

基于语义的位置服务隐私保护综述

马明杰¹, 杜跃进^{1,2}, 李风华¹, 刘佳文³

(1. 中国科学院信息工程研究所信息安全国家重点实验室, 北京 100093;
2. 阿里巴巴集团安全部, 浙江 杭州 310052;
3. 中央财经大学信息学院, 北京 100081)

摘要: 位置服务发展引起的隐私泄露问题已不容忽视。基于语义的位置隐私保护是位置隐私保护工作中的重要组成部分。首先总结了近年来基于语义的位置隐私保护研究工作, 包括隐私保护设计中的关键因素、语义保护可行性、相关概念等; 随后讨论了基于语义的位置隐私保护机制、语义相似度测量方法、隐私保护度量方法等; 最后展望了基于语义的位置隐私保护未来的研究方向。

关键词: 信息安全; 语义; 隐私保护; 基于位置服务

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2016.00088

Review of semantic-based privacy-preserving approaches in LBS

MA Ming-jie¹, DU Yue-jin^{1,2}, LI Feng-hua¹, LIU Jia-wen³

(1.State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China;
2. Security Department, Alibaba Group, Hangzhou 310052, China;
3. School of Information, Central University of Finance and Economics, Beijing 100081, China)

Abstract: The development of location-based services (LBS) caused serious privacy concerns. Semantic-based privacy-preserving approaches is an important part of privacy protection in LBS. The recent work on location privacy protection based on semantics were summarized, including key factors in design of privacy protection, the feasibility of semantic protection, related concepts. Then semantic-based location privacy-preserving mechanisms, measurement methods of semantic similarity, evaluation index of privacy for semantic location privacy were discussed. Finally, some future research directions of semantic-based location privacy protection were prospected.

Key words: information security, semantics, privacy-preserving, LBS

1 引言

随着智能可穿戴设备、车联网等新技术的出

现, 移动互联网得到迅速发展。2016 年, QuestMobile 发布的《春季 APP 实力榜》^[1]显示: 截止到 2016 年 3 月, 国内移动互联网月度活跃用

收稿日期: 2016-09-20; 修回日期: 2016-11-06。通信作者: 李风华, lfh@iie.ac.cn

基金项目: 国家自然科学基金委—广东联合基金资助项目 (No.U1401251); 国家高技术研究发展计划 (“863” 计划) 基金资助项目 (No.2015AA016007); 国家自然科学基金—青年科学基金资助项目 (No.61502489); 国家 “核高基” 科技重大专项基金资助项目 (No.2015ZX01029101); 中国科学院大学生创新实践训练计划基金资助项目 (No.1188005088)

Foundation Items: The Key Program of the National Natural Science Foundation of China-Guangdong Union Foundation (No.U1401251), The National High Technology R&D Program of China (863 Program) (No.2015AA016007), The National Natural Science-Youth Science Foundation of China (No.61502489), The National Science and Technology Major Project of China (No.2015ZX01029101), The Innovative Practice Project of College Students in Chinese Academy of Sciences (No.1188005088)

户数突破 9.27 亿, CNIC 发布的《第 37 次中国互联网络发展状况统计报告》^[2]显示, 网络预约出租车用户规模为 9 664 万人。由此可见位置服务的用户在国内非常庞大。

位置服务是把双刃剑, 位置服务的普及在给人们带来便利的同时, 也给人们的隐私保护带来了严峻考验。2014 年 7 月, 中央电视台报道了苹果搜集用户位置隐私的问题, 报道中称苹果公司详细记录了用户的移动轨迹以及位置, 并将其存储在未加密的数据库中。2014 年, 英国综合性日报《卫报》(The Guardian) 也曝出即使用户已经选择关闭定位功能, 匿名应用 Whisper 也会泄露用户的地理位置。在国内, 微信“附近的人”功能使其他用户可轻松定位用户位置, 因此导致越来越多的犯罪事件。由此, 位置服务中的隐私保护刻不容缓。

针对位置服务引起的隐私泄露问题, 国内外学者提出了多种解决方案。现有基于位置服务的隐私保护方法主要分为数据加密、伪地址、空间转换以及匿名区域等。这些方法大多针对位置本身, 缺乏对位置本身与用户的关系以及位置背后代表的意义的探究, 难以捕捉人类活动在现实生活中的意义, 为此, 越来越多的研究者开始研究基于语义的位置隐私保护, 以期实现更深层次的保护效果。

2 位置服务隐私保护概述

用户在使用位置服务时, 会在一定程度上泄露其位置信息, 有可能会泄露用户的敏感信息。例如, 根据用户历史位置数据, 很容易推断出用户经常出沒的地点、经常行走的路线等, 从而推断出用户的家庭住址、人际关系、身体状况等敏感信息。

2.1 隐私保护关键因素

在考虑设计隐私保护问题上往往要考虑以下关键因素。

1) 隐私保护与服务质量间的平衡。在 LBS 中, 隐私保护与服务质量的平衡是需要谨慎处理的重要问题。用户要获得位置服务, 需要将自己的位置信息以某种方式提交给服务商或其他用户, 这就存在泄露隐私的风险。用户要求的服务质量越高, 泄露具体信息的风险越高。例如, 要

求查询 100 m 内的工商银行和 1 000 m 内的银行服务所需的用户地理位置精度以及攻击者可判断下一刻用户位置的确定性是不一样的。同理, 细粒度的隐私保护在一定程度上也会牺牲服务质量。如何权衡服务质量和隐私之间的关系是保护位置隐私需要关注的问题。

2) 隐私保护效果与资源消耗间的平衡。用户在发起位置服务请求时, 一般对服务的响应速度有一定的要求, 如果响应速度过慢, 势必会影响用户体验。现有的隐私架构无论是 P2P 架构还是 TTP 架构, 都需要在保证隐私效果的同时, 考虑资源消耗问题。

基于位置的服务具有查询和数据更新频率高等特点, 因此传统的数据隐私保护方法并不适用。除去加密的方法, 现有的基于位置服务的隐私保护技术主要分为 3 类。

1) 伪造数据。主要思想是用户要么发送与真实位置相关的假数据, 要么发送包含有真实位置的多个数据给位置服务提供商以保护其真实位置。伪造数据可通过泛化位置数据^[3,4]、添加噪声(如差分隐私方法)^[5~7]、基于语义的隐私保护等方法实现^[8~37]。

