第44卷第7期 中国科学技术大学学报 Vol. 44, No. 7
2014年7月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jul. 2014

文章编号:0253-2778(2014)07-0544-10

ARBA:基于分解重构技术的LBS隐私保护方法

林 瑜[1],韩建民[1],于 娟[2],贾 洞[1],詹皇彬[1]

(1.浙江师范大学数理与信息工程学院,浙江金华 321004;2.复旦大学计算机科学与技术系,上海 200433)

摘要:现有的匿名化方法多采用时空伪装技术,该技术计算负担重,LBS响应延迟时间长,导致LBS服务质量低.为此,提出了分解重构的匿名化方法,该方法首先对接收到的LBS查询集进行分组,形成满足匿名模型的等价类,然后对每个等价类根据不同的策略进行分解和重构,生成新的匿名查询集.此外,面向多种隐私需求,提出了一系列匿名模型,并进一步提出了基于分解重构技术的匿名模型的实现算法MBFAA.实验表明,提出的重构分解技术可以有效地实现各种匿名模型.

关键词:k-匿名;位置 l-多样性;查询 m-多样性;分解重构

中图分类号:TP311 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2014.07.002

引用格式:Lin Yu, Han Jianmin, Yu Juan, et al. A novel anonymization method based on anatomy and reconstruction in LBS privacy preservation[J]. Journal of University of Science and Technology of China, 2014,44(7):544-553,562.

林瑜,韩建民,于娟,等. ARBA:基于分解重构技术的LBS隐私保护方法[J]. 中国科学技术大学学报,2014,44(7):544-553,562.

A novel anonymization method based on anatomy and reconstruction in LBS privacy preservation

LIN Yu[1], HAN Jianmin[1], YU Juan[2], JIA Jiong[1], ZHAN Huangbin[1]

(1. *Mathematics, Physics and Information Engineering College, Zhejiang Normal University, Jinhua* 321004, *China*;
2. *Department of Computer Science and Technology, Fudan University, Shanghai* 200433, *China*)

Abstract: Most of the existing methods are realized by temporal and spatial cloaking techniques. However, these cloaking-based methods are disadvantageous due to their high computation loads and long response delays, which lowers service quality. To address these problems, a novel technique, anatomy and reconstruction, was proposed. This technique first partitions the LBS query set into several equivalence classes, making sure that each equivalence class satisfies the given anonymity constraints. Then it reconstructs the LBS queries in each equivalence class according to the predefined strategies separately, and generates a new set of anonymous queries. Considering various privacy requirements, a series of anonymity models were proposed, and a unified anonymization algorithm MBFAA was introduced to realize these models. Experimental results show that the proposed method can effectively implement all the anonymity models.

Key words: k-anonymity; location l-diversity; query m-diversity; anatomy and reconstruction

收稿日期:2014-03-21;**修回日期**:2014-06-15

基金项目:国家基金项目(61170108,6110019),浙江省自然科学基金(Y1100161,LQ13F020007)资助.

作者简介:林瑜,女,1990年生,硕士. 研究方向:隐私保护. E-mail: linyu@zjnu.net

通讯作者:韩建民,博士/教授. E-mail: hanjm@zjnu.cn

0　引言

随着 GPS、移动计算和无线通信技术的发展,基于位置服务(location-based service,LBS)得到了快速的发展.LBS 服务是通过移动运营商的无线网络(如 GSM 网、CDMA 网)或外部定位方式(如 GPS)获取移动终端的位置信息,并在 GIS 平台的支持下,为用户提供服务的一种增值业务.LBS 应用涉及人们生活的各个领域.例如,健康、交通、工作、娱乐等.非营利全球信息安全组织(ISACA)的一项调查显示,有 58%的智能手机用户在使用 LBS 应用[1];另外,美国著名市场研究机构 Pyramid Research 预测,全球 2015 年基于位置的服务市场将达到 103 亿美元,相比于 2010 年的 28 亿美元[2],增长速度惊人.可见,LBS 是一个具有广泛应用背景和广阔市场前景的产业.然而,人们在享受 LBS 服务的同时,个人隐私信息也会受到一定的威胁.2010 年 7 月份的调查表明 55%的 LBS 用户担忧他们位置隐私的泄漏[3],尤其是基于位置的社会网络(location-based social network,LBSN)的兴起[4],使得个人隐私问题变得更加突出.

攻击者通过获取 LBS 查询来获取用户的隐私信息,获取用户 LBS 查询的途径主要有 2 种:①通过窃听用户与 LBS 之间通信信道获取;②从 LBS 服务提供者(location-based service provider,LBSP)获取.前者可以通过信道加密、匿名通信等成熟的保密信道技术来获得有效的防护.后者则需要设计有效的隐私保护方法来进行防范,这也是目前 LBS 隐私保护领域研究的热点.即在假定 LBSP 不可信的前提下设计相应的隐私保护方法,达到既不泄露用户隐私又不影响用户享用 LBS 服务的目的.LBS 涉及的隐私可分为 2 类:位置隐私和查询(内容)隐私[5].位置隐私是指与用户当前位置相关的隐私,或者可以由用户位置推导出的隐私.查询(内容)隐私则是指与 LBS 服务请求内容相关的隐私,比如,若用户请求最近的教堂位置,攻击者就可以推断出用户的宗教倾向.现有的 LBS 隐私保护研究工作中,以位置隐私保护研究为主,位置隐私保护通过破坏用户与位置信息的关联来实现,而查询隐私保护则是通过破坏用户与查询信息的关联来实现.

位置 k-匿名[6-8]是当前实现 LBS 隐私保护的主要方法.位置 k-匿名是指查询用户的位置与其他至少 $k-1$ 个用户的位置不可区分,使攻击者无法推断出发出该查询用户的精确位置,以达到保护用户位置隐私的目的.时空伪装是实现位置 k-匿名的主要技术,其思想是将用户的精确位置扩大为一个至少包含 k 个用户的伪装区域(cloaking region,CR),然后用 CR 代替用户的精确位置发送给 LBSP.现有的基于时空伪装技术的匿名化算法大体可分为 2 类:依赖数据的伪装算法和依赖空间的伪装算法[9].依赖数据的伪装算法是基于用户的位置与查询发起者的距离来构造伪装区域,它又可分为基于距离的伪装算法[10-12]和 k-桶的伪装算法[13-14].依赖空间伪装算法基于包含用户的整个区域来生成伪装区域,又可分为基于网格的伪装算法[15-16]和基于区域的伪装算法[17-18].

