

一种基于类四叉树的位置 K-匿名算法

金福生, 叶子石, 宋红

(北京理工大学 软件学院, 北京 100081)

摘要: 通过对 Casper 模型在匿名度和位置服务精准度方面优化, 提出了一种基于类四叉树的位置 K-匿名算法. 该算法采用自底向上的机制, 在扩大扫描区域时, 对请求用户所在网格的周边网格进行优先扫描, 然后再进行层级和空间分辨率上的提高. 实验结果表明, 该算法能够更好地提高位置服务的质量和匿名区域的准确度.

关键词: 类四叉树; 位置 K-匿名; Casper 优化; 自底向上

中图分类号: TP 391 **文献标志码:** A **文章编号:** 1001-0645(2014)01-0068-05

A Similar Quadtree Based on Location K-Anonymity Algorithm

JIN Fu-sheng, YE Zi-shi, SONG Hong

(School of Software, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Through optimizing the degree of anonymity and accuracy of location-based services of the Casper model, a similar quadtree based location K-anonymity algorithm is proposed. The algorithm uses a bottom-up mechanism, and in expanding the scanning area, the grids located around requesting user's grid are first scanned, then the level and spatial resolution are increased. Experiment results show the proposed algorithm can improve the quality of location-based services and the accuracy of the anonymous areas.

Key words: similar quadtree; location K-anonymity; optimized Casper; bottom-up

基于位置的服务(location-based service, LBS)从本质上来说是一种通过各种方式获取移动终端用户位置, 并与用户提出请求的位置有关的服务^[1]. 然而在用户享受 LBS 服务的同时, 其隐私也面临泄露的危险, 因此位置隐私保护应运而生^[2-3]. 享受服务与位置隐私保护是对矛盾体, 即享受基于位置的高效服务需要提供精确的位置, 而位置隐私保护需要隐藏用户位置^[4]. 目前已有的位置隐私保护模型通常假设用户具有统一的隐私保护需求, 而没有意识到对位置隐私保护来说, 不同的用户有不同的匿名需求, 因此不能确保 LBS 服务本身的精确性和有效性^[5]. 例如, 有些用户需要相对准确的 LBS 服务和很少的隐私保护, 而另一些使用 LBS 的用户则不希望被广告或第三方服务打扰因此偏爱高质量的隐私保护, 这时用户需要的是一种更加个性化的隐

私保护模型, 文献[6]首先注意到了这一点.

目前国内外几种主要的位置 K-匿名模型有: B Bamba 等^[7]根据位置信息二维化的特点, 提出的基于网格化的 BUGC(bottom-up grid cloaking)模型, TDGC(top-down grid cloaking)模型和 HGC(hybrid grid cloaking)模型等几种位置 K-匿名模型; MF Mokbel 等^[8]则根据空间网格预测思想, 提出了底层单元格分步扫描的 Casper 匿名模型; P. Zacharouli 等^[9-11]根据移动请求结点的历史运动情况分析, 提出了基于轨迹和连续查询的匿名模型; B. Gedik 等^[12]则根据不同用户的匿名请求, 提出了个性化的 Clique-Cloak 位置 K-匿名模型. 这些模型从上述主要矛盾出发, 完成了基本的位置 K-匿名工作. 其中基于轨迹和连续查询的匿名模型以及 Clique-Cloak 模型能够满足个性化的用户匿名需

收稿日期: 2013-09-22

基金项目: 国家自然科学基金资助项目(61003065)

作者简介: 金福生(1977—), 男, 博士, 讲师, E-mail: jfs21cn@bit.edu.cn.

求,但是在面对大数据量的匿名请求时处理速度过慢,容易影响用户体验;并且在一些简单需求的应用场景中,过于个性化的匿名需求往往显得多余和不必要。而基于网格化的位置 K-匿名模型则通常用于解决大量、统一的匿名请求,例如,一种通用的应用场景,其快速的处理时间和较高的匿名成功率能够满足一般的匿名要求。其中 Casper 模型是 M. F. Mokbel 等^[13]提出的一种基于网格化的位置 K-匿名模型,其通过一种对事先划定好的区域进行网格化的划分,并且将每个网格的信息以四叉树形式逐层上升储存,直到这一网格形式重新划归为一个区域。用户的匿名区域返回的是这些网格和子网格中的一个,最小不会小于这一区域地图的最小网格分辨率,最大则不会大于程序本身确定的匿名区域大小。Casper 模型的优势是能够快速划归匿名区域,以及其抗攻击性高,能够适应当今对于位置 K-匿名简单而又广泛的需求。然而 Casper 模型的缺点也很明显:其匿名区域过大而显得不够准确,并且使用户丢失了一些位置服务质量。