2) 空间转换。其主要思想是将位置信息转换到另一个空间, 即将位置信息转换为包含查询和数据解码的空间关系, 如利用 Hilbert 曲线转换形成隐蔽空间^[38,39]。

3) 匿名区域。其主要思想是将用户的位置模糊成 K -匿名空间。匿名空间的构成有 2 种形式: 一种是以 P2P (peer-to-peer)^[40]的方式构成用户间的匿名空间; 另一种是以路网为基础, 构成路网的匿名空间^[41,42]。

表 1 列举了现有的位置服务的隐私保护方案, 并对比了现有主流方案的优缺点。现有的主流方案根据是否依靠可信第三方 (TTP, trusted third-party), 可以分为包含可信第三方和点对点结构两类方法, 其中, 依靠可信第三方的方法占大多数, 但有一定程度的风险; 而不依靠第三方的方法, 又受到各种技术上或开销上的限制。其中, 基于语义的隐私保护方法因其计算量比较大, 属于依靠可信第三方的方法。该方法虽然同样有依靠可信第三方的缺点, 但其在数据可用性以及

表 1 现有位置服务隐私保护方案比较								
分类	隐私保护方法	优点	缺点	隐私保护等级	是否依靠可信第三方	能否应用到连续场景	是否基于密码	K-匿名
伪地址	泛化	简单易实现	易受到攻击	低	是	是	否	是
	差分隐私	保护等级高	实现效率不高	高	是	否	否	否
	基于语义的隐私保护	数据可用性高	计算过程繁琐	高	是	是	否	是
空间转换	Hilbert 曲线	抵御推理攻击	计算开销大	低	是	否	否	是
匿名区域	P2P	无第三方泄露风险	受距离的限制	高	否	是	否	是
	Mix-zone	简单易实现	受选择区域大小影响较大	低	是	是	否	是

隐私保护等级上都远胜过其他的保护方法。相较于差分隐私来说，基于语义的方法更能适用于实际的环境中，相较于同样能实现高等级隐私保护的 P2P 方法，基于语义的保护方法不受距离限制，对用户来说更加通用。

2.2 基于语义的位置保护的优势

针对位置服务的隐私保护问题，很多学者提出了解决方案。例如通过对原始数据扰动和匿名产生假位置^[3]，在同一地方发送包含不同服务属性的多个 dummy 查询，迷惑攻击者^[4]，以及经典的如 P2P 空间匿名的 K-匿名方法^[40]等，这些方法在一定程度上保护了用户的隐私，但无法应对具有一定背景知识的攻击者的推理攻击。举例说明，如图 1 所示，假设 Mary 晚上 12 点在北京医院发起 LBS 服务，若采取 K-匿名方法对其进行隐私保护，因医院人口密度较大，发起服务时生成的匿名区域可能就在医院内部，从而泄露用户隐私；若采用空间转换技术，因空间较小，最后的结果也会集中在医院附近，导致隐私泄露；若采用 L-多样性的方法，匿名区域可能包括北京医院、东单公园、礼品店、同仁医院 4 个位置，符合 3-多样性；但是攻击者结合背景知识，考虑到晚上 12 点以及停留时间，完全可以排除东单公园以及礼品店，进而推断出 Mary 在医院。在此背景下，很多研究者开始研究基于语义位置的隐私保护。

基于语义的位置隐私保护技术较其他技术有如下优点。

- 1) 隐私保护水平高，且有较高的实现效率。该方法能有效地抵抗推理攻击。不同于差分隐私，该方法可以应用在实际的隐私保护中，实现效率高。
- 2) 应用场景广泛。该方法不仅可以应用于单点

位置，也可以用于连续的位置保护即轨迹保护。基于语义的隐私保护不仅可用于位置保护，在网络搜索^[43]、数字医疗^[44]等领域也取得了较好的效果。



图 1 北京医院实例

2.3 相关概念

1) 相似性

相似性是基于语义的位置隐私中的重要概念，用来衡量位置等事物之间的相像程度。在具体方案中对相似度会有不同的定义。

目前，在位置服务中主要有 2 种相似性：语义相似性和地理位置相似性。假设 Alice、Bob、John 这 3 个用户某一天的轨迹如表 2 所示。

表 2 不同用户某天轨迹			
时间	地点		
	Alice	Bob	John
白天	闵庄（办公）	中关村（办公）	香山（娱乐）
傍晚	闵庄健身房（娱乐）	清华附近 KTV（娱乐）	闵庄附近麦当劳（餐馆）
晚上	玉泉路（住宅区）	西二旗（住宅区）	玉泉路医院（医疗）

由表 2 可知，虽然 Alice 与 Bob 的活动范围在地理位置上有一定的差异，但其活动轨迹根据

语义信息可总结为：办公地点（白天）→娱乐场所（傍晚）→住宅区（晚上），在轨迹上其所做的事情类似，这种情况可以认为 Alice 与 Bob 两者轨迹语义相似性较高，地理位置相似性较低。对比 Alice 与 John，两者的活动范围围绕在闵庄、玉泉路附近，但两者轨迹上的位置语义不相同，即 2 个人活动的意义不相同，可认为两者地理相似性度较高，而语义相似性较低。

2 个位置语义相似表示在这 2 个位置有相似用户以相似的概率在相似的时间从事相似活动，如麦当劳和肯德基语义相似而学校和酒吧语义不相似。另一方面，语义相似度也是相对的，例如，麦当劳、肯德基、全聚德虽同为餐厅，但麦当劳和肯德基之间的语义相似度比全聚德要高。语义相似性的衡量根据用户或作者提取特征的不同而有所不同。语义特征可提取自上下文信息、时间或使用目的等信息。文献[8]认为位置的语义特征可用使用时间和停留时间表示。使用时间为位置的活跃时间，如餐厅一般在饭点人比较多；停留时间则表示用户在位置停留时间长短，如工作地地点的停留时间比餐厅的停留时间长。一般情况，提取的位置语义特征越多，语义分类越精确，保护用户的位置隐私能力越强。