时空伪装技术多是基于第三方可信匿名器的,该技术存在以下不足:

(Ⅰ)可信匿名器的负担过重.匿名器需要完成两个方面的工作,一是负责对移动用户的服务请求进行匿名化处理,构造出伪装区域,发送给 LBSP;二是对 LBSP 发送过来的候选查询结果集进行筛选,得到精确的查询结果并发送给用户.在 LBS 查询比较多的情况下,匿名器的负担过重.

(Ⅱ)网络负载重.时空伪装方法中,匿名器将伪装区域发送给 LBSP,LBSP 需要返回该伪装区域满足查询请求的候选集.若伪装区域较大,候选集也较大,则会增加网络负担,尤其是用户比较稀疏的情境.

(Ⅲ)服务质量差.由于构造伪装区域的算法复杂,匿名器需要一定的时间处理,造成系统响应时间延迟,匿名器筛选出来的结果也有可能不是最优结果,因此在效率和精度上都对服务质量造成影响.

(Ⅳ)易受连续查询攻击.在连续 LBS 中,若攻击者知道某些连续 LBS 查询是由某一用户发出的,通过获取不同时刻伪装区域的交集,可以推断出发送查询的用户,这种攻击称为连续查询攻击[19].

为此,我们提出了一种新的匿名化技术——分解重构技术,并基于该技术提出一个通用的匿名化算法.本文的主要贡献包括:①提出了基于分解重构技术的匿名化方法(anatomy and reconstruction based anonymization,ARBA).该方法的思想是:匿名器将分组得到的 k 个用户的 LBS 查询按照属性{ID,location,query}进行分解,再根据用户隐私约束采取不同的策略将 k 个用户的 ID,位置信息 location 与请求 query 进行重构,并把重构后的查询集发给 LBSP.②面向不同的隐私需求,提出了一系

列隐私保护匿名模型.③提出基于分解重构技术的通用匿名化算法.④通过仿真实验说明了ARBA的有效性.

相对时空伪装技术,ARBA方法的优势在于:

(Ⅰ)匿名查询构造简单,ARBA方法构造匿名查询时,只是将用户的ID,位置信息location与请求query重新组合,匿名器无需运行复杂的构造伪装区域的算法,因此计算负担轻,响应速度快.

(Ⅱ)筛选查询结果精确简单.匿名器只需筛选出用户真实的查询结果返回给用户,摒弃其他构造的虚假查询的结果.

(Ⅲ)可抵制连续查询攻击,ARBA方法不需要构造伪装区域,所以无法对其进行连续查询攻击.

(Ⅳ)匿名模型构造灵活,ARBA方法将LBS查询分解,然后又重构成新的查询集,可以根据隐私保护的需求,灵活地构造满足各种匿名约束的查询集,实现各种匿名模型.

1 预备知识

1.1 相关符号

LBS的用户隐私约束是个性化的,用户可以在不同的情境对不同的查询定义不同的隐私约束.比如,普通用户在超市并不在乎自己的位置隐私,普通用户在道路上查找附近的餐馆或加油站也不在乎自己的查询隐私,这些情况下LBS查询可能就不需要隐私约束.当一个用户处于医院或教堂,或者查找他附近的医院或教堂,而他又不希望他的位置和查找内容被泄漏,这种情况下他的位置隐私和查询隐私就需要保护,同时他的LBS查询需要设置相应的隐私约束.

用户隐私约束包括2类:①匿名级别:指不可区分个体、位置或查询的个数,比如,位置k-匿名中的k;②隐私类别:位置隐私或查询隐私.

LBS查询请求q可形式化为一个4元组,q=(Uid,Loc,Qry,Cont),其中:

Uid:用户标识;

Loc:Loc=(pos,t),t时刻的位置信息pos;

Qry:用户查询内容;Qry可以形式化为一个4元组,Qry=(obj,dist,time,prof).其中obj为需要查找的目标,dist为查找目标与Loc的距离(0表示距离Loc最近的查找目标),time为查询时间,prof为查询是否需要基于用户资料(1为基于用户资料完成查询,0为不基于用户资料完成查询).比如:查询"返回距离我500 m内的朋友"可形式化为(friend,500,now,1);"返回距离我最近的加油站"可形式化为(gasstation,0,now,0).

Cont:用户隐私约束,Cont可形式化为一个2元组,Cont=(anony_level,privacy_type),其中anony_level为匿名级别,privacy_type为隐私类别,01表示保护查询隐私,10表示保护位置隐私,11表示保护查询隐私和位置隐私.

比如,用户u_1要找距离他100 m以内的医院,要求保护查询隐私和位置隐私,匿名级别为3,可描述为(useruid,loc,qry,cont)=(u_1,Loc,(hospital,100,now,0),(3,11)).

为描述方便,我们引入以下符号:

D:LBS查询集,即在给定的时间段中,匿名器接收到的LBS查询集合;

E_i:LBS查询集的第i个等价类;

q(Uid,Loc,Qry,Cont):LBS查询,简写为q;

q(Uid):查询q的用户标识Uid的值;

q(Loc):查询q的位置区域Loc的值;

q(Qry):查询q的请求内容Qry的值;

q(Cont):查询q的隐私约束Cont的值;

q(UidLoc):查询q的用户标识Uid及Loc的值;

q(LocQry):查询q的用户标识Loc及Qry的值;

对于LBS查询q=(Uid,Loc,Qry,Cont),位置隐私关注的是q(Uid)与q(Loc)之间的关系,而查询隐私关注的是q(Uid)与q(Qry)之间的关系.