本文以 Casper 模型为基础,采用自底向上的方法,通过对当前网格层的邻域扫描,减少了匿名区域的范围,提高了 K-匿名的精确性和用户的位置服务质量。该算法在一定程度上解决了网格化算法对位置服务质量的高损耗问题,同时保留了其匿名成功率高、处理速度快的特点,弥补了现有技术的不足。

1 基于类四叉树的位置 K-匿名算法

1.1 用户信息表结构

本文提出的类四叉树算法,以 Casper 算法为基础,并对其现有的信息表参数进行了优化。与 Casper 模型相比,类四叉树算法引入了记录用户请求次数的 RNO 参数、请求时间的 Time 参数、记录用户需求延时的 dTime 参数以及标志参数 Active。

用户信息表中每条用户记录由 (UID, RNO, CID, Time, Location, dTime, Profile) 结构元组构成。其中,UID 参数代表移动用户的身份标识,此参数具有唯一性,任意两个移动用户的 UID 在同一时刻不能相同;RNO 代表同一移动用户发出的 LBS 请求的编号,RNO 参数通过记录用户在现有信息信息表中出现的次数来衡量一个用户的请求频率,请求数越多,RNO 值越大;CID 代表用户发出请求时所在的网格区域中最底层的单元格编号,匿名服务器通过用户信息表和网格信息进行关联;Time 参数

表示用户发出请求的时间;dTime 参数表示请求延迟期;Location 参数中包含 X,Y 两个坐标值,分别对应移动用户发出请求时网格图上的空间位置;Profile 参数表示移动用户的隐私需求属性,包含两个子值 K 和 A,具体以 (K,A) 的形式表达,其中 K 值代表匿名区域最小匿名度,表示如果最终匿名区域中最少包含 K 个用户,则移动用户被识别出的概率为 $1/K$,根据第三方应用的需求不同,用户的 K 值由算法统一确定,A 值表示匿名需求的最小面积;此外,还包含一个布尔型变量 Active 用来判断元素状态。

Time 和 dTime 参数用来判断某个用户的活跃期,当一个移动用户请求查询时,其申请时间 Time 会被记录,如果在 dTime 延迟期内此用户没有进行下一次请求,则用户信息将被视作未激活元素,Active 置为 F,直到其被二次激活,同时重置 RNO;如果在延迟期内有下一次请求,则按顺序继续记录。在位置 K-匿名中,如果匿名集的 K 个元素全部是较为活跃的元素,将增加 K-匿名成功率、减小被识别出的概率;反之则容易泄露隐私信息。

1.2 Casper 算法与类四叉树算法

传统的 Casper 算法在构建金字塔形网格状图形区域后,对于发出请求的移动用户自顶向下扫描网格图,也就是 Top-Down 式遍历四叉树的一支,直到找到不符合用户 K 值和需求的当前层网格非叶子结点或叶子结点,则返回此网格的上一层网格坐标作为匿名集区域,完成对请求用户的 K-匿名,结果通常是标准的矩形匿名框。此方法的优点在于匿名速度非常快;而缺点也是很明显的,匿名结果会是超大的匿名框,并且匿名成功率低。

本文提出的类四叉树算法则采用自底向上 (bottom-up) 的扫描方式,虽然这样做的结果牺牲了一部分匿名速度,但是却极大地提高了匿名框的精度以及匿名成功率。算法主要步骤为:

① 从事先划分好存入数据库的底层单元格信息开始扫描,根据其匿名需求确定匿名集;

② 选择行或列进行迭代,遍历当前结点的兄弟结点,判断条件成立则返回此最小匿名框 (不一定是标准矩形),否则继续;

③ 直到在距底层 3 层以内的网状平面 (分辨率再低则没有意义) 找到成立的网格区域,或者返回匿名失败信息。

此方法通过先遍历同层的 3 个兄弟结点,再向

上遍历父亲结点的方法,缩小了匿名框的面积,提高了匿名的空间分辨率和 LBS 服务精准度。

2 仿真实验

2.1 实验环境

为了验证模型的有效性,方便将类四叉树算法与其他模型在相同环境下进行比较,本文选用了位置 K-匿名算法测试中知名的 NG 网络结点生成器^[14-15] (the network-based generator of moving objects) 作为实验平台。NG 网络结点生成器模拟的网络特征遵循了现实生活中的道路交通规律,其中,时