2) 停留点

分析轨迹的语义信息时，停留点是一个很重要的特征。文献[9]给出了停留点定义，如图 2 所示，停留点表示用户曾经逗留过的位置。对于未经处理的 GPS 轨迹，每个停留点都有特别的含义，如人们生活或工作的地方、餐厅等。停留点往往反映了用户的出行目的、喜好、工作性质等。停留点有 2 种形式。

① 单点停留。它指用户在某个地点停留超过一定时间。图 2 中 P_3 点是一个典型的单点停留。单点停留一般发生在用户进入一个建筑并失去信号直到信号再次出现。

② 区域停留。用户在某个区域内停留超过一定时间。图 2 中用户 P_5 、 P_6 、 P_7 、 P_8 在一定区域范围内逗留一段时间。这种情况可能发生在旅游景区区、住宿区、办公区等地点。

停留点的提取是保护轨迹隐私中的一个重要部分，停留点选取不当，可能会泄露用户的敏感信息。

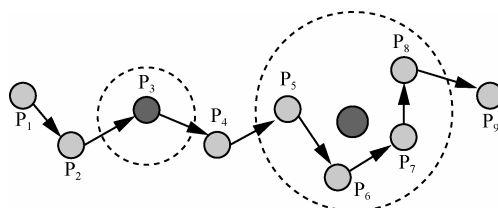


图2 停留点示意

3 基于语义的位置隐私保护机制

基于语义的位置隐私保护主要分为单点位置隐私保护和轨迹隐私保护两类，无论哪一类，其保护过程都主要分为 4 个阶段。

1) 建模。无论保护单点位置还是轨迹隐私首先要对未处理的数据进行建模。建模的具体过程包括数据清理^[45]、地图匹配^[46]、压缩数据^[47]等。隐私保护在建模的过程中需要有特殊的处理过程，例如，单点中针对路网环境中的位置分类，根据位置敏感与否、敏感度多少等，将城市网络分割成多个子图，构建单点模型；轨迹中则将真实轨迹分割成停留点和路径 2 个部分，构建轨迹模型。

2) 注解。根据位置上下文信息注解位置信息^[48~51]。注解的值可以是属性值（如火车），也可以是一个事件或目的的值（如娱乐场所）。注解可以从未处理的数据中直接计算，也可以通过从上下文中抽取等方法获得^[52]。单点的注解主要是对位置点的实际含义进行描述；而轨迹的注解则较为复杂，主要分为 2 种方式：一种是根据 POI 来注解^[48,49]，另一种是根据活动的停留点注解^[50,51]。现有的注解存在以下问题：注解的分类现在缺少一个强有力的规范；注解精确度有待提高；如何结合地理信息，使其更符合现实生活等。这些都是需要关注解决的问题。

3) 语义挖掘。语义挖掘的目的是为了更好地定义行为^[53]或位置代表的意义^[54]。根据语义相似性判别方法（如聚类、分类等），获得相应的语义类，或用相应的算法提取行为知识等，这些方法都可以用于语义挖掘。

3.1 基于语义的单点位置隐私保护

基于 LBS 的单点位置主要包括用户身份、位置坐标、查询内容等。基于语义的单点位置隐私是根据位置本身的语义实现保护位置信息的目的。

Damiani 等^[10,11]首次在单点位置隐私保护中引入语义的方法，并命名为语义感知模糊技术。该方法先将位置初步分类，计算位置所在区域是否小于定义阈值，若小于则该区域为模糊区域，大于则合并该区域的邻近单元，使其小于定义值。文献[12]用语义模型作为输入数据的规则说明，用一个中间系统调节服务提供商和数据资源之间的平衡，以保护用户隐私。

为保证即使用户位置隐私泄露，敌人也不能以很高的概率将用户和敏感信息关联，Xiao 等^[13]提出了另一种有效机制，即 *p-sensitivity*。针对该机制，用户在相同的匿名集中将拥有相同的匿名区域，并确保每个匿名集中至少有 K 个用户，以保护用户隐私。论文采用 *PE-Tree* 的方法将查询分为敏感和非敏感两类，并考虑了匿名用户位置的语义信息和查询差异性，完善了 Damiani 等方案中位置分类较粗糙的问题。

虽然 Xiao 等将查询分为敏感和非敏感，但并未考虑位置之间的差异，保护效果并不理想。为进一步阐释位置的实际含义，Haav 等^[14]采用本体论的方法定义了个人语义空间关系以体现位置差异性；Xue 等^[15]给出了语义位置的概念，利用位置的语义差异进行隐私保护。Xue 等将语义位置定义为地图中集合了相同背景信息（如年龄、性别、活动等）人的区域。文章提出利用位置差异性的概念解决一些现实生活中的情况，如要求用户身份的私人服务等，同时为提高 K -匿名的保护力度，文章需要保证每个查询至少与 L 个不同的语义位置相关。

文献[16~18]对敏感度进一步细化，如商场的敏感度要比医院的敏感度低，并提出了 *PROBE* 架构。*PROBE* 架构中用户可根据自身意愿选择相应的敏感项，其具体的实现过程如图 3 所示。首先用户通过 *Profile* 编辑器编辑自己的隐私简档，如用户可以从一个预先定义的列表中选择敏感特征类型，并制定敏感首选项。用户选择比较敏感的位置通过减少位置信息精确度保护其隐私性，而那些敏感度较低的位置，为获得更好的服务质量则可以适当降低隐私保护等级。位置匿名服务器将输入的隐私简档通过模糊引擎处理生成模糊区域。用户发起服务请求时，位置转换应用在客

户端检查用户位置，如果用户的精确位置在匿名区域，则用匿名区域代替精确的位置。

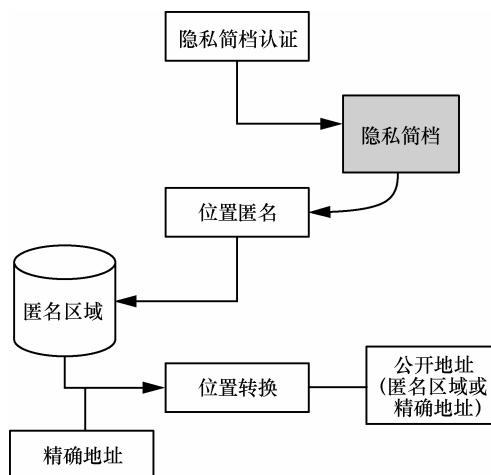


图 3 PROBE 架构

在考虑位置语义的同时，Che 等^[19]开始对隐私保护的架构进行调整，其在考虑语义位置能够影响用户位置分布因素的基础上，提出 *SALS* 位置分享架构，允许用户以 *P2P* 的方式与其他用户合作。