1.2 基于可信匿名器的体系结构

目前多数位置k-匿名的研究工作是基于可信第三方匿名器的,即在移动设备和LBSP之间建立一个可信的匿名器来完成位置匿名化工作.可信匿名器主要由3部分组成:伪装引擎、查询库及查询结果过滤器.在可信匿名器的体系结构下,时空伪装的匿名化过程为:匿名器接收到LBS查询,首先为查询更换用户id,如果该用户是第一次发出LBS查询,则匿名器用一个新的假id′替换用户的id;否则,使用该用户以前的假id′作为用户的id;然后伪装引擎根据该LBS查询的隐私约束,生成伪装区域CR,将伪装后的查询请求q以(id,id′,CR,qry)的形式存入查询库.CR是用户的伪装区域.匿名器生成匿名查询q'=(id′,CR,qry),将q'发送给LBSP.LBSP执行查询q',并把查询结果(q',R)发送给匿名器.q'

为查询标识,R 为查询结果候选集.查询结果过滤器根据查询库中对应的查询 $q=(\mathrm{id},\mathrm{id}',\mathrm{CR},\mathrm{qry})$,对查询结果候选集进行筛选,生成精确的查询结果,并把查询结果返回给用户,其过程如图 1 所示.

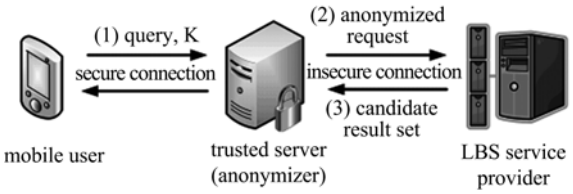

图 1　基于可信第三方匿名器的体系结构

Fig. 1　Architecture based on the trusted third party anonymity

2　分解重构技术

2.1　框架

分解重构是实现位置 k-匿名的一种新方法,也是基于可信匿名器的体系结构,其思想不同于时空伪装方法,它无需构造匿名区域,而是把每个 LBS 查询按属性分解,然后再组合成多个查询,包括一个真实的查询和多个虚假的查询.类似时空伪装方法,匿名器接收到 LBS 查询,首先需要将用户 id 替换成假 id′.该匿名化过程如图 2 所示,可分为三个步骤:

Step 1　分组.匿名器将接收到的查询分成若干个等价类,每个等价类至少有 k 个 LBS 请求.

Step 2　分解.将每个等价类的 LBS 请求根据分解策略按属性进行分解.

Step 3　重构.将分解后的 LBS 请求根据相对应的重构策略重新组成多个查询,然后发送给 LBSP.

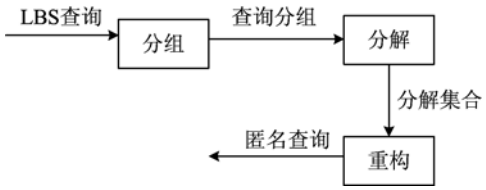

图 2　分解重组模型

Fig. 2　Anatomy and reconstruction model

2.2　分组

定义 2.1　设在给定的时间间隔内,匿名器共收到 n 个 LBS 请求,记为 $D=\{q_1,q_2,\cdots,q_n\}$,将这些查询按某种策略划分为 m 个等价类 $\{E_1,E_2,\cdots,E_m\}$,满足 $\bigcup_{i=1}^{m}E_i=D$ 且 $\forall\, i,j(1\leqslant i\neq j\leqslant m)$,$E_i\cap E_j=\varnothing$,此过程称为分组,$E_i$ 为经过分组的等价类.

通常匿名器在 Δt 时间间隔内会接收到许多查询且时间间隔比较短,因此我们假设在这个时间间隔内,匿名器不会收到同一用户的两个查询.

分组的目标是在满足隐私约束的条件下,分组中 LBS 查询的个数尽可能的小,以减少重构的虚假查询个数.分组应满足以下原则:

(Ⅰ)同一等价类的用户隐私类别尽量相同,因为不同的隐私类别需要采用不同的分解和重构策略;

(Ⅱ)同一等价类的用户匿名级别尽量相同,等价类的大小应至少等于等价类中查询的最大隐私级别.

(Ⅲ)同一等价类下的位置信息尽可能不同,以抵制位置同质性攻击;

(Ⅳ)同一等价类下的位置信息差距不能过大,以防构造出无效的查询.

(Ⅴ)同一等价类下的敏感查询所占的比率不能太高,以抵制查询隐私泄密.

考虑到 Δt 时间间隔内,匿名器接收到的查询有时不够多,以致无法建立满足匿名约束的分组,为此,我们在匿名器中增加了历史 LBS 查询缓冲器,缓冲器的大小可以根据需要设置,存储一段时间内发送过来的匿名查询,如图 3 所示.在当前查询数量不足的情况下,从历史缓冲器中提取查询,提取的策略有 2 种.①随机策略:匿名器从历史 LBS 查询缓冲器中随机筛选出一部分的查询并与新到达的 LBS 查询结合生成 LBS 查询集.②滑动窗口策略:该策略在历史 LBS 查询缓冲器上建立一个时间滑动窗口,滑动窗口总是接近最新时刻,然后将滑动窗口中的查询与新到的查询结合生成 LBS 查询集.

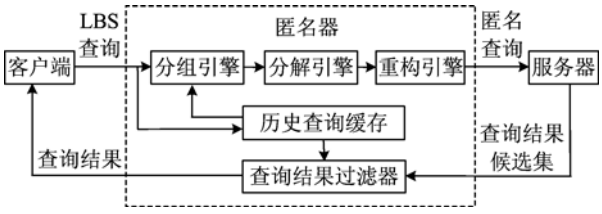

图 3　基于分解重构技术的 LBS 匿名系统结构

Fig. 3　Architecture based on anatomy and reconstruction technique

2.3　分解与重构

定义 2.2　分解.把分组得到的等价类 E_i 中的每个查询 $q\{\mathrm{Uid},\mathrm{Loc},\mathrm{Qry},\mathrm{Cont}\}$,垂直划分为若干部分的过程称为分解.分解的策略有 3 个:

分解策略 1:把查询 q 分解为 $q(\mathrm{Uid})$,$q(\mathrm{Loc})$,$q(\mathrm{Qry})$三部分,构成 3 个集合,分别记为:集合 S_{Uid},集合 S_{Loc},集合 S_{Qry};

分解策略 2：把查询 q 分解为 q(Uid)，q(LocQry)两部分，构造 2 个集合，分别记为：集合 S_{Uid}，集合 S_{LocQry}；

分解策略 3：把查询分解为 q(UidLoc)，q(Qry)两部分，构造 2 个集合，记为：集合 S_{UidLoc} 和 S_{Qry}；

定义 2.3 重构. 设 E_i 为分组后生成的等价类，S_{Uid}，S_{Loc}，S_{Qry}，S_{LocQry}，S_{UidLoc} 为 E_i 分解后生成的集合，由这些集合构造新的查询集的过程，称为重构. 重构的策略有 3 个：

重构策略 1：$D_1 = S_{Uid} \times S_{Loc} \times S_{Qry}$；

重构策略 2：$D_2 = S_{Uid} \times S_{LocQry}$；

重构策略 3：$D_3 = S_{UidLoc} \times S_{Qry}$.