间因素和道路交通状况会影响移动用户的速度和道路选择,因此可以为本文的实验提供高质量的仿真环境。

实验主机操作系统为 Windows7 操作系统,内存空间为 3 GB,双核处理器,每个处理器的主频为 2.1 GHz。

模拟软件选择了奥登堡市地图作为实验用图,地图的尺寸为 $23\,572 \times 26\,915$ 。

2.2 实验数据

软件主要模拟指定区域上移动网络中车流的运动情况,实验环境参数的具体设定如表 1 所示。

表 1 实验环境参数

Tab. 1 Experimental environment parameters

参数	初始移动 结点数目	每个时间戳 增加结点数目	初始外部 对象数目	每个时间戳增 加对象数目	最大运行 时间/s	移动结点的 类别数目	外部对象的 类别数目	移动物体最大速 度/($\text{km} \cdot \text{h}^{-1}$)	移动结点汇 报概率/ms
多车、快速	200	30	0	0	20	6	3	90	1 000
中量、中速	100	20	0	0	20	6	3	50	1 000
少车、慢速	50	10	0	0	20	6	3	40	1 000

2.3 仿真测试

测试使用 NG 网络结点生成器在奥登堡 (Oldenburg) 环境下模拟生成的移动信息数据,将全部地图分为 256×256 个单元格,每个单元格大小为 94×106 ,并在服务器端取不同 K 值 (5, 10, 20) 分别进行测试。

本算法以 Casper 模型为基础进行改进,旨在满足用户匿名需求和效率的前提下,提高匿名精确性,减小用户的位置服务质量损失,因此实验从与此相关的相对匿名度和空间分辨率两方面着手,与目前流行的 BUGC 算法、HGC 算法、Casper 算法进行对比。

2.3.1 相对匿名度测试

相对匿名度是指任意一个服务请求所包含的位置信息被匿名处理后,所在匿名集合的匿名元素总量与移动用户所要求的匿名 K 值的比值,在大于 1 的基础上,相对匿名度越高匿名效果越好,同时系统花费的代价也更高。本测试中的相对匿名度是通过在每个有效结点进行匿名区域的元素统计并与其 K 值相除,最后计算平均值得出。测试结果如图 1 和图 2 所示。

由图 1~2 易知,在 K 取不同值时,与其它 3 个算法相比,类四叉树算法的相对匿名度较稳定;而在固定 K 值、增加最大空间公差的情况下,类四叉树算法也能够保持稳定的相对匿名度,不会因为条件

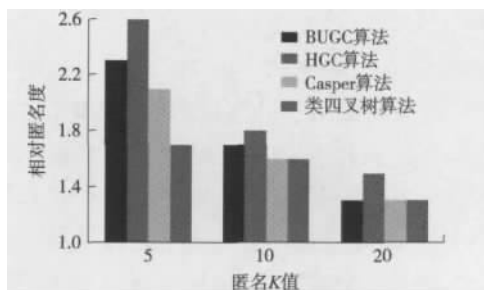


图 1 400 m × 400 m 空间公差情况下不同 K 值的相对匿名度

Fig. 1 Relative anonymity degree in 400 m × 400 m space tolerance and different K value

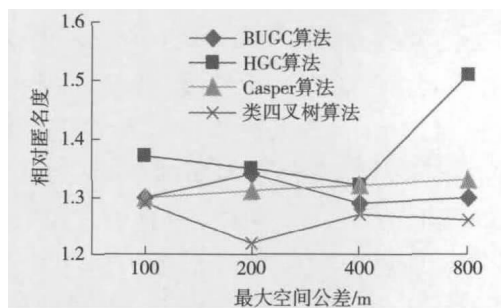


图 2 $K=10$ 情况下不同空间公差的相对匿名度

Fig. 2 Relative anonymity degree in different space tolerance when K equals ten

的变化而出现大的波动,从而影响 LBS 服务和匿名保护之间的平衡。由于基于网状图的 K -匿名保护算法从已经划分好的空间区域选取匿名区域,因此这些算法整体的相对匿名度较高、抗攻击性强。类四叉树算法以 Casper 模型为基础,采用自底向上的

