16: 　　if $G_i.P=p_x$ then
17: 　　$\mathbb{G}=\mathbb{G}\cup$(random choose G_j in $G_i.H$ and $G_j\cap\mathbb{G}=\varnothing$);
18: 　break;
19:return $\mathbb{G}$;
// Procedure FindPOI
FindPOI($*\,b,p$)
1:for each entry$_i$ in b do
2: 　if entry$_i\rightarrow$child$\neq\varnothing$ then
3: 　　FindPOI(entry$_i\rightarrow$child,p);
4: 　if $p\neq0$ then
5: 　　$\mathbb{G}=\mathbb{G}\cup$entry$_i.H$;
6: 　　$p=p-|$entry$_i.H|$;
7: 　else
8: 　　return $\mathbb{G}$;

5　细粒度的好友搜索算法

5.1　基于 Z 序的数据编码

除了位置信息外,在社交网络系统中,用户会发布一些可公开的个人基本信息(如年龄、性别、爱好等)以便于“好友”对其的搜索. 本文提出一种基于 Z 序的编码方案对用户社交数据进行编码.

定义 3(Z 序数据编码)　给定 M 维用户社交网络数据集 $E(Q,A,F)$,其中 $Q=\{q_1,q_2,\cdots,q_n\}$ 为用户对象集,$A=\{a_1,a_2,\cdots,a_M\}$ 为 M 维数据属性,F 为 Q 与 A 之间的映射关系集 $F=\{f_i:Q\rightarrow V_i(i\leqslant M)\}$,其中 V_i 为 a_i $(i\leqslant M)$的值域. 将每一个 V_i 映射为$[0,2^{\lceil\log_2 d\rceil}-1]$的值域空间($d=\max\{|V_1|,\cdots,|V_M|\}$),对每一条用户记录 $q_i((b^1_{\lceil\log_2 d\rceil-1}b^1_{\lceil\log_2 d\rceil-2}\cdots b^1_0),\cdots,(b^M_{\lceil\log_2 d\rceil-1}b^M_{\lceil\log_2 d\rceil-2}\cdots b^M_0))$ 实施 Z 序编码,使得

$$Z(q_i)=(\overbrace{b^1_{\lceil\log_2 d\rceil-1}\cdots b^M_{\lceil\log_2 d\rceil-1}}^{\text{第1组}}\overbrace{b^1_{\lceil\log_2 d\rceil-2}\cdots b^M_{\lceil\log_2 d\rceil-2}}^{\text{第2组}}\cdots\overbrace{b^1_0\cdots b^M_0}^{\text{第}\lceil\log_2 d\rceil\text{组}}).$$

性质 1(聚簇性)　在 Z 序编码下,两条数据拥有的公共 Z 序编码越长,其数据各个属性的相似度也越高.

定义 4(支配操作)[18]　称一个点 q_i 在 M 维空间上支配另一个点 q_j,当且仅当 $\forall a_r\in A$,有 $q_i.a_r\leqslant q_j.a_r$,并且 $\exists a_t\in A$ 使得 $q_i.a_t<q_j.a_t$.

性质 2(单调性)　在基于 Z 序编码的 M 维空间上,若 q_i 支配 q_j,则在 Z 序空间填充曲线上 q_i 的位置先于 q_j.

根据定义 3 可知,每一条 M 维用户数据对应一个 $M\lceil\log_2 d\rceil$ 位 Z 序编码,每一个编码被分为$\lceil\log_2 d\rceil$组,每组 M 位. 从第 1 组到第$\lceil\log_2 d\rceil$组可以看作是对用户发布数据的逐层细化拟合. 例如图 4 展示了一个二维数据集所对应的 Z 序编码,其中 $p_3(1,6)((001,110))$的 Z 序编码为$\overline{0}101\ \underline{10},\overline{01},01,\underline{10}$ 是其对应的 3 组 2-bit 子编码;同理,$p_4(3,5)$ 的 Z 序编码为$\overline{0}110\ \underline{11}$. 可以看到,在 Z 序编码下,$p_3$ 和 p_4 拥有共同的前缀$\overline{01}$,且都处于区域Ⅱ.

5.2　查询剪枝策略

为了便于描述,文中以对 2 维社交网络数据集的最近邻(Nearest Neighborhood,NN)查询为例介绍剪枝策略.