用户的 LBS 查询请求经过分解重构后，用户信息就会与多个位置、多个查询重构出多个新的 LBS 查询，再将重构得到的新 LBS 查询发送给 LBSP. 这样攻击者即使获得了这些查询信息，也无法推断出发出该查询的具体用户以及该用户的真实位置，从而保护了个体的位置隐私和查询隐私.

2.4 分解重构方法的性能分析

分解重构技术的 3 个步骤是有机的整体，分解策略与重构的 3 个策略相对应，通过不同的重构策略来满足隐私约束. 本质上，重构分解技术是通过在查询集中添加虚假查询来实现隐私保护的，其带来的负面作用就是会生成大量的虚假查询，增加 LBSP 的负担.

下面对分解重构技术进行分析，假设分组中的 LBS 查询个数为 k，这 k 个查询的 q(Uid)，q(Loc)，q(Qry)均不相同，并且假设经过分组处理后产生的分组已经满足匿名级别的约束. 下面从安全性和真实查询率 2 个角度来分析 3 种重构策略.

(Ⅰ) 重构策略 1

重构策略 1 的查询集 $D_1 = S_{Uid} \times S_{Loc} \times S_{Qry}$，其优势是安全性高，每个位置都有 k 个不可区分的个体，攻击者推断出个体的具体位置的概率不高于 $1/k$，实现了位置隐私保护；另外，任何一个查询也对应 k 个不可区分的个体，攻击者推断出发出该查询个体的概率不高于 $1/k$，实现了查询隐私保护. 重构策略 1 也可以抵制一定的背景知识，例如，若攻击者具备某个体位置的背景知识，则推断出该个体发出查询的内容概率也不高于 $1/k$. 即在位置隐私泄密的情况下，重构策略 1 的匿名仍旧可以保护查询隐私.

该策略构造 $k*k*k$ 个查询，即个体在每个位置(k 个位置)发出 k 个查询. 其中只有 k 个查询是真实的，其他 $k*k*k-k$ 个查询是虚假的. 真实查询占整个查询集的比率是 $1/(k*k)$.

该策略的缺点是产生了大量的虚假查询，增加了 LBSP 的负担，一个改进的方法是，采用随机技术，从 $k*k*k$ 个查询中选出部分查询(包括真实查询)构成匿名查询集.

(Ⅱ) 重构策略 2

重构策略 2 的查询集 $D_2 = S_{Uid} \times S_{LocQry}$，每个位置对应 k 个不可区分的个体，满足 k-匿名约束，攻击者推断出个体的具体位置的概率不高于 $1/k$，实现了位置隐私保护；任一个查询均对应 k 个不可区分的个体，隐私攻击者推断出发出该查询的个体的概率不高于 $1/k$，实现了查询隐私保护. 不同于策略 1，若位置隐私泄密，策略 2 的匿名集的查询隐私也就泄密了，因为该策略下，查询位置和查询内容是一一对应的.

该策略将构造 $k*k$ 个查询，k 个个体中的某个体在每个位置(k 个位置)发出 1 个查询，其中只有 k 个查询是真实的，$k*k-k$ 个查询是虚假的. 真实查询占整个查询集的比率是 $1/k$.

(Ⅲ) 重构策略 3

重构策略 3 的查询集 $D_3 = S_{UidLoc} \times S_{Qry}$，每个位置对应 1 个用户，没有实现位置 k-匿名，因此不能保护个体的位置隐私，任一个查询也对应 k 个不可区分的个体，隐私攻击者推断出发出该查询的个体的概率不高于 $1/k$，实现了查询隐私保护，所以该策略只适合保护查询隐私. 即使在位置隐私泄密的情况下，仍旧可以保护查询隐私.

该策略将构造 $k*k$ 个查询，k 个个体中的每个个体在所在位置发出 k 个查询，其中只有 k 个查询是真实的，$k*k-k$ 个查询是虚假的. 真实查询占整个查询集的比率是 $1/k$.

3 种策略的性能分析见表 1.

表 1 3 种重构策略的性能分析

Tab. 1 Performance analysis of three strategies

策略	查询集大小	真实查询率	位置隐私	查询隐私	位置隐私泄密的查询隐私
重构策略 1	$k*k*k$	$1/(k*k)$	Yes	Yes	Yes
重构策略 2	$k*k$	$1/k$	Yes	Yes	No
重构策略 3	$k*k$	$1/k$	No	Yes	Yes

3 匿名模型

分解重构方法将 LBS 查询分解为多个部分，然后再重新组合，在实现匿名模型方面比时空伪装方

法更加灵活,下面讨论利用分解重构方法构造的匿名模型.

3.1 位置 k-匿名模型

定义 3.1 位置 k-匿名. 当一个用户的位置信息与其他至少 $k-1$ 个用户的位置信息不可区分时,则称该用户是位置 k-匿名的.

位置 k-匿名可以保护个体的位置隐私,重构策略 1 和策略 2 均能实现位置 k-匿名.

3.2 位置 l-多样性

若 k 个个体处于不同的物理位置,但在语义上是相同的,比如用户可能同处于一个医院大楼的不同位置,这样即使满足位置 k-匿名,攻击者仍旧可以推断出发出该请求的用户在医院,用户的位置隐私泄漏了,我们称这种攻击为位置同质性攻击,位置 l-多样性可以用来抵制位置同质性攻击.

定义 3.2 位置 l-多样性. 当等价类中的位置信息在语义上至少有 l 种不同位置,则称该用户满足位置 l-多样性.

利用分解重构技术实现位置 l-多样性,需要在分组阶段,将至少 l 个语义上位置不同的查询放到同一个等价类中,即让每个等价类都至少存在 l 个语义上不同的位置. 再根据隐私约束的需要,选择其中的一个重构策略实现位置 l-多样性.