扫描机制,先遍历同层的3个兄弟结点再向上遍历父亲结点的方法,缩小了匿名框的面积,在继承基于网状图的K-匿名保护算法优点的前提下,提高了匿名的空间分辨率和LBS服务精度。

2.3.2 空间分辨率测试

空间分辨率是指满足移动用户关于匿名需求的最小空间面积与匿名算法所获得的最终匿名区域的比值。在满足移动用户位置隐私保护要求的前提下匿名框的区域范围越小,这意味着该匿名框更接近理论上的最优解。即空间分辨率越大越好。此属性在一定程度上与相对匿名度是矛盾的。测试中设置的LBS服务精度即用户的需求位置查询精度为 100×100 。当请求结点能够成功匿名时,算出其空间分辨率;而整体的空间分辨率为所有可以成功匿名结点的平均值,因此当空间公差为 100×100 时,空间分辨率全部为1。测试结果如图3和图4所示。

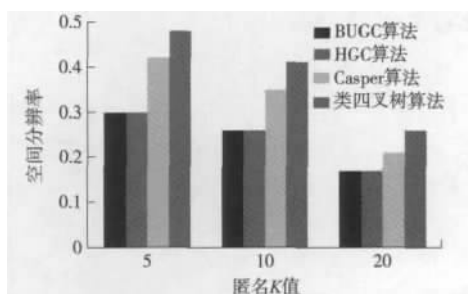


图3 400 m×400 m空间公差情况下不同K值的空间分辨率

Fig. 3 Spatial resolution in 400 m×400 m space tolerance and different K value

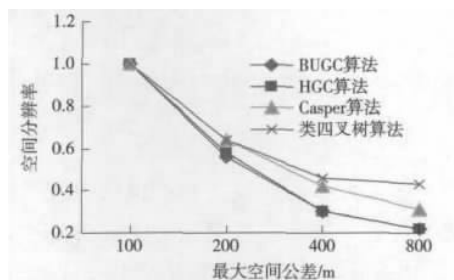


图4 K=10情况下不同空间公差的空间分辨率

Fig. 4 Spatial resolution in different space tolerance when K equals ten

由图3~图4可见,传统的基于网格的K-匿名算法其空间分辨率随着最大公差的增大急剧下降,是因为可成功匿名的结点随着公差增加而变多。而类四叉树算法则在Casper算法的基础上增加了最小匿名框选择的机制,并且随着公差增大,相邻单元格的结点元素增多,因此组建匿名框的可能更大,所以空间分辨率随着公差增大而趋于稳定,并

且整体高于传统算法。

3 结 论

在分析常见位置匿名算法的基础上,以典型的网格位置K-匿名模型Casper为基础,提出了类四叉树的K-匿名算法,分别从相对匿名度和空间分辨率两个方面评价算法的精确度性能和位置服务质量。实验结果表明,本文提出的类四叉树算法与现有的基于网格的位置K-匿名算法相比,提高了相对匿名度和空间分辨率,减小了由匿名保护带来的信息损失,可更好地平衡基于位置服务和个人位置隐私保护之间的关系。

参考文献:

- [1] Mokbel M F. Privacy in location-based services: state-of-the-art and research directions[C] // Proceedings of 2007 International Conference on Mobile Data Management. [S.l.]: IEEE, 2007:228-228.
- [2] Krumm J. A survey of computational location privacy[J]. Personal and Ubiquitous Computing, 2009,13(6): 391-399.
- [3] 潘晓,肖珍,孟小峰.位置隐私研究综述[J].计算机科学与探索,2007,1(3):268-281.
Pan Xiao, Xiao Zhen, Meng Xiaofeng. Survey of location privacy[J]. Computer Science and Exploration, 2007,1(3):268-281. (in Chinese)
- [4] Liu L. From data privacy to location privacy: models and algorithms[C] // Proceedings of the 33rd International Conference on Very large Data Bases. [S.l.]: VLDB Endowment, 2007:1429-1430.
- [5] Chow C Y, Mokbel M F, He T. A privacy-preserving location monitoring system for wireless sensor networks[J]. IEEE Transactions on, Mobile Computing, 2011, 10(1):94-107.
- [6] Wang T, Liu L. Privacy-aware mobile services over road networks[J]. Proceedings of the VLDB Endowment, 2009,2(1):1042-1053.
- [7] Bamba B, Liu L, Pesti P, et al. Supporting anonymous location queries in mobile environments with privacygrid[C] // Proceedings of the 17th International Conference on World Wide Web. [S.l.]: ACM, 2008:237-246.
- [8] Chow C Y, Mokbel M F, Aref W G. Casper: query processing for location services without compromising privacy[J]. ACM Transactions on Database Systems (TODS), 2009,34(4):24.

