基于位置敏感哈希分割的空间 K-匿名共匿算法

侯士江[1,2] 张玉江[2] 刘国华[3]

(燕山大学信息科学与工程学院 秦皇岛 066004)[1] (燕山大学工业设计系 秦皇岛 066004)[2]

(东华大学计算机科学与技术学院 上海 200051)[3]

摘 要 空间 K-匿名技术主要用于隐私保护,防止个人信息泄露。目前的主要方法都基于用户-匿名器-基于位置的服务(location based services,LBS)模型。提出了一种基于位置敏感哈希分割的空间 K-匿名共匿算法。这种算法在保距性和共匿性方面都可以满足要求,而且算法具有适度的计算复杂度。最后,针对有效性(最小化匿名空间区域)和效率(构建代价)做了实验,证明所提出的算法具有良好的性能。

关键词 基于位置的服务,空间 K-匿名,隐私保护,空间数据库

中图法分类号 TP309.2 **文献标识码** A

Spatial K-Anonymity Reciprocal Algorithm Based on Locality-sensitive Hashing Partition

HOU Shi-jiang[1,2] ZHANG Yu-jiang[2] LIU Guo-hua[3]

(Department of Information Science&Engineering, Yanshan University, Qinhuangdao 066004, China)[1]

(Department of Industrial Design, Yanshan University, Qinhuangdao 066004, China)[2]

(Department of Computer&Technology, Donghua University, Shanghai 200051, China)[3]

Abstract Spatial K-anonymity is an important measure for privacy to prevent the disclosure of personal data. The main methods are based on the model of User-Anonymizer-LBS. This paper proposed a spatial K-anonymity algorithm based on locality-sensitive hashing partition. The algorithm is shown to preserve both locality and reciprocity with moderate computation complexity. Finally, aimed on effectiveness(minimum anonymizing spatial region size) and efficiency(construction cost), the experimental results verify that the proposed method has high performance.

Keywords Location-based services, Spatial K-anonymity, Privacy protection, Spatial databases

1 引言

随着便携式定位装置的发展,移动设备的应用领域如实时交通监控、找出最佳骑行路线、3G 宽带服务等得到了广泛的关注。在参与式传感中,用户既可以提供宝贵的信息(数据报告)也可以检索(依赖所在位置)信息(查询)。隐私保护是数据共享中一个很重要的问题。在参与式传感中,隐私担忧来自两个方面。第一方面是在数据发布过程中,通常不能披露详细的个人信息。已经提出了许多技术如通过数据转换[1,2]来解决这一问题。第二方面源于查询过程,如用户提交的基于位置的查询(例如,"最近的酒吧?")。位置隐私主要集中在两点:隐藏用户位置;隐藏用户身份,这可以避免用户主体与行为之间的关联(例如,"谁正在查询最近的酒吧?")。我们的工作针对第二方面的问题。

K-匿名作为隐私保护措施首先由 Sweeney 等[3]提出,用来防止个人数据的泄露。如果一个表满足 K-匿名,那么这个表中的每条记录都是不可分辨的,每条记录至少有(K−1)条其它记录与之具有同样的准标识符属性。在位置隐私语境中,可以将位置属性视为准标识符。位置隐私的 K-匿名意味着用户的位置无法从至少(K−1)其他用户中区分出来。实现位置隐私的 K-匿名之通常的做法是引入一个可信服务器,称之为匿名器(anonymizer,AZ),匿名器负责删除用户的 ID 和构建一个匿名空间区域(anonymizing spatial region,ASR),这个区域包含用户本身和至少(K−1)个邻近用户。用 ASR 取代查询用户的真实坐标位置,这就是所谓的"隐匿"。空间 K-匿名(spatial K-anonymity,SKA)要求即使在最坏的情况下(即攻击者掌握所有用户的位置),其能够识别出查询用户 U 的概率也不超过 1/K。