根据 Z 序编码的性质 1,Z 序编码可以以组为单位迭代细分. 每一条记录的 Z 序编码可以看作是$\lceil\log_2 d\rceil$层(组)2-bit 子编码构成的树形结构,其中第 $i(0<i\leqslant\lceil\log_2 d\rceil)$层编码是第 $i-1$ 层区域的“田”字形划分(如图 4 所示,00 对应区域Ⅰ,01 对应区域Ⅱ,10 对应区域Ⅲ,11 对应区域Ⅳ).

当接收到好友检索请求,LBS 服务器首先对其进行 Z 序编码,并以此为目标点,搜索其在空间中的支配最近邻对象. 为了减少候选记录的数量,本文提出以下逐层剪枝规则:

(1)当目标点位于区域Ⅰ,则剪切同层的区域Ⅱ,Ⅲ,Ⅳ.

(2)当目标点位于区域Ⅱ,则剪切同层的区域Ⅲ,Ⅳ.

(3)当目标点位于区域Ⅲ,则剪切同层的区域Ⅱ,Ⅳ.

(4)当目标点位于叶节点所在区域,则剪切由 Z 序曲线填充的该区域中位于目标点之后的对象.

考虑 $M(M>2)$维社交网络数据集的最近邻查询,其可以划归为对$\binom{M}{2}$个 2 维数据集的最近邻查询候选集交集操作的支配最近邻. 从表面看,这是一个空间消耗极大的操作,然而在现实社交好友搜索中,查询者的

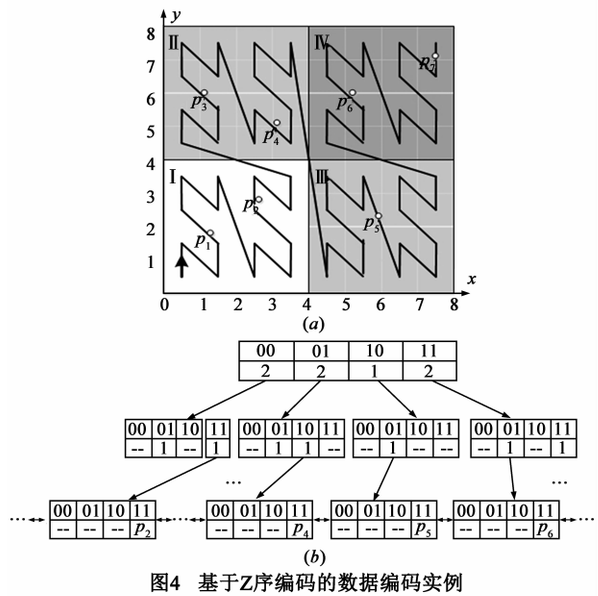

图4 基于Z序编码的数据编码实例

搜索条件往往不高于4条($M\leqslant 4$),此外,这里对高维多值社交属性集数据的最近邻操作所占用的空间消耗进行概率分析.

假设用户各个属性的数据值在各种的值域空间分布是独立均衡的. 给定 M 维用户社交网络数据集 $E(Q,A,F)$,有$\binom{M}{2}$个2维子数据集的支配最近邻对象集. 考虑搜索深度为 $x(1<x\leqslant|Q|)$,则目标点的支配最近邻同时出现在$\binom{M}{2}$个2维子数据集的 Top $-x$ 条记录的概率 $\mathscr{P}=\left(\frac{x}{|Q|}\right)^M$,即服从二项分布 $BD(|Q|,\mathscr{P})$. 根据 De Moivre-Laplace 定理[19],$BD(|Q|,\mathscr{P})$能够转化为正态分布 $ND(\mu,\sigma^2)$,其中 $\mu=|Q|\mathscr{P}$,$\sigma^2=|Q|\mathscr{P}(1-\mathscr{P})$. 由于 $\mathscr{P}(\mu-4\sigma<X\leqslant\mu+4\sigma)=99.997\%$,有$(1=\mu-4\sigma)\Rightarrow(|Q|^2+16|Q|)\mathscr{P}^2-18|Q|\mathscr{P}+1=0$,对其的求解得:$\mathscr{P}=\frac{-b+\sqrt{b^2-4ac}}{2a}$,其中 $a=|Q|^2+16|Q|$,$b=-18|Q|$,$c=1$. 据此,$x=|Q|\left(\frac{-b+\sqrt{b^2-4ac}}{2a}\right)^{\frac{1}{M}}$,即只需对各2维子数据集中支配最近邻对象集的前$|Q|\left(\frac{-b+\sqrt{b^2-4ac}}{2a}\right)^{\frac{1}{M}}$条记录实施全局支配最近邻操作.