3.3 查询 m-多样性

若等价类中所有的查询对象都是相同的,比如说都是查找最近的医院,那么即使满足位置 k-匿名,攻击者还会推断出个体的查询对象,我们称这种攻击为查询同质性攻击. 另外,并非所有的查询对象都是敏感的,比如若查询对象是医院,查询对象是敏感的,若查询对象是加油站,则查询对象是不敏感的. 查询对象的敏感性一般可以由用户来定义,若用户发出查询时,要求保护查询隐私,则该查询对象是敏感的,称该查询为敏感查询;否则是不敏感的. 为抵制查询同质性攻击,我们提出查询 m-多样性模型.

定义 3.3 查询 m-多样性. 设 E_i 为分组得到的等价类,若 E_i 中查询对象 obj 不少于 m 种,则称敏感查询对象 obj 满足查询 m-多样性.

利用分解重构技术实现查询 m-多样性,需要在分组阶段,使每个等价类至少存在 m 种不同的查询对象.

3.4 复合约束匿名模型

基于以上的定义,根据隐私需求,可以定义一些复合的匿名模型.

定义 3.4 (k,l)-匿名. 若分组得到的等价类 E_i 至少包含 k 个不同的 LBS 请求查询,且 E_i 满足位置 l-多样性,则称 E_i 是满足(k,l)-匿名的.

定义 3.5 (k,m)-匿名. 若分组得到的等价类 E_i 至少包含 k 个不同的 LBS 请求查询,且 E_i 满足查询 m-多样性,则称 E_i 是满足(k,m)-匿名的.

定义 3.6 (l,m)-匿名. 若分组得到的等价类 E_i 满足位置 l-多样性和查询 m-多样性,则称 E_i 是满足(l,m)-匿名的.

定义 3.7 (k,l,m)-匿名. 若分组得到的等价类 E_i 至少包含 k 个不同的 LBS 请求查询,且 E_i 满足位置 l-多样性和查询 m-多样性,则称 E_i 是满足(k,l,m)-匿名的.

以上复合约束匿名模型都是在分组阶段通过构造满足隐私约束的等价类来实现的.

易知,(k,l,m)-匿名模型是其他 6 种匿名模型的超集,当 $l=m=1$ 时,(k,l,m)-匿名模型退化为位置 k-匿名,当 $k=m=1$ 时,(k,l,m)-匿名模型退化为位置 l-多样性,当 $k=l=1$ 时,(k,l,m)-匿名模型退化为查询 m-多样性. 类似地,当 $m=1$,(k,l,m)-匿名退化为(k,l)-匿名模型;当 $l=1$,退化为(k,m)-匿名模型;当 $k=1$,退化为(l,m)-匿名模型.

4 基于分解重构技术的匿名化算法

4.1 基于网格的位置 l-多样性

为简化位置语义的定义工作,本文提出基于网格的位置 l-多样性的定义,该方法首先将用户区域按照一定的粒度划分为多个网格,认为不同网格的位置语义是不同的,基于网格来实现位置 l-多样性.

4.2 基于查询对象类别的查询 m-多样性

查询 m-多样性主要用来保护查询隐私,并非所有的 LBS 查询都需要保护查询隐私,LBS 查询的查询隐私是否需要保护是用户提交查询时在隐私约束中定义的,因此,查询 m-多样性主要针对需要保护查询隐私的 LBS 查询.

通常情况下,判断查询对象是否语义相同,不能简单地利用字符串匹配来实现,比如:距离我 100 m 以内的医院和距离我 500 m 以内的诊所,虽查询对象用的是不同词,但在语义上是相似的. 我们将敏感查询对象按语义分类,如敏感查询对象分为以下 6 类:健康状况、宗教信仰、政治倾向、性偏好、娱乐习惯、其他等,基于敏感查询对象类别实现查询 m-多样性.

4.3 基于多维桶技术的 (k,l,m)-匿名化算法

实现(k,l,m)-匿名模型,需要考虑 3 个匿名参

数：k,l,m. 分组生成的等价类需要满足的约束有 3 个：①等价类中查询的个数不少于 k；②等价类中查询的 Loc 信息在语义上至少有 l 个不同的值；③等价类中查询对象 obj 不少于 m 种；④等价类中查询的隐私类别相同. 匿名化算法的思想是构造满足上述 3 个约束的等价类. 由定义 3.4 可知，只需在分组阶段让每个等价类满足一定的约束，就可利用重构策略生成满足 (k,l,m)-匿名的查询集.

构造等价类时，需要考虑 LBS 查询中两个属性的约束：LBS 查询的 Loc 属性和查询对象 obj 属性. 我们采用分桶技术来构造等价类，其思想是首先构造一个 2 维桶，然后按照某策略从不同的桶中取出 k 个查询构成等价类，循环这个过程，直到无法构成满足约束的新的等价类. 构造 2 维桶的方法如下：Loc 属性和 obj 属性各对应桶的一维，每个 LBS 查询的这两个属性值分别对应到相应的桶中，记桶为 buk＜Loc,obj＞，用 size(buk＜Loc,obj＞)表示桶中查询个数. 例如，假设用户区域经过网格划分后，分为 4 个区域，分别是 1,2,3,4；obj 的类别分为 6 类，分别是健康状况(h)、宗教信仰(r)、政治倾向(p)、性偏好(s)、娱乐习惯(b)、其他(o)，分别用 h,r,p,s,b,o 表示，其他类表示用户不需要保护查询隐私. 设表 2 为当前得到的查询，所有查询采用 1.1 节的形式化方法，则表 3 是对表 2 查询分桶的结果.