最先提出的空间 K-匿名算法有 Casper Cloak[4]、Interval Cloak[5]。Casper 和 Interval Cloak 算法仅对均匀数据是安全的。现有证明安全的技术有 Hilbert Cloak[6],其已经应用在 P2P 系统(Peer-to-Peer system)上。Hilbert Cloak 使用希尔伯特空间填充曲线(Hilbert space filling curve)[7]将 2-D 空间映射为 1-D 值。同时,用户位置隐私保护在其他相关问题中也得到研究。Clique Cloak[8]合并了时空隐匿,概率性隐匿(Probabilistic Cloaking)[9]不适用于 SKA 的概念。Khoshgozaran 和 Shahabi[10]开发了一种 1-D 转换和加密技术,其隐匿了空间数据和 LBS 查询。Hoh 和 Gruteser[11]通过连续收集位

到稿日期:2012-10-20 返修日期:2013-02-21 本文受国家自然科学基金(61070032)资助。

侯士江(1978−),男,博士生,讲师,主要研究方向为信息安全、计算机辅助设计,E-mail:shjhou@ysu.edu.cn;刘国华(1966−),男,教授,博士生导师,主要研究方向为 BMP、数据库理论、信息安全。

置样本来隐藏用户的轨迹。文献[12-14]关注于通过加密技术获得查询隐私保护而不依赖 SKA。然而，这种加密协议即使在并行架构下也会引起很高的通信和计算代价。文献[15,16]研究了路网环境下的匿名查询问题。

2 系统模型

与文献[5]类似，我们假设攻击者可以：(1)拦截 ASR，(2)知道匿名器所使用的隐匿算法，(3)获得了所有用户的最新位置。第(1)个假设意味着 LBS 不可信或者匿名器与 LBS 之间的通信通道不安全。第(2)个假设是由于数据安全技术通常是公开的。第(3)个假设是出于这样的事实，即用户经常从同一地点(家里、办公室)进行基于位置的查询，而这些用户位置数据可以通过物理观察、三角测量、电话目录等方法确定。

为了解决这些问题，现有的空间 K-匿名系统大多采用图 1 所示模型。(1)用户 U 向匿名器发送查询请求和所需的匿名度 K；(2)匿名器移除用户 U 的 ID 并建立一个匿名集(anonymizing set，AS)，AS 包含了 U 以及至少($K-1$)个邻近区域的其他用户，包含了 AS 的封闭空间区域即为 ASR；(3)匿名器向 LBS 提交 ASR 并储存空间数据(即查询请求)；(4)LBS 处理查询并向匿名器返回候选结果集；(5)匿名器移除错误的查询结果并向用户 U 返回真实结果。

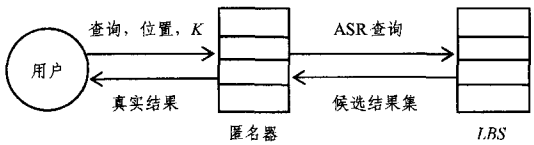
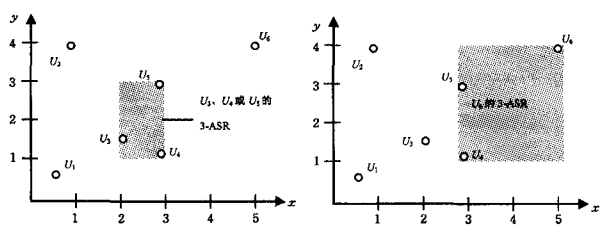


图 1 空间匿名系统模型

3 SKA 的共匿性(reciprocity)

简单生成包括 K 个用户的 ASR 并不能充分满足 SKA 的要求。以中心隐匿(Center Cloak)算法为例，假设用户提交了一个查询，算法搜索距离其最近的($K-1$)个其他用户，然后用最小外包矩形(minimum bounding rectangle，MBR)封装这 K 个用户形成 ASR。如果查询来自 U_3、U_4 或 U_5，中心隐匿算法将产生图 2(a)中所示的 ASR。然而，如果查询源于 U_6，ASR 将是图 2(b)中的阴影矩形。显然后者违反了 SKA，因为攻击者能够确定查询是由 U_6 发布，因为其他用户不可能产生同样的 ASR。具体来说，由 U_4 或 U_5 产生的 $K=3$ 的 AS 都会包含 U_3(如图 2(a)所示)，而不会包含 U_6。大多数现有的隐匿算法都有同样的问题。