(下转第76页)

同CT图像进行分割,结果显示本文算法在系统鲁棒性、分割精度、系统运行时间比先前方法都有很大的改进.该算法可适用于多组数据,分割效果良好,可以为后续肝脏肿瘤体积的测定提供必要的技术支持,具有一定的临床应用价值.

参考文献:

- [1] Jemal A, Bray F, Center M M, et al. Global cancer statistics[J]. *A Cancer Journal for Clinicians*, 2011, 61(2):69-90.
- [2] 周康荣,杨军,严福华.小和微小肝癌的影像学诊断和进展[J].*放射学实践*,1999,14(4):273-274.
Zhou Kangrong, Yang Jun, Yang Fuhua. Imaging diagnosis and progress of small and micro liver cancer[J]. *Radiology Practice*, 1999, 14(4):273-274. (in Chinese)
- [3] 翟伟明,胡成文,张伟宏,等.基于动态自适应体素生长的肺部CT图像3维分割算法[J].*中国图象图形学报*, 2005,10(10):1269-1274.
Zhai Weiming, Hu Chengwen, Zhang Weihong, et al. A dynamic adaptive 3D voxel-growing segmentation algorithm for pulmonary CT images[J]. *China Journal of Image and Graphics*, 2005, 10(10):1269-1274. (in Chinese)
- [4] Seo K S. Automatic hepatic tumor segmentation using

composite hypotheses[J]. *Image Analysis and Recognition*, 2005, 3656:922-929.

- [5] Park Seung-Jin, Seo Kyung-Sik, Park Jong-An. Automatic hepatic tumor segmentation using statistical optimal threshold[J]. *Lecture Notes in Computer Science*, 2005, 3514:275-283.
- [6] Nugroho H A, Ihtatho D, Nugroho H. Contrast enhancement for liver tumor identification[C]// *Proceedings of the MIDAS Journal-Grand Challenge Liver Tumor Segmentation (MICCAI Workshop)*. New York: Springer, 2008.
- [7] Wong Damon, Liu Jiang, Yin Fengshou, et al. A semi-automated method for liver tumor segmentation based on 2D region growing with knowledge-based constraints [J]. *MICCAI Workshop*, 2008, 41:43-159.
- [8] Abate A F, Michele N, Cenny T, et al. IME an image management environment with content-based access [J]. *Image and Vision Computing*, 1999, 17:976-980.
- [9] Huang J. Color-spatial image indexing and applications [D]. [S.l.]: Cornell University, 1998.
- [10] Rangayyan R M. Biomedical signal analysis[M]. New York: Wiley, 2002.

(责任编辑:刘芳)

(上接第71页)

- [9] Chow C Y, Mokbel M F. Enabling private continuous queries for revealed user locations[M]. Berlin, Heidelberg: Springer, 2007:258-275.
- [10] Gkoulalas-Divanis A, Verykios V S. A free terrain model for trajectory K-anonymity[C]// *Proceedings of Database and Expert Systems Applications*. Berlin, Heidelberg: Springer, 2008:49-56.
- [11] Gkoulalas-Divanis A, Verykios V S, Mokbel M F. Identifying unsafe routes for network-based trajectory privacy[C]// *Proceedings of SDM*. Sparks, Nevada, USA: SIAM, 2009:942-953.
- [12] Gedik B, Liu L. Protecting location privacy with personalized k-anonymity: Architecture and algorithms

[J]. *IEEE Transactions on Mobile Computing*, 2008, 7(1):1-18.

- [13] Mokbel M F, Chow C Y, Aref W G. The new Casper: query processing for location services without compromising privacy[C]// *Proceedings of the 32nd International Conference on Very Large Data Bases*. Seoul, Korea: ACM, 2006:763-774.
- [14] Brinkhoff T. Generating traffic data[J]. *IEEE Data Eng Bull*, 2003, 26(2):19-25.
- [15] Brinkhoff T. A framework for generating network-based moving objects[J]. *GeoInformatica*, 2002, 6(2): 153-180.

(责任编辑:刘芳)