5.3 基于粒度化可视约束的重匿名策略

为了对不同角色的搜索用户提供不同粒度的视图,用户在发布个人的社交网络数据前需向 LBS 服务器提交其对不同角色用户设定的数据访问控制条件,即粒度化可视约束.

定义5(粒度化可视约束) 给定目标用户 q 及其好友列表描述(Q,R,f,h),其中 $Q=\{q_1,\cdots,q_n\}$ 表示用户 q 的好友列表,目标用户 q 将 Q 划分为 l 个角色 $R=\{r_1,\cdots,r_l\}$ $(l\leqslant n)$,(f,h)是一对对偶操作:$f:Q\rightarrow R$ 表示由好友 ID 到角色的映射,$h:R\rightarrow Q$ 表示由角色到其相应的好友 ID 的映射且 $\forall r_i,r_j,h(r_i)\cap h(r_j)=\varnothing$. 用户 q 对角色 r_i 的可视化约束可以表示为一个五元组 $S_{q\rightarrow r_i}=(s_{r_i},t_{r_i},m_{r_i},k_{r_i},e_{r_i})$,$s_{r_i}$ 和 t_{r_i} 分别表示对角色 r_i 可搜索 q 的空间及时间约束,$(m_{r_i},k_{r_i},e_{r_i})$是设定角色 r_i 对 q 定位的重匿名化约束且 $m_{r_i}\geqslant m$,$k_{r_i}\geqslant k$,$e_{r_i}\geqslant e$((m,k,e)是对其位置信息设定的初次匿名参数).

这里,用户好友列表 Q 可以包含一个默认好友,与之对应的角色被设定为“游客”,即开辟一个与陌生人交互的接口.

LBS 服务器将用户的可视化约束内嵌到其相应的个人信息之中,并将(s_{r_i},t_{r_i})作为查询筛选条件,即当用户 q_i 满足搜索用户 q 的查询请求时,LBS 服务器检测 q_i 对 q 设定的搜索时空约束,若满足,则使用匿名参数$(m_{r_i},k_{r_i},e_{r_i})$对 ASR_{q_i} 实施重匿名,并将结果返回用户. 重匿名算法类似于算法1,由于篇幅有限,此处省略详细的算法描述.

6 实验结果及分析

实验配置为服务器采用 Intel Xeon E3 处理器,内存32GB. 位置匿名器和 LBS 服务器分别采用3个基于 Map-Reduce 的计算节点分布式计算平台. 本实验将从匿名算法性能以及好友搜索实时性方面进行实验验证.

6.1 匿名性能

本文使用仿真数据模拟现实社交网络中用户 Check-In 的位置信息,为模拟用户在空间区域内的均匀分布,数据产生器在 1000×1000 的空间区域内随机地生成用户的位置及 POI 的坐标,生成的数据库包含100K个用户位置及10K个 POIs. 实验中设置匿名请求提交窗口分别为100、300、500、700、900,其中位置匿名请求的提交用户是随机选择的,匿名参数为 $k=10$,$m=3$,$e=1.3$(参数 m 和 e 仅适用于(k,m,e)-匿名算法). 图5~7分别展示了在上述设置中,(k,m,e)-匿名算法在执行效率、隐私保护程度及匿名区域可用性三个方面与经典的 k-匿名算法及离线最优算法的比较. 图5说明,在同等条件下,随着请求数量的增加,本文提出的匿名算法执行时间呈线性增长,由于 k-匿名算法没有考虑对参数 m 和 e 的限制,因此执行时间开销总是比本文所提匿名算法要小,然而正如图6中所展示的,(k,m,e)-匿名算法在隐私保护的程度上远高于同等条件下 k-匿名算法对隐私的保护程度(这里使用用户在 POI 间的分布熵来描述 ASR 中用户位置隐私的保护程

度,其中熵越高,说明用户在 POI 间的分布越均衡). 图 7 展示了匿名区域可用性(即 ASR 的面积),为了便于比较,这里对实验结果进行归一化处理. 结果表明,虽然从统计的角度文中所提算法所生成 ASR 总是大于 k-匿名算法所产生的 ASR,但两者的相差并不大,可以推知其依然满足现实社交网络中对位置可用性的需求.