(a) U_3、U_4 或 U_5 的 3-ASR　　(b)U_6 的 3-ASR

图 2 空间隐匿示例

为了解决这类问题，Kalnis 等[6]引入了共匿性(reciprocity)的概念，这是 SKA 的充分条件。共匿性要求对于给定的 K 值，其用户子集一直群组在一起，相当于 AS 中的任一用户都存在于 AS 中的其他用户的 ASRs 中。其定义如下：

定义 1(共匿性，Reciprocity)　假定用户 U 提交一个查询，匿名度为 K，匿名集为 AS，匿名空间区域为 ASR，如果 AS 满足：(1)其包含 U 及至少($K-1$)个其他用户；(2)AS 中的每个用户对于特定的 K 值产生同样的匿名集 AS，那么就认为 AS 满足共匿性。

如果 AS 满足共匿性，那么 AS 中的任一用户产生 ASR 的概率都是 $1/|AS|$，$|AS|$ 是 AS 的基数。由于 $|AS|\geq K$，因此准确识别查询用户的概率不会超过 $1/K$。如图 2(a)所示，$AS=\{U_3,U_4,U_5\}$ 满足共匿性，因为无论是 U_3 还是 U_4 或 U_5 提交的 $K=3$ 的查询都是 $AS=\{U_3,U_4,U_5\}$。因此攻击者不能识别出具体的用户。相反，如图 2(b)中所示，$AS=\{U_4,U_5,U_6\}$ 不能满足共匿性，因为 U_4 和 U_5 的 AS 是图 2(a)中的 $\{U_3,U_4,U_5\}$(即 U_6 不在 U_4 和 U_5 的 3-ASR 中)。如果所有的 AS(即对于每个用户和 K)都满足共匿性，那么隐匿算法是共匿的，能够证明共匿算法是安全的。

除了安全之外，空间隐匿方法还应该高效且有效。高效是指产生 ASR(匿名器端)的代价应该最小化，以满足更好的可扩展性和更快的服务。有效性指的是 ASR 的面积应该最小化。具体来说，大的 ASR 会引起高的处理负担(LBS 端)和网络代价(LBS 与匿名器之间要传输更多的数据)。共匿性通常会导致较大的 ASRs，所以它会对效率产生负面的影响。

4 位置敏感哈希(LSH)

满足共匿性要求的简单匿名方法是将邻近的用户分割成桶。例如，给定数据集 S，我们随机选择一点 q 和它的($K-1$)个最近邻居一起组成桶。这个过程递归进行，直到所有的点都被分配到桶。这种方法有两个缺点：首先，它产生了数据碎片并且不具有保距性。分在同一个桶中的用户在原始的数据集中可能并没有邻近关系，尤其是对于 K 值较大的情况。如图 3(a)所示，匿名度 $K=4$ 且首先选择点 q。q 和它的 3 个邻居形成匿名区域，如图中矩形所示。将这 4 个点移出数据集后，剩余点的匿名区域很大。图 3(b)所示的分割方式显然更为可取。其次，时间复杂度高，很容易发现该方法需要 $O(\frac{n^2}{K}\log(\frac{n^2}{K}))$ 的运行时间。接下来我们提出一种基于位置敏感哈希(locality-sensitive hashing，LSH)的方法，它用较低的时间复杂度就可实现理想的分割[17]。