表 2 查询内容

Tab. 2 Queries

id	查询内容
q_1	$(u_1,1,(h,100,now,0),(3,11))$
q_2	$(u_2,2,(h,200,now,0),(3,11))$
q_3	$(u_3,3,(h,300,now,0),(3,11))$
q_4	$(u_4,1,(r,100,now,0),(3,11))$
q_5	$(u_5,3,(r,200,now,0),(3,11))$
q_6	$(u_6,4,(p,100,now,0),(3,11))$
q_7	$(u_7,4,(p,200,now,0),(3,11))$
q_8	$(u_8,4,(s,100,now,0),(3,11))$
q_9	$(u_9,1,(b,100,now,0),(3,11))$

表 3 2 维桶及容量

Tab. 3 Bucket and capacity of two dimensions

分区	身体状况 h	宗教信仰 r	政治倾向 p	性偏好 s	娱乐习惯 b	其他 o
1	$\{q_1\}$	$\{q_4\}$			$\{q_9\}$	
2	$\{q_2\}$	$\{q_5\}$				
3	$\{q_3\}$					
4			$\{q_6,q_7\}$	$\{q_8\}$		

从各桶中选择查询生成等价类可以依据最大桶优先策略(maximal-bucket first, MBF)[20]，即优先从 size(buk＜Loc,obj＞)最大的桶中选择查询. 例如，针对表 3 的 2 维桶的情况，首先选择桶中查询个数最大的桶，即桶＜4，p＞，包括查询 $\{q_6,q_7\}$，任取一个，比如取 q_6，然后将该桶所在的行和列屏蔽，从未屏蔽的最大桶中再选一个查询，比如选择 q_1，剩下只有桶＜2，r＞未屏蔽，从中选择 q_5，即构造了一个等价类 $\{q_6,q_1,q_5\}$，然后清空桶的屏蔽记录，再按最大桶优先策略构造下一个等价类，直到剩余非空桶的个数不足 k 个，结束循环. 然后在满足 (k,l,m)-匿名的前提下，将剩余的查询加入已经构造好的等价类中. 最后依据某策略对每一个等价类的元组进行分解和重构.

最大桶优先匿名算法(maximal-bucket first anonymity algorithm, MBFAA)描述见算法 4.1.

算法 4.1 最大桶优先分组匿名算法(MBFAA)

输入：查询集 D，匿名化参数 k,l,m

输出：匿名查询集

```
/**分桶分组阶段**/
①根据 Loc 值和 obj 值，构造二维桶 buk＜Loc,obj＞；
②循环执行
③  对所有桶设未屏蔽标记，分组 E=∅；
④  equi_class_size=max(k,l,m)
⑤  for i=1 to equi_class_size
⑥    if 存在未被屏蔽的非空桶；
⑦      在未屏蔽桶中选择最大的桶 buk，从中提取一个查询 q；
⑧      E=E+{q}；
⑨      buk=buk−{q}；
⑩      size(buk)=size(buk)−1；
⑪      if (i＜=l)
⑫        屏蔽 q 在 loc 维上的所有桶；
⑬      if(i＜=m)
⑭        屏蔽 q 在 obj 维上的所有桶；
⑮    else 结束分桶过程；
⑯    end if
⑰  end for
⑱  将所选 equi_class_size 个查询构成新的等价类；
⑲直到非空桶的个数小于 equi_class_size；
/**剩余查询处理阶段**/
⑳for 剩余元组
㉑  循环将剩余查询添加到已经完成的各等价类中，使各等价类大小近似相等；
㉒end for
```

/ * * 分解重构阶段 * * /
㉓选择分解策略对各个等价类进行分解;
㉔选择重构策略重新构造新的查询集;
㉕返回新的查询集

设初始查询集的规模为 n,算法第①步构造多维桶,需要扫描所有的查询,时间复杂度为 $O(n)$;步骤②～⑲,循环从每个桶中选一个查询构造等价类,步骤⑳～㉒,处理剩余的查询,每个查询处理一次,时间复杂度为 $O(n)$;步骤㉓,每个查询分解一次,时间复杂度为 $O(n)$;步骤㉔按策略 1 重构,可以构造 $n*k*k$ 个查询,按策略 2 和策略 3,可构造 $n*k$ 个查询,时间复杂度为 $O(n*k*k)$ 或 $O(n*k)$,所以整个算法的时间复杂度约为 $O(n)$.

易知,通过设置 MBFAA 算法中的 k,l,m,可以实现位置 k-匿名,位置 l-多样性,查询 m-多样性,(k,l)-匿名,(k,m)-匿名,(l,m)-匿名,(k,l,m)-匿名.

5　评估模型

本文主要从安全性和服务质量 2 个角度对 ARBA 方法进行评估. 这两个指标是相互矛盾的,对于同一组查询,一般匿名查询集越大,攻击者识别出真实查询的概率就越小,安全性越强,但由于查询集大,虚假查询过多,系统负担重,响应时间长导致服务质量差. 下面分别讨论安全性和服务质量的评估方法.

5.1　安全性评估模型

安全性是指攻击者从匿名查询中推断出发出查询用户的位置或发出查询的用户身份的概率,这个概率越小越安全.

定义 5.1　真实查询率. 真实查询率是指真实查询在整个匿名查询集中的比率,可用式(1)来度量.

$$\text{Security}=\frac{N_{\text{real}}}{N_{\text{anon}}} \tag{1}$$

式中,N_{real}表示真实查询个数,N_{anon}表示匿名查询集的大小. ARBA 方法安全性可以用真实查询率来度量. 相对而言,重构策略 1 生成的匿名查询集的真实查询的概率比较低,因此比较安全.

5.2　服务质量评估模型

服务质量可以从用户获得查询结果的时间延迟和精度 2 个角度衡量,一般延迟越小,查询结果越精确,服务质量越好,本文利用以下 2 个指标度量 ARBA 的服务质量.

(Ⅰ) 平均等价类大小

从查询结果的精度角度上分析,ARBA 方法比时空伪装方法的查询结果更精确,因为 ARBA 方法返回的是真实查询的结果,而时空伪装的方法返回的是匿名器在候选集筛选出来的结果.

由于 ARBA 发送了很多虚假查询,会增加 LBSP 服务器负担,延迟响应时间,因此从服务质量的角度,应尽量减少虚假查询的比率,而减少虚假查询的比率的方法之一就是尽量减少等价类的大小. 为此,用平均等价类大小作为 ARBA 服务质量评估的方法之一,其度量方法如下:

$$\text{Avereisize}=\frac{\sum_{i=1}^{g}|e_i|}{g} \tag{2}$$

式中,$|e_i|$为等价类 E_i 的大小,g 为等价类的个数.