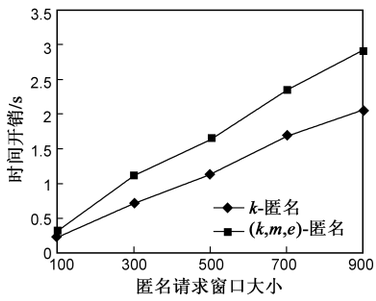

图5　(k,m,e)-匿名与k-匿名算法执行时间对比

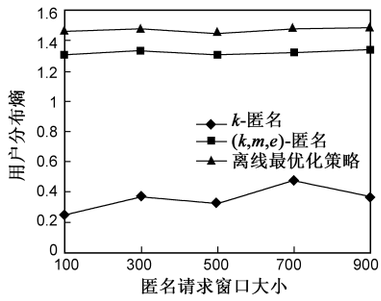

图6　(k,m,e)-匿名与k-匿名算法隐私保护程度对比

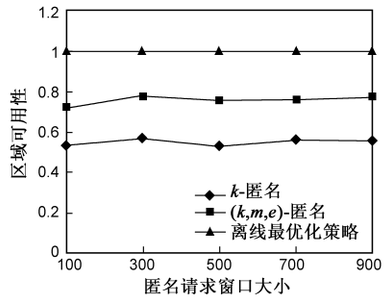

图7　(k,m,e)-匿名与k-匿名算法匿名区域可用性对比

6.2　搜索实时性

图 8 展示了基于 Z 序编码的执行效率,在上述实验环境下,用户数量设置为 10K、20K、30K、40K、50K,其中,每一个用户包含 5 个属性且属性的取值范围[1, 50]. 结果表明,对 50K 用户的一次 Z 序编码的压力测试大约需要 10s,基本满足现实场景中对编码效率的需求.

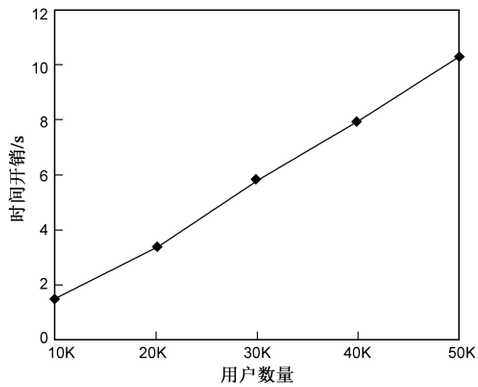

图8　Z序编码执行效率

图 9 展示了好友搜索执行效率,实验中设置搜索请求提交窗口分别为 100、300、500、700、900,为了便于比较,我们引入了经典的位置存储结构(基于 B^+-tree 的索引(B^x-tree[12]),基于 quad-tree 的索引(STRIPES[13]))作为参照,实验结果表明本文使用的基于 Z 序编码的搜索算法及剪枝策略在搜索效率方面明显优于经典的搜索算法且具有良好的实时性.

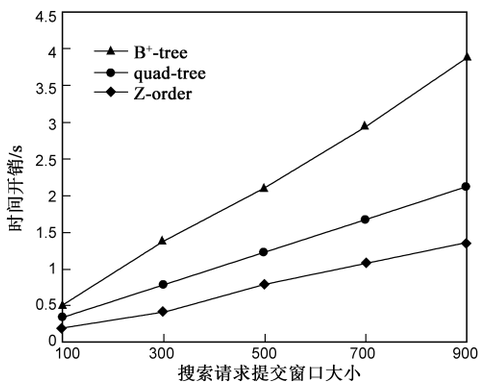

图9　好友搜索执行效率

7　结束语

本文采用重匿名技术,在保证用户位置信息可用性的前提下实现对用户位置信息进行粒度化的隐私保护. 与 k-匿名相比,基于(k,m,e)-匿名技术的隐私保护策略防止攻击者通过用户分布对用户位置信息进行推理攻击,仅附加较低的性能惩罚. 在好友搜索阶段,采用基于 Z 序的数据编码及相应的剪枝策略,大幅降低了对高维多值社交网络数据进行最近邻搜索时的时空开销.

参考文献

[1] Kido H, Yanagisawa Y, Satoh T. An anonymous communication technique using dummies for location-based services [A]. Proceedings of International Conference on Pervasive Services (ICPS05)[C]. USA:IEEE Press,2005. 88 – 97.

[2] 刘华玲,郑建国,孙辞海. 基于贪心扰动的社交网络隐私保护研究[J]. 电子学报,2013,41(8):1586 – 1591.
Liu Hua-ling, Zheng Jian-guo, Sun Ci-hai. Privacy preserving in social networks based on greedy perturbation[J]. Acta Electronica Sinica, 2013, 41(8): 1586 – 1591. (in Chinese)

[3] Rinku Dewri. Exploiting service similarity for privacy in location based search queries[J]. IEEE Transactions on Parallel and Distributed Systems,2013,25(2):374 – 383.