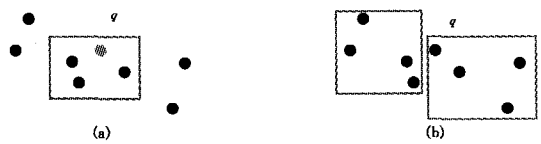


图 3 简单 NNs 分割产生碎片问题

LSH 与一般哈希函数的区别在于其位置敏感性，也就是散列前的相似点经过哈希之后，也能够在一定程度上相似，并且具有一定的概率保证。相似性检索一般通过 K 近邻或近似近邻查询来实现。

定义 2(位置敏感)　设集合 S，$\forall p,q\in S$，距离记为 $D(p,q)$，p 的 r 邻域记为 $O(p,r)$。从 S 映射到 W 的函数族 $\mathcal{H}=\{h:S\to W\}$ 称为对距离 D 是 (r_1,r_2,p_1,p_2)-位置敏感的，如果满足以下两个条件：

(1)if $p\in O(q,r_1)$，then $\Pr_{\mathcal{H}}(h(q)=h(p))\geq p_1$；

(2)if $p\notin O(q,r_2)$, then $\Pr_*(h(q)=h(p))\leqslant p_2$。

其中，$p_1>p_2$，$r_1<r_2$。

本文采用文献[17]中所提出的基于 p-稳态分布的 LSH 族。策略如下，设输入向量为 v_1 和 v_2，独立于 p-稳态分布 X 的向量 a 被选择。在 a 上的投影距离为$(a\cdot v_1-a\cdot v_2)$，分布为 $\|v_1-v_2\|_p X$。将投影点分到尺寸为 K 的等宽的桶中，当 $h_a(q)=q\cdot a$ 时，向量 a 为 LSH 函数。相似点被分到同一个桶中的概率很高，反之却不能成立，即同一个桶中的点有可能相距很远。可以使用多个哈希函数来解决这种问题，即 q 被哈希函数族 L 取代，$g_l(q)=\langle h_1(q)\rangle, l=1,2,\cdots,L$。如图 4 所示，多个哈希函数能够使数据点更好地分离。点 p_1，p_2，p_3 和 p_4 投影到 a 上很接近，而投影到 b 上却更为分散。使用向量 a，b 的目的是将 p_1，p_2 映射到同一个桶，p_3 和 p_4 映射到另一个桶，如图 4 所示，映射时会有一个错误率 ε。

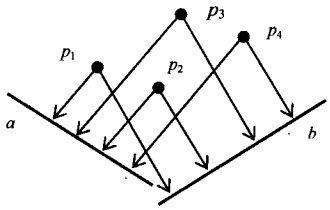


图 4　用两个哈希函数 a，b 哈希平面中的 4 个点

5　基于位置敏感哈希分割的匿名算法

在空间隐匿问题中使用 LSH 主要归功于它的保距特性，将数据集分割为至少包含 K 个元素的组。使用单个哈希函数，同一个桶中包含远离点的概率比较高，所以我们用哈希函数族 L 来代替。基于位置敏感哈希分割的匿名算法(LSH-based partition cloaking，LSH_BPC)见算法 1。算法首先建立将 S 经过哈希函数族 L 散列形成的序列表 $l_1,\cdots,l_L$，接下来每个序列表被分割成大小为 K 个元素的桶(最后一个桶可能会多于 K 个元素)。为了避免产生碎片，每次都从第一个序列表 l_1 中的第一个元素 q 开始，序列表中所有包含 q 的桶中的元素形成集合 Ω，然后再从 Ω 中找出 q 的$(K-1)$个最近邻居。由于 LSH 的特性，q 的最近 NNs 在 Ω 中的可能性很大，而且 Ω 的尺寸也不会很大。

算法 1　LSH_BPC(U,K, S)