(Ⅱ) 查询冗余率

MBFAA 算法将查询集划分为多个等价类,并要求每个等价类应满足一定的匿名约束,比如对于 (k,l,m)-匿名模型,等价类需要包含至少 k 个不同的查询,且满足位置 l-多样性和查询 m-多样性,而实际等价类所包含的查询个数会比模型要求的多,即含有冗余查询,影响服务质量. 冗余查询越多,服务质量越差,为此我们引入查询冗余率的定义. 设每个等价类的实际查询个数为 $|e_i|$,等价类的个数为 g,则 7 种匿名模型查询冗余率的度量方法如下:

位置 k-匿名模型:

$$\text{Red\_ratio}=\sum_{i=1}^{g}\frac{|e_i|-k}{|e_i|} \tag{3}$$

位置 l-多样性模型:

$$\text{Red\_ratio}=\sum_{i=1}^{g}\frac{|e_i|-l}{|e_i|} \tag{4}$$

查询 m-多样性模型:

$$\text{Red\_ratio}=\sum_{i=1}^{g}\frac{|e_i|-m}{|e_i|} \tag{5}$$

(k,l)-匿名模型:

$$\text{Red\_ratio}=\sum_{i=1}^{g}\frac{|e_i|-\max(k,l)}{|e_i|} \tag{6}$$

(k,m)-匿名模型:

$$\text{Red\_ratio}=\sum_{i=1}^{g}\frac{|e_i|-\max(k,m)}{|e_i|} \tag{7}$$

(l,m)-匿名模型:

$$\text{Red\_ratio}=\sum_{i=1}^{g}\frac{|e_i|-\max(l,m)}{|e_i|} \tag{8}$$

(k,l,m)-匿名模型:

$$\text{Red\_ratio}=\sum_{i=1}^{g}\frac{|e_i|-\max(k,l,m)}{|e_i|} \quad (9)$$

6 实验结果与分析

6.1 实验环境与配置

实验采用的 PC 机内存为 2 GB,奔腾双核处理器主频为 3 GHZ.我们设计一个匿名器,每 2 s 匿名器匿名化一次查询,匿名器设置一个历史查询缓冲区,可存储10 000个历史查询,并设置一个当前查询缓冲区,存储当前待匿名化的查询.若匿名器接收到的新查询个数小于 N_{min},则等待下一次处理;若个数小于阈值 N_{qry},则从历史查询缓冲区中选择部分查询放入当前查询缓冲区.

本文设计了一个查询生成器,用于批量生成查询,设查询格式 q=(uid,loc,(obj,dist,time,prof),privacy_type),每 2 s 生成一次查询,每次生成查询个数少于 100,用户个数小于10 000,位置网格为 10 * 10=100(4.1 节),查询对象为 6 类(4.2 节).

首先利用查询生成器随机生成一批查询,分组引擎根据不同的匿名模型进行分组,重构引擎再根据不同的重构策略调用相对应的分解引擎进行查询重构,实验设置参数 l=3,m=3,N_{min}=10,N_{qry}=100.

6.2 安全性分析

7 种匿名模型中,k-匿名模型、(k,l)-匿名模型、(k,m)-匿名模型、(k,l,m)-匿名模型更有实际意义,因此实验比较这 4 种匿名模型.

图 4～6 为三种重构策略下,k 从 3 变化到 10,4 种匿名模型的真实查询率变化情况.其中,k 是用户指定的匿名约束,即不可区分查询的个数.从实验结果可以看出,重构策略 1 的真实查询率最低,因为策略 1 将查询按照{ID,location,query}属性分解并采用全相连形式构造匿名查询,构造虚假查询最多,安全性最高.在同一种策略下,(k,l,m)-匿名模型真实查询率最低,安全性最高,因为该模型在分组过程中需要更多的查询以同时满足位置 k-匿名、位置 l-多样性和查询 m-多样性.由于该实验位置参数网格设定为 100,查询对象设置为 6 类,而需要满足的约束分别为位置 3-多样性和查询 3-多样性,显然位置 l-多样性更容易满足,因此 k-匿名模型、(k,l)-匿名模型的真实查询率比(k,m)-匿名模型、(k,l,m)-匿名模型较高.随着 k 增大,3 种策略下 4 种匿名模型的真实查询率均逐渐减小.

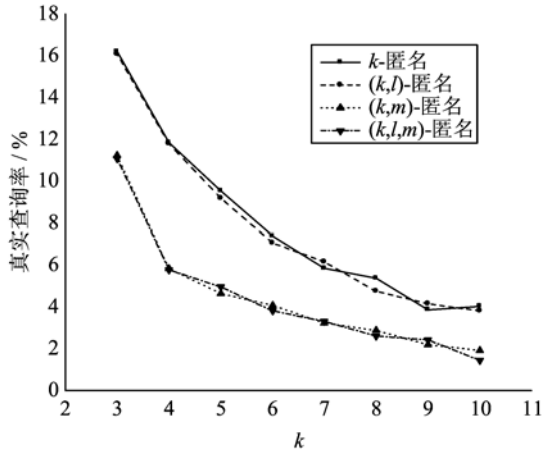

图 4 重构策略 1 真实查询率

Fig. 4 Real query rate of policy 1

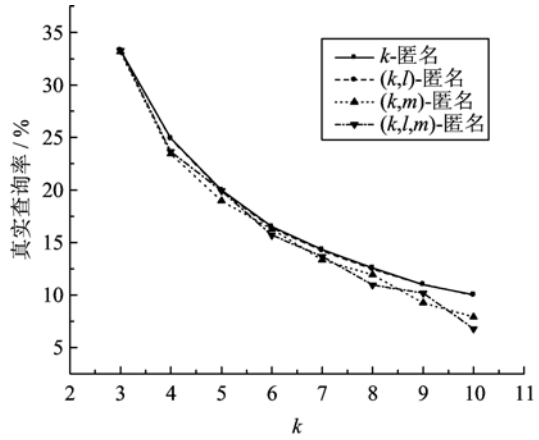

图 5 重构策略 2 真实查询率

Fig. 5 Real query rate of policy 2

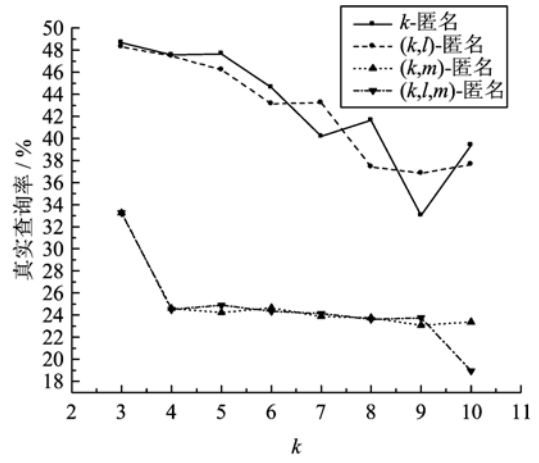

图 6 重构策略 3 真实查询率

Fig. 6 Real query rate of policy 3

6.3 服务质量分析

服务质量从平均等价类大小和查询冗余率角度进行度量,这两个指标与重构策略无关.