[4] Gruteser M, Grunwal D. Anonymous usage of location-based services through spatial and temporal cloaking[A]. Proceedings of the International Conference on Mobile Systems Applications and Services[C]. USA:IEEE Press,

2013. 163 – 168.
[5] Mokbel M F, Chow C Y, Aref W G. The new casper: Query processing for location services without compromising privacy[A]. Proceedings of the International Conference on Very Large Data Bases[C]. USA: VLDB Press, 2006. 763 – 774.
[6] Bhuvan B, Ling L, et al. Supporting anonymous location queries in mobile environments with privacy grid[A]. Proceedings of the 17th International Conference on World Wide Web[C]. USA: ACM Press, 2008. 237 – 246.
[7] 王丽娜,彭瑞卿,等. 个人移动数据收集中的多维轨迹匿名方法[J]. 电子学报,2013,41(8):1653 – 1659.
Wang Li-na, Peng Rui-qing, et al. Multi-dimensional trajectory anonymity in collecting personal mobility data[J]. Acta Electronica Sinica, 2013, 41(8): 1653 – 1659. (in Chinese)
[8] 叶阿勇,林少聪,马建峰,许力. 一种主动扩散式的位置隐私保护方法[J]. 电子学报,2014,43(7):1362 – 1368.
Ye A-yong, Lin Shao-cong, Ma Jian-feng, Xu Li. An active diffusion based location privacy protection method[J]. Acta Electronica Sinica, 2014, 43(7): 1362 – 1368. (in Chinese)
[9] Chen Q, Hu H, Xu J. Authenticating top-k queries in location-based services with confidentiality[A]. Proceedings of the International Conference on Very Large Data Bases[C]. USA: VLDB Press, 2014. 49 – 60.
[10] Xiong X, Aref W G. R-trees with update memos[A]. Proceedings of the 22nd International Conference on Data Engineering (ICDE06)[C]. USA: IEEE Press, 2006. 22.
[11] Saltenis S, Jensen C S, Leutenegger S T, Lopez M A. Indexing the positions of continuously moving objects[A]. Proceedings of ACM SIGMOD[C]. USA: ACM Press, 2000. 331 – 342.
[12] Jensen C S, Lin D, Ooi B C. Query and update efficient B^{+}-tree based indexing of moving objects[A]. Proceedings of the International Conference on Very Large Data Bases[C]. USA: VLDB Press, 2004. 768 – 779.
[13] Patel J M, Chen Y, Chakka V P. Stripes: An efficient index for predicted trajectories[A]. Proceedings of ACM SIGMOD[C]. USA: ACM Press, 2004. 637 – 646.
[14] Berg M de, Kreveld M van, Overmars M, Schwarzkopf O. Computational Geometry: Algorithms and Applications (2nd edition)[M]. Berlin: Springer-Verlag, 2000.
[15] Hu L, Ku W-S, Bakiras S, Shahabi C. Spatial query integrity with voronoi neighbors[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4): 863 – 876.
[16] Kalins P, Ghinita G, Mouratidis K, Papadias D. Preventing location-based identity inference in anonymous spatial queries[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(12): 1719 – 1733.
[17] Ghinita G, Kalnis P, Khoshgozaran A, Shahabi C, Tan K-L. Private queries in location based services: Anonymizers are not necessary[A]. Proceedings of the ACM SIGMOD International Conference on Management of Data[C]. New York: ACM, 2008. 121 – 132.
[18] Lee K, Lee W, Zheng B, Li H, Tian Y. Z-sky: an efficient skyline query processing framework based on z-order[J]. The VLDB Journal, 2010, 19(3): 333 – 362.
[19] William Feller. An Introduction to Probability Theory and Its Applications (Vol. 1, 3rd Edition)[M]. USA: Wiley, 1968.

作者简介

周志刚 男,1986 年 2 月出生,山西太原人,博士生,现就读于哈尔滨工业大学计算机系,主要研究方向为网络与信息安全、位置隐私保护、云安全等.
E-mail: zzgisgod@ sina. com

张宏莉 女,1973 年出生,吉林榆树人,博士、教授、博士生导师,国家计算机信息内容安全重点实验室副主任,计算机网络与信息安全技术研究中心副主任,信息产业部十一五科技规划组专家. 主要研究方向为网络信息安全、网络测量、并行处理.