Li 等^[20]将现实的路网情况也考虑进方案，提出了一种新的隐私保护匿名架构保护路网环境中的敏感位置。在二层非集中架构下，将路网图转化为以路段交叉点和语义位置为中心点的 *Voronoi* 分割图，每个顶点的 *Voronoi* 分割即是该顶点的主域；其次，系统引入一个记录各主域内移动用户数量的数据中心（*DDC, dominance data center*），查询用户个性化地选择一个舒适区域，分割区域“有选择”地扩展以构建语义安全的匿名区域。根据信息熵有选择地扩展 *Voronoi* 分割区域，生成语义安全的匿名区域，该方法可有效解决路网环境下可能存在单路段攻击和语义推断攻击的问题。Pan 等^[21]利用在路网上的查询语义实现了个人隐私保护。

在对位置的语义进行更准确的阐释和隐私保护构架调整的基础上，后来的学者开始从不同层面不同维度提出各自的保护方案。文献[22,23]提出在数据库层面保护位置隐私的语义 *Bob-Tree* 的方法，其中 *Bob-Tree* 是一个在节点上含有语义信息的索引架构。文献[24]解释了用包含时空信息的语义信息推断用户行为的过程，提出了在语义

LBS 中, 利用决策树防止语义行为泄露的保护技术。文献[25]用本体论分类方法实现语义位置模糊技术, 比传统的基于几何学的位置模糊技术保护好。文献[26]提出了语义模糊的方法, 用一个信任第三方将位置坐标转化为高度匿名的位置特征。文献[27]引入了贝叶斯推理的方法, 用该方法对位置语义进行建模, 并证明了语义隐私度量的细节。

越来越多的研究者开始关注基于位置服务的语义隐私保护, 但是现有的研究方案仍然有各自的不足: 1) 语义分类大多比较粗糙, 不能精确区分位置语义的不同, 如虽然同为医院, 但传染病医院和口腔医院的敏感度不同; 2) 隐私保护的构架仍然没有一个公认模型, 现有构架依然存在资源消耗和隐私泄露的问题。

3.2 基于语义的轨迹隐私保护

轨迹是指用户在连续查询的状况下, 按时间排序位置序列。因此, 轨迹隐私保护可以认为是由单点隐私保护拓展而来, 是位置隐私的延伸与发展。

文献[9]对语义轨迹进行了形式化表述, 定义了其概念数据模型, 为后来的研究者提供了一种表示移动目标行为的方法。该模型将语义轨迹定义为停走序列, 即停靠点代表的是位置, 移动的为路径。

Monreale^[8]等进一步对位置进行分类, 生成泛化用户访问地址, 进而生成匿名轨迹数据集, 来保证攻击者推断出用户 ID 以及用户访问敏感位置的概率不会高于数据拥有者给定的阈值。同样是将敌手获得信息限制在某个阈值下, Lee 等^[28]提出通过对用户在位置的停留时间的观察, 来挖掘位置语义的方法。文章利用停留时间作为语义特征提取出位置可以提供什么类型的服务, 将其作为位置语义, 保护用户隐私。该方法的提出基于以下的认知, 即用户想要保护位置隐私是为了保护他们在这个位置做了什么, 如果一个轨迹上在某个地点超过了一个时间阈值, 就认为是在做有意义的事情, 该位置就定义为停靠点, 而停留时间长短则可以反映在该位置做的事情的种类, 进而判断位置的语义, 如判断该场所是餐厅还是工作场所。文章同时在位置隐私中定义了 θ -安全,

引入 EMD 距离计算语义相似性。具体实现方法是利用 EMD 距离计算两位置语义的概率分布, 生成位置语义图, 图中顶点为位置集群, 即一类语义相似的位置集合, 边的权重为 2 个不同聚类间的 EMD 距离。根据用户所在位置的集群标签进行扩展, 形成匿名区域 CA, 选出有最小 EMD 的区域, 如果其值小于 θ 则认为匿名区域符合 θ -安全。虽然文章很好地考虑了位置语义的选取, 但没有考虑路网环境。

文献[29]实现了对实时轨迹中敏感数据的保护, 在对城市网络位置匿名时融入语义轨迹的概念。文章首先对城市网络进行建模, 将城市根据语义敏感度分割成若干个区域, 由此形成以语义为基础的子图 G_i , 每一个城市图由若干个子图组成。用户请求服务时, 如果用户所在位置区域为敏感区域, 则需要匿名处理后发出请求。文献[30]则采用语义匿名和代理建立 SAP 树来取代传统的对中心匿名服务器的依赖, 满足位置服务要求并保护轨迹隐私。

然而上述方法都易受到几种位置隐私中常有的攻击, 即关系推断攻击、相似性攻击和推理攻击。为解决这几种攻击, 下面几种方法被提出。

为抵抗关系推断攻击, 即攻击者根据模糊轨迹推断出用户间的社会关系, 文献[31]提出一种基于语义树 (semantic tree) 的方法。作者根据匿名算法建模模糊区域, 并将此作为语义区域, 根据位置流行度和用户时空之间的相似度分配每个区域上的权值。该方法同时还支持多种层次位置的语义保护。文献[32]提出 SST (semantic space translation) 算法, 采用不同的策略以获得隐私保护与数据实用性的平衡。

文献[33]提出一种基于区域语义规则的数据保护方法来抵御相似性攻击。

Vincent 等^[34]为解决推理攻击, 提取出真实位置轨迹的地理位置和语义特征, 以合成可信的假轨迹。作者认为其合成的假轨迹与真实轨迹有很多相同且有用的静态特征, 可以用于地理数据分析。具体实现过程为: 首先随机抽样真实用户轨迹数据集, 成为种子数据集, 利用语义相似性量化 2 个移动模型的相似性, 构建位置语义图。将拥有高语义相似的分为一个组, 代表一个位置语