输入：查询点 U，匿名度 K，数据集 S

输出：MBR(T)// U 的 ASR

1. if |S|<2K
2. 　Return MBR(S)
3. 生成哈希函数族 L，向量从 a 高斯分布中选择
4. 依据 S 的哈希值计算并维护 L 排序表{$l_1,\cdots,l_L$}
5. T←∅
6. while 排序表中的元素个数≥2K and U∉T do
7. 　T←∅
8. 　将 l_i 分割为容量为 K 的桶，i=1,…,L
9. 　q←l_1 中的第一个元素
10. 　Ω←∅
11. 　for i=1 to L do
12. 　　b←l_i 中包含 q 的桶中的元素
13. 　　Ω←Ω∪b
14. 　end
15. 　T←q∪(Ω 中 q 的(K−1)个最近邻居)
16. 　从 L 排序表中移除 T 中的元素
17. end
18. if U∈T Return MBR(T)
19. else Return MBR(L)

算法 LSH_BPC 第 3－4 行生成排序表 L，时间代价为 O($Ln\log n$)，其中 $n=|S|$。第 15 行要求将 Ω 中的元素进行排序，代价为 O($LK\log LK$)，实际上远低于此。第 6－17 行的循环至多执行$\lfloor n/K\rfloor$次。因此最坏情况下算法的时间复杂度为 O($Ln\log n$)。

以图 4 示例进行分析，假设 $U=p_3$，$K=2$，算法首先生成两个哈希函数族 $l_1=\{p_1,p_3,p_2,p_4\}$，$l_2=\{p_1,p_2,p_3,p_4\}$，然后将 l_1、l_2 分为容量为 2 的桶 $l_1=\{\langle p_1,p_3\rangle,\langle p_2,p_4\rangle\}$，$l_2=\{\langle p_1,p_2\rangle,\langle p_3,p_4\rangle\}$。这时，$q=p_1$，$\Omega=\{p_3,p_2\}$，$T=\{p_1,p_2\}$，从 L 排序表中移除 T 中的元素后仅剩下$\{p_3,p_4\}$，元素个数<K，返回 MBR(p_3，p_4)，算法终止。

6　实验结果及分析

实验参照了计算几何库 CGAL(http://www.cgal.org/)中的方法。采用 Windows 操作系统，Core2 Duo 1.7Ghz 处理器。实验中，n 个用户位置随机分布在 1000×1000 的区域内。图 5 给出了在 $n=1000$、哈希函数族 $L=20$ 的情况下，ASR 与整个面积的百分比与 K 值的关系。结果显示，ASR 尺寸基本上随 K 值线性变化。图 6 给出了在 $L=10$、$K=10$、n 在区间[500,2000]变化时的情况。

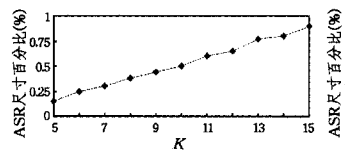


图 5　隐匿尺寸百分比与 K 值的关系

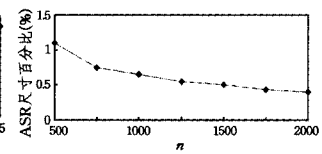


图 6　ASR 尺寸与 n 的关系

图 7 给出了当用户数分别为 1000、5000、1000，$L=20$ 时，运行时间随 K 值变化的情况。结果显示运行时间与 K 值成反比关系。

为了评估哈希函数的个数不同所产生的影响，我们随机生成了 1000 个位置点，L 在区间[2,45]之间变化。如图 8 所示，L 越大，ASR 尺寸越小。当 $L>10$ 之后，隐匿尺寸显著地降低，并稳定下来。

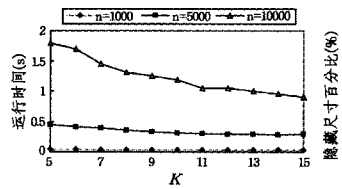


图 7　运行时间与 K 值的关系

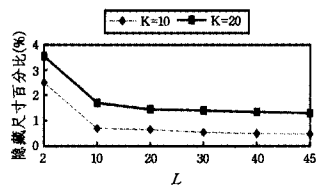


图 8　L 对 ASR 尺寸的影响

结束语　在参与式传感中，用户既可以提供宝贵的信息(数据报告)，也可以检索(依赖所在位置)信息(查询)。隐私保护是数据共享中一个很重要的问题。位置隐私保护通常使用空间 K-匿名技术。目前的主要匿名方法都基于用户-匿名器-LBS 模型。文中提出了一种基于位置敏感哈希分割的空间 K-匿名共匿算法，该算法具有适度的计算复杂度。最后，实验证明了算法在有效性(最小化匿名空间区域)和效率(构建代价)方面具有良好的性能。