(Ⅰ)平均等价类大小

图 7 为 k 从 3 变化到 10,4 种匿名模型的平均等价类大小的变化情况.从图 7 可以看出,等价类的大小随着 k 近似呈线性变化.与真实查询率类似,位置 l-多样性更容易满足,在同一种策略下,k-匿名模型、(k,l)-匿名模型的平均等价类大小比(k,m)-匿名模型、(k,l,m)-匿名模型小.

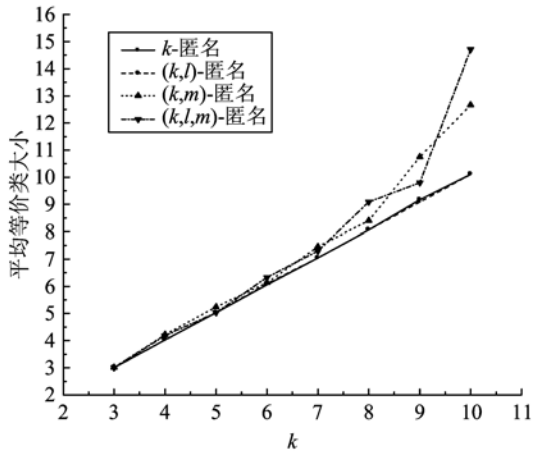

图 7　参数 k 变化下的平均等价类大小

Fig. 7　Average size of equivalence class of three strategies

(Ⅱ)查询冗余率

图 8 为 3 种重构策略下,k 从 3 变化到 10,4 种匿名模型的查询冗余率变化情况.从图 8 可以看出,4 种匿名模型的冗余查询率都不高.这是因为等价类大小接近于匿名约束值,比如在(k,l,m)-匿名模型中,等价类大小$|e_i|$与 $\max(k,l,m)$值接近,所以由式(9)可以得出该匿名模型的查询冗余率不高.另外,当 k 变化时,分组引擎产生等价类的大小具有一定的随机性,导致冗余率大小具有一定的不稳定性.随着 k 增大,冗余率大体呈相应增长趋势.与真实查询率类似,4 种匿名模型中,位置 l-多样性更容易满足,因此(k,m)-匿名模型、(k,l,m)-匿名模型的查询冗余率较 k-匿名模型、(k,l)-匿名模型高.

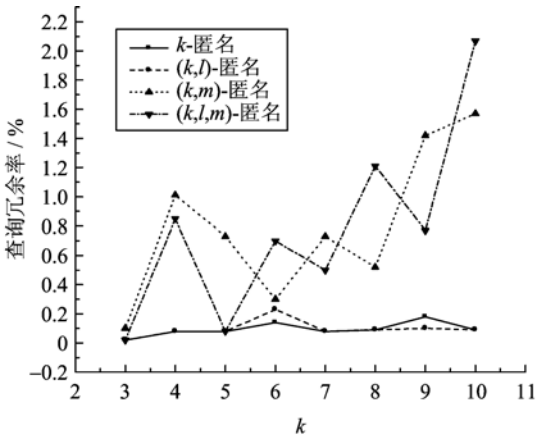

图 8　参数 k 变化下的查询冗余率

Fig. 8　Query redundancy of three strategies

7　结论

针对现有的基于位置时空伪装技术的 LBS 隐私保护方法的不足,本文提出了分解重构的匿名技术,同时提出了一组匿名模型集,并提出了实现分解重构技术的匿名模型的实现算法.实验表明,本文提出的重构分解技术可以很好地实现各种匿名模型.

参考文献(References)

[1] Yankee Group, eMarketer. ISACA: LBS services use increases, but privacy concerns remain [EB/OL]. http://www.199it.com/archives/51801.html.

[2] Sythoff J T, Morrison J. Location-Based Services Market Forecast 2011-2015 [M]. Brazil: Pyramid Research Publisher, 2011.

[3] Marist Poll. Marist Institute for Public Opinion (MIPO): Half of social networks online concerned about privacy [EB/OL]. http://maristpoll.marist.edu/714-half-of-social-networkers-online-concerned-about-privacy/.

[4] 谈嵘, 顾君忠, 杨静, 等. 移动社交网络中的隐私设计研究[J]. 软件学报, 2010, 21(ZK): 298-309.

[5] Shin K G, Ju X E, Chen Z G, et al. Privacy protection for users of location-based services wireless communications [J]. IEEE Wireless Communications, 2012, 19(1): 30-39.

[6] Gruteser M, Grunwald D. Anonymous usage of location-based services through spatial and temporal cloaking[C]// Proceedings of the 1st International Conference on Mobile systems, Applications and Services. ACM Press, 2003: 31-42.

[7] Gedik B, Liu L. Location privacy in mobile systems: A personalized anonymization model[C]// Proceedings of the 25th International Conference on Distributed Computing Systems. Columbus, USA: IEEE Press, 2005, 620-629.

[8] Kalnis P, Ghinita G, Mouratidis K, et al. Preventing location-based identity inference in anonymous spatial queries[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(12):1 719-1 733.

[9] Gkoulalas-Divanis A, Kalnis P, Verykios V S. Providing k-anonymity in location based services[J]. ACM SIGKDD Explorations NewsLetter, 2010, 12(1): 3-10.

[10] Bettini C, Wang X S, Jajodia S. Protecting privacy against location-based personal identification [C]// Proceedings of the 2nd VLDB Workshop. Trondheim, Norway: Springer, 2005: 185-199.

(下转第 562 页)