义类，不考虑位置的地理特性，只考虑不同的人以相同的方式访问某个地方将会分成相同的类，即用户访问概率、访问时间、以相似的方式从该地方转换到另一个地方的概率等因素相似，就认为其语义相似。为保护种子数据集用户的隐私，文章生成的假轨迹需要测试，测试其与种子痕迹的地理相似性，确保生成的假轨迹与真实轨迹语义相似地理非相似。文章很好地保护了种子用户的隐私，使攻击者不能确认种子用户轨迹或 ID。

针对其他类型的位置服务隐私泄露，文献[35]考虑到位置签到服务导致的用户私人信息泄露问题，设计了推测用户动机的机器学习模型。根据该模型预测出来的动机，来选择对用户分享的位置和话语的扰动程度。但该文使用的数据集较片面，只采集了美国的 Foursquare 用户数据。文献[36]考虑到完全依赖语义的基于位置的云应用的隐私保护问题，通过语义匿名进行有效的保护。文章通过融合位置匿名和失真的时态数据达到了更高层次的隐私保护，但没有将语义位置和物理位置相结合，即不能应用到实际场景中。文献[37]考虑到对路网的覆盖问题，提出了 CloSed 模型，设计了能够覆盖路网不同语义区域的匿名集。该模型能够满足服务质量的需求，很好地平衡服务质量和隐私需求。

3.3 小结

现有的基于语义的位置服务隐私保护工作正处于起步阶段，还未形成成熟的理论体系。在基于单点和基于轨迹的位置隐私中，都有许多尚未解决的问题，这些都是目前的研究热点。表 3 将位置隐私保护工作分为单点和轨迹 2 个类别，分别归纳了基于语义的位置隐私保护所采用的技术方法，并总结了现有工作的优缺点。虽然方法多种多样，但都未达到理想的保护位置隐私的效果。

4 语义相似度测量方法

为实现多层次的位置语义，体现位置间语义差异大小，采用位置语义相似度测量方法测量位置间的语义差异。语义相似度测量即测量位置间语义的不同。语义相似的具体实例见 2.3 节中的相似性定义。为实现语义聚类，首先要计算位置或轨迹间的聚类，目前有很多种计算相似的方法，如欧几里得距离、共同序列距离、KL 散度等，但这些距离应用到语义相似性上时有其局限性。例如，文献[28]提到当用 KL 散度计算以停留时间为特征的语义聚类时，其表现较差。语义相似度的测量方法有很多，本文只介绍 2 种比较有代表性的方法。

表 3 基于语义的位置服务隐私保护工作的优缺点

类别	方法	优点	缺点
单点	语义感知 ^[10,11]	能抵抗空间背景知识攻击	保护效果较差
	本体论 ^[12,14,25]	攻击者很难辨别真实位置	未考虑路网情况
	树算法 ^[13,22~24]	算法简单易理解	算法受内存大小限制
	Location diversity ^[15]	给出了语义位置概念，一定程度上保护现实生活的敏感位置	文章假设敏感位置是一个静态集合，可能会产生不符合现实情况的问题
	特征提取 ^[16~18]	在考虑位置流行度的基础上，用户可预先定义列表	未考虑路网情况
	SALS ^[19]	隐私度高	计算开销大
	Voronoi 图 ^[20]	解决路网环境和 P2P 环境的隐私保护问题，有效解决单路段攻击	未考虑连续 LBS 情况以及用户上下文信息
	语义位置模糊匿名 ^[21,26]	易实现	依靠可信第三方
	贝叶斯 ^[27]	实现贝叶斯方法建模位置语义	依靠可信第三方
	B.E.S ^[9]	定义了停走模型	特征较为简单，保护力度不够
轨迹	语义匿名 ^[8,32,33,35~37]	可抵御相似性攻击	保护强度受匿名集大小影响
	区域 EMD ^[28]	采用停留时间模型，考虑语义位置选取	未考虑路网信息
	图论 ^[29]	能够很好地保护敏感位置	保护程度和服务质量受敏感位置分布的影响
	树算法 ^[30,31]	能抵御推断攻击	计算开销大
	相似性聚类 ^[34]	不依赖真实用户轨迹，可有效保护种子用户的隐私	未考虑轨迹中的时间因素

4.1 基于 EMD 距离的测量方法

EMD 是在区域上 2 个概率分布距离的度量^[55], 数学上称为 Wasserstein 度量标准。EMD 距离最早用来解决运输问题^[56], 它反映一种概率分布转换到另一种概率分布时, 移动权值所需的最小工作量。EMD 在调节距离上的优势可以用来捕捉语义的不同^[28]。

文献[57]采用 EMD 的方法保护发布数据的隐私取得了良好效果, 基于此, 为更好地运用 EMD 距离, 文献[28]将 EMD 的地面距离 d_{ij} 设置为停留时间长短, 计算 EMD 距离作为位置对语义差异的大小, 2 个位置对之间的 EMD 距离越大, 表示 2 个位置对越不同。例如, 用户在位置 P、位置 Q、位置 R 停留的时间分别为 1 h、3 h 和 9 h, 计算两两位置对之间的 EMD 距离, 得到 $D_{\text{EMD}}(P,Q) < D_{\text{EMD}}(P,R)$, 即 P 与 Q 的相似度比 P 与 R 的相似度要高。文献[34]能够在计算 2 个移动模型不同的过程中将位置间的距离函数包含在内, 采用 EMD 距离方法测量的语义不相似性, 也取得了良好效果。

4.2 基于余弦相似度的测量方法

基于余弦相似度测量是通过测量 2 个向量内积的夹角余弦来度量它们的相似性, 可用在任何维度的向量比较中, 因此在相似度测量上使用广泛, 尤其是在文本相似度上有较好应用。

文献[58]计算余弦相似度时首先将位置中的语义特征映射到向量空间, 形成位置语义与向量数据的映射, 计算向量间的差异大小。语义特征可根据位置停留时间^[9]、用户访问目的^[34]、周围环境等信息作为输入的语义特征, 将这些语义特征以向量的形式呈现, 计算两者之间的相似度, 从而得到位置之间的语义相似。

余弦相似度的定义如下: 假设 2 个位置向量的语义值分别为 $\text{sem}(A)$ 、 $\text{sem}(B)$, 其相似性可表示为

$$\text{sim}(A,B) = \frac{\text{sem}(A)\text{sem}(B)}{\|\text{sem}(A)\| \|\text{sem}(B)\|}$$

余弦的取值范围为[0,1], 2 个位置语义相似度越高, 其值越接近于 1, 相似度越低, 越接近于 0。

基于 EMD 距离和余弦相似度测量语义相似度都是基于向量距离的语义测量的方法, 这 2 种

方法在测量语义相似度方面表现较为优异, 能够客观地反映位置的特点, 但是不能很好地解决数据稀疏问题。

5 基于语义的隐私保护度量方法

隐私度量是评价隐私方案保护力度的重要指标。隐私方案的隐私度越高, 其隐私保护程度越好, 用户被披露敏感信息的风险越小。现有基于语义的位置隐私保护度量模型都基于信息熵、K-匿名等理论, 根据选择的保护方法不同可选择不同的度量方法。

5.1 基于信息熵度量方法

信息熵一般用来衡量一个随机变量出现的期望值。一个变量的信息熵越大, 需要越多的信息才能确定变量。信息熵可以看作是随机变量的平均不确定度, 因此可用于度量隐私保护力度。一般情况下, 信息熵值越大, 隐私保护程度越高。

概率分布为 $p_1, p_2, \dots, p_n$ 的信息熵^[59]可定义为

$$H = -\sum_{i=1}^n p_i \log p_i$$

学者在信息熵的基础上对位置隐私度量做出一些改进, 如文献[20]提出语义位置流行度度量。语义位置流行度是基于信息熵的度量方法, 其与访问者数量以及访问频率有关^[60]。语义位置流行度在 2 个区域访问者数量相同的情况下, 用户越分散, 则流行度越高; 单个用户访问频率相同的情况下, 用户数量越多, 位置流行度越高, 隐私保护水平越高。文献[20]定义了构建以语义流行度为基础的匿名区域, 假设一个匿名区域中敏感语义流行度占总语义流行度的比值小于某一个事先定义的安全门限 θ , 则称这个区域为一个 θ -语义安全隐匿区域。 θ 越小, 其保护程度越高。

用信息熵来度量隐私保护程度的方法, 虽然简单有效, 但是该方法忽略了数据之间的相关性。

5.2 K-匿名

K-匿名是根据匿名集的大小度量隐私保护强度的方法。K-匿名因其简单明了等优势广受青睐。K-匿名的核心是将敏感项进行泛化处理, 使其隐匿于其余 K-1 条用户或位置信息中, 使攻击者确认用户真实位置的概率不会高于 $\frac{1}{K}$ 。一般情况

下, K 的值越大, 其隐私保护强度越高, 但服务质量也会受影响。文献[15]为抵御语义推断攻击, 利用位置语义多样性, 确保其生成的隐匿空间中至少包含 L 个不同的语义位置, 攻击者获得正确位置的语义概率为 $\frac{1}{L}$ 。对攻击者来说, 这种位置属性更加多样化的保护方案, 更难以推断。

K -匿名的度量方法应用面比较窄, 只能用于度量那些 K -匿名隐私机制的方法。同时, 有研究表明, K -匿名能对查询隐私有效度量, 而对位置隐私度量效果不好^[61]。

6 结束语

现有基于语义的位置隐私研究较少, 因此存在很多亟待解决的问题。

1) 语义相似特征多样化。现有基于语义的位置隐私保护方案未明确提出语义的概念, 没有一个规范能够明确语义的含义, 且现有方案大多数只提取语义的一个特征, 如停留时间等作为语义的标识, 对位置语义的分类来说不够精确。位置的语义特征可以从时间、用户行为模式、位置本身具有的特性等中提取。

2) 个性化隐私保护方案。在位置服务中, 每个用户要求的隐私保护程度不同, 如用户的工作地点是涉密单位, 则其对工作地点的隐私保护比在非涉密单位用户的要求要高; 不同年龄层、不同性别用户对相同地点的隐私保护程度要求也不尽相同。同时, 即使对于同一用户, 随着时间的变化, 其对隐私保护的要求也可能改变。因此, 针对用户需求的不同产生相应的保护策略与保护力度是未来研究的重要方向。

3) 位置语义标注自动化。现有方案对未知语义标注多是基于人工标注或粗略标注的形式, 这样的方式繁琐且容易出现问题。对于数据集中出现的新位置, 如何根据数据集中已有的数据对位置进行精准标记并分类, 实现位置语义自动标注也是未来研究位置隐私保护的重要基础。

4) 隐私保护与现实环境中的位置相结合。现实的位置信息比理论要复杂得多, 不仅要考虑路网信息, 还需考虑路况等。不同位置的地理环境不一样, 如有的地区以平原为主, 路况较为简单,

而有的地区以山区为主, 道路崎岖, 路况复杂, 引起的运动轨迹、行驶速度等也不尽相同。因此, 如何将基于语义的隐私保护应用到现实场景中, 是一个亟待解决的问题。

基于语义的位置隐私保护因结构较为复杂、涉及多方面知识等问题, 目前已知的国内外相关研究较少, 但其在隐私保护方面所起的作用不容忽视。本文概括了现有的基于语义的位置隐私保护近几年来研究成果与进展, 旨在为基于语义的位置隐私保护发展尽绵薄之力。

参考文献:

- [1] QuestMobile. 春季 APP 实力榜[R]. 北京: QuestMobile, 2016. QuestMobile. Spring APP power rankings[R]. Beijing: QuestMobile, 2016.
- [2] CNNIC. 中国互联网络发展状况统计报告[R]. 北京: 中国互联网络信息中心, 2016. CNNIC. Statistical report on internet development in China[R]. Beijing: China Internet Network Information Center, 2016.
- [3] HARA T, SUZUKI A, IWATA M, et al. Dummy-based user location anonymization under real-world constraints[J]. IEEE Access, 2016, 4: 673-687.
- [4] NIU B, GAO S, LI F H, et al. Protection of location privacy in continuous LBSs against adversaries with background information[C]//The International Conference on Computing, Networking and Communications (ICNC'16). 2016: 1-6.
- [5] NGO H, KIM J. Location privacy via differential private perturbation of cloaking area[C]//IEEE 28th Computer Security Foundations Symposium(CSF'15). 2015: 63-74.
- [6] XIAO Y, XIONG L. Protecting locations with differential privacy under temporal correlations[C]//The 22nd ACM Conference on Computer and Communications Security (CCS'15). 2015: 1298-1309.
- [7] CHATZIKOKOLAKIS K, PALAMIDESI C, STRONATI M. A predictive differentially-private mechanism for mobility traces[C]//The 14th Privacy Enhancing Technologies Symposium (PETS'14). 2014: 21-41.
- [8] MONREALE A, TRASARTI R, RENSO C, et al. Preserving privacy in semantic-rich trajectories of human mobility[C]//The 3rd ACM SIGSPATIA; International Workshop on Security and Privacy in GIS and LBS(SPRINGL'10). 2010: 47-54.
- [9] SPACCAPIETRA S, PARENT C, DAMIANI M L, et al. A conceptual view on trajectories[J]. Data & Knowledge Engineering, 2008, 65(1): 126-146.
- [10] DAMIANI M L, BERTINO E, SILVESTRI C. Protecting location privacy through semantics-aware obfuscation techniques[C]//Joint iTrust and PST Conferences on Privacy, Trust Management and Security (IFIPTM '08). 2008: 231-245.
- [11] DAMIANI M L, SILVESTRI C, BERTINO E. Semantics-aware obfuscation for location privacy[J]. Journal of Computing Science and Engineering, 2008, 2(2): 137-160.
- [12] LIOUDAKIS G V, DELLAS N L, KOUTSOLOUKAS E, et al. A semantic framework for privacy-aware access control[C]//The In-

- ternational Multiconference on Computer Science and Information Technology(IMCSIT'08). 2008: 813-820.
- [13] XIAO Z, XU J, MENG X. P-Sensitivity: a semantic privacy-protection model for location-based services[C]//The International Conference on Mobile Data Management Workshops(MDMW'08). 2008: 47-54.
- [14] HAAV H M, KALJUEE A, LUTS M, et al. Ontology-based retrieval of spatially related objects for location based services[C]//On the Move to Meaningful Internet Systems: (OTM'09). 2009: 1010-1024.
- [15] XUE M, KALNIS P, PUNG H.. Location Diversity: enhanced privacy protection in location based services[C]//The 4th International Symposium on Location and Context Awareness (LoCA'09), Tokyo. 2009: 70-87.
- [16] DAMIANI M L, BERTINO E, SILVESTRI C. The PROBE framework for the personalized cloaking of private locations[J]. Transactions on Data Privacy, 2010, 2(3): 123-148.
- [17] DAMIANI M L. Analyzing semantic location cloaking techniques in a probabilistic grid-based map[C]//The 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS' 10). 2010: 522-523.
- [18] DAMIANI M L, SILVESTRI C, BERTINO E. Fine-grained cloaking of sensitive positions in location-sharing applications[J]. IEEE Pervasive Computing, 2011, 10(4): 64-72.
- [19] CHE Y, CHIEW K, HONG X, et al. SALS: semantics-aware location sharing based on cloaking zone in mobile social networks[C]//The First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems(MobiGIS'12). 2012: 49-56.
- [20] LI M, QIN Z, WANG C. Sensitive semantics-aware personality cloaking on road network environment[J]. International Journal of Security and Its Applications, 2014, 8(1):133-146.
- [21] PAN X, WU L, PIAO C, et al. P³RN: personalized privacy protection using query semantics over road networks[C]//The International Conference on Web-Age Information Management (WIAM'14), Macau. 2014: 323-335.
- [22] THI-BAO T L, DANG T K. Semantic Bob-tree: a new obfuscation technique for location privacy protection[C]//The International Conference on Advances in Mobile Computing & Multimedia (MoMM'12). 2012: 281-284.
- [23] LE T T B, DANG T K. Semantic-aware obfuscation for location privacy at database level[C]//The International Conference on Information and Communication Technology(ICT'13). 2013: 111-120.
- [24] OH Y, JUNG K, PARK S, et al. A privacy preserving technique to prevent sensitive behavior exposure in semantic location-based service[C]//The 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems(KES'14). 2014, 35: 318-327.
- [25] ELKHODR M, SHAHRESTANI S, CHEUNG H. A semantic obfuscation technique for the internet of things[C]//The International Conference on Communications Workshops(ICC'14). 2014: 448-453
- [26] YANG J, ZHU Z, SEITER J, et al. Informative yet unrevealing: semantic obfuscation for location based services[C]//The 2nd Workshop on Privacy in Geographic Information Collection and Analysis (GeoPrivacy'15). 2015.
- [27] WU Z, CHEN Z, ZHU J, et al. Location semantics protection based on bayesian inference[C]//The International Conference on Web-Age Information Management (WIAM'15). 2015: 297-308.
- [28] LEE B, OH J, YU H, et al. Protecting location privacy using location semantics[C]//The 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '11). 2011: 1289-1297.
- [29] EMRE Y, MARIA L D, OSMAN A., et al. Privacy-preserving sharing of sensitive semantic locations under road-network constraints[C]//The 13th International Conference on Mobile Data Management(MDM'12). 2012: 186-195.
- [30] LI M, CAO S, QIN Z. A trajectory preserving method based on semantic anonymity proxy[C]//IEEE International Conference on Dependable, Autonomic and Secure Computing(DASC 2013). 2013: 16-20.
- [31] KOU Q, TIAN Y, SONG Z, et al. Privacy preserving social tie discovery based on cloaked human trajectories[C]//The 7th International Workshop on Hot Topics in Planet-scale mobile computing and online Social networking (HOTPOST'15). 2015: 13-18.
- [32] HAN P, TSAI H P. SST: privacy preserving for semantic trajectories[C]//IEEE International Conference on Mobile Data Management (MDM 15). 2015: 80-85.
- [33] MUBARK A A, ELABD E, ABDULKADER H. Semantic anonymization in publishing categorical sensitive attributes[C]//The International Conference on Knowledge and Smart Technology(KST 2016). 2016: 89-95.
- [34] VINCENT, BINDSCHAEDLER, REZA, et al. Synthesizing plausible privacy-preserving location traces[C]//The 37th IEEE Symposium on Security and Privacy(S&P'16). 2016: 546-563.
- [35] BILOGREVIC I, HUGUENIN K, MIHAILA S, et al. Predicting users' motivations behind location check-ins and utility implications of privacy protection mechanisms[C]//The 22nd Annual Network and Distributed System Security Symposium (NDSS'15). 2015.
- [36] BARAK O, COHEN G, TOCH E. Anonymizing mobility data using semantic cloaking[J]. Pervasive & Mobile Computing, 2016, 28(C): 102-112.
- [37] LI Y. Semantic-aware location privacy preservation on road networks[C]//Database Systems for Advanced Applications (DAS-FAA'16). 2016: 314-331
- [38] NGO H, KIM J. Location privacy via differential private perturbation of cloaking area[C]//Computer Security Foundations Symposium (CSF'15). 2015: 63-74
- [39] THAT D H T, POPA I S, ZEITOUNI K, et al. PAMPAS: privacy-aware mobile participatory sensing using secure probes[C]//The 28th International Conference on Scientific and Statistical Database Management (SSDBM'16). 2016: 1-12.
- [40] NOVAK E, LI Q. Near-private, proximity based location sharing[C]//The Conference on Computer Communications (INFOCOM'14). 2014: 37-45.
- [41] PALANISAMY B, LIU L. Effective mix-zone anonymization techniques for mobile travelers[J]. Geoinformatica, 2014, 18(1): 135-164.
- [42] ZUBERI R S, AHMAD S N. Secure mix-zones for privacy protection of road network location based services users[J]. Journal of Computer Networks & Communications, 2016, 2016(3): 1-8.
- [43] GERVAIS A, SHOKRI R, SINGLA A, et al. Quantifying Web-search privacy[C]//The 2014 ACM Sigsac Conference on Computer and Communications Security(CCS'14). 2014: 966-977.

- [44] SU X, HYYSALO J, RAUTIAINEN M, et al. Privacy as a service in digital health[J]. Computers and Society, 2016: 1-6.
- [45] SAFI H, ASSEMI B, MESBAH M, et al. Trip detection with smartphone-assisted collection of travel data[J]. Transportation Research Record Journal of the Transportation Research Board, 2016, 2594: 18-26.
- [46] CETATEANU A, LUCA B A, POPESCU A A, et al. A novel methodology for identifying environmental exposures using GPS data[J]. International Journal of Geographical Information Science, 2016: 1-17.
- [47] TRAJCEVSKI G. Compression of spatio-temporal data[C]//The 17th IEEE International Conference on Mobile Data Management[MDM'16]. 2016: 4-7
- [48] MOUSAVI S M, HARWOOD A, KARUNASEKERA S, et al. Geometry of interest (GOI): spatio-temporal destination extraction and partitioning in GPS trajectory data[J]. Journal of Ambient Intelligence & Humanized Computing, 2016: 1-16.
- [49] GRAAFF V, DE B R A, VAN K M. Automated semantic trajectory annotation with indoor point-of-interest visits in urban areas[C]//The 31st Annual ACM Symposium on Applied Computing(SAC'16), Pisa. 2016: 552-559
- [50] HUANG G, HE J, ZHOU W, et al. Discovery of stop regions for understanding repeat travel behaviors of moving objects[J]. Journal of Computer & System Sciences, 2015, 82(4): 582-593.
- [51] ILARRI S, STOJANOVIC D, RAY C. Semantic management of moving objects: a vision towards smart mobility[J]. Expert Systems with Applications An International Journal, 2015, 42(3): 1418-1435.
- [52] CHRISTINE P, STEFANO S, CHIARA R, et al. Semantic trajectories modeling and analysis[J]. ACM Computing Surveys, 2013, 45(4): 1-32.
- [53] SIQUEIRA F D L, BOGORNY V. Discovering chasing behavior in moving object trajectories[J]. Transactions in GIS, 2011, 15(5): 667-688.
- [54] PANAGIOTAKIS C, PELEKIS N, KOPANAKIS I, et al. Segmentation and sampling of moving object trajectories based representativeness[J]. IEEE Transactions on Knowledge & Data Engineering, 2011, 24(99): 1328-1343.
- [55] RUBNER Y, TOMASI C, GUIBAS L J. A metric for distributions with applications to image databases[C]//The International Conference on Computer Vision(ICCV'98). Bombay, 1998: 59-66.
- [56] RUBNER Y, TOMASI C, GUIBAS L J. The earth mover's distance as a metric for image retrieval[J]. International Journal of Computer Vision, 2000, 40(2): 99-121.
- [57] LI N, LI T, VENKATASUBRAMANIAN S. T-closeness: privacy beyond k -anonymity and l -diversity[C]//The International Conference on Data Engineering(ICDE'07). 2007: 106-115.
- [58] 侯洁. 基于位置语义的位置服务(LBS)隐私保护研究[D]. 成都: 电子科技大学, 2015.
- HOU J. Research on location semantics based privacy-preserving technology for location-based services[D]. Chengdu: School of Computer Science and Engineering, 2015.
- [59] SHANNON C E. A mathematical theory of communication[J]. The Bell System Technical Journal, 1948, 27(3): 379-423.
- [60] 李敏. 基于位置服务的隐私保护研究[D]. 成都: 电子科技大学, 2014.
- LI M. Research on privacy protection in location based services[D]. Chengdu: School of Computer Science and Engineering, 2015.
- [61] SHOKRI R, TRONCOSO C, DIAZ C, et al. Unraveling an old cloak: k -anonymity for location privacy[C]//The ACM Workshop on Privacy in the Electronic Society(WPES'10). 2010: 115-118.

作者简介:



马明杰 (1987-), 女, 山东莱芜人, 中国科学院信息工程研究所博士生, 主要研究方向为信息安全、隐私保护。



杜跃进 (1972-), 男, 宁夏银川人, 博士, 阿里巴巴集团安全部技术副总裁, 中国科学院信息工程研究所客座教授、博士生导师, 主要研究方向为版权与数据资产保护、安全态势感知技术。



李凤华 (1966-), 男, 湖北浠水人, 博士, 中国科学院信息工程研究所副总工、研究员、博士生导师, 主要研究方向为网络与系统安全、信息保护、隐私计算。



刘佳文 (1995-), 女, 陕西咸阳人, 中央财经大学本科生, 主要研究方向为网络安全、隐私保护。