参 考 文 献

[1] Agrawal D, Aggarwal C C. On the design and quantification of privacy preserving data mining algorithms [C]// Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. 2001:247-255

[2] Ahmadi H, Pham N, Ganti R, et al. Privacy-aware regression modeling of participatory sensing data [C]// Procee-dings of the 8th ACM Conference on Embedded Networked Sensor Systems. 2010:99-112

[3] Sweeney L. *K*-anonymity: a model for protecting privacy [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5):557-570

[4] 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. 2006:763-774

[5] 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. 2003:31-42

[6] 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):1719-1733

[7] Butz A R. Alternative algorithm for hilbert's space- filling curve [J]. IEEE Transactions on Computers, 1971, C-2(4):424-426

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

[9] Cheng R, Zhang Y, Bertino E, et al. Preserving user location privacy in mobile data management infrastructures [J]. Privacy Enhancing Technologies, 2006, 4258:393-412

[10] Khoshgozaran A, Shahabi C. Blind evaluation of nearest neighbor queries using space transformation to preserve location privacy [J]. Advances in Spatial and Temporal Databases, 2007, 4605:239-257

[11] Hoh B, Gruteser M. Protecting location privacy through path confusion [C]// First International Conference on Security and Privacy for Emerging Areas in Communications Networks. 2005:194-205

[12] Ghinita G, Kalnis P, Khoshgozaran A, et al. Private queries in location based services: anonymizers are not necessary [C]// SIGMOD. 2008:121-132

[13] Khoshgozaran A, Shahabi C, Shirani-Mehr H. Location privacy: going beyond k-anonymity, cloaking and anonymizers [J]. Knowledge and Information Systems, 2011, 26(3):435-465

[14] Ghinita G, Kalnis P, Kantarcioglu M, et al. Approximate and exacthybrid algorithms for private nearest-neighbor queries with database protection [J]. GeoInformatica, 2011, 15(4):699-726

[15] Mouratidis K, Yiu M L. Anonymous query processing in road networks [J]. IEEE Transaction on Knowledge and Data Engineering, 2010, 22(1):2-15

[16] Chow C Y, Mokbel M F, Liu X. Query-aware location anonymization for road networks [J]. GeoInformatica, 2011, 15(3): 571-607

[17] Datar M, Indyk P. Locality-sensitive hashing scheme based on *p*-stable distributions [C]// Proceedings of the Twentieth Annual Symposium on Computational Geometry. 2004:253-262

(上接第 71 页)

[6] Kumar S, Lai T H, Balogh J. On k-coverage in a mostly sleeping sensor network[C] // Proceedings of the 10th Annual International Conference on Mobile Computing and Nerworking. Philadelphia, PA, USA, 2004:144-158

[7] Meguerdichian S, Koushanfar F, Potkonjak M, et al. Coverage problems in wireless Ad-hoc sensor networks[C]// Proceedings of the IEEE Conference on Computer Communications. Anchorage, Alaska, 2001:1380-1387

[8] Fang Q, Gao J, Guibas L. Locating and bypassing routing holes in sensor networks[C]// Proceedings of the IEEE Conference on Computer Communications. Hong Kong, China, 2004:2458-2468

[9] Wang G, Cao G, La Porta T. Movement-assisted sensor deployment[C] // Proceedings of the IEEE Conference on Computer Communications. Hong Kong, China, 2004:2469-2479

[10] So Man Cho A, Ye Y. On solving coverage problems in a wireless sensor network using voronoi diagrams[C]// Proceedings of the International Workshop on Internet and Network Economics (WINE). LNCS 3828. Hong Kong, China, 2005:584-593

[11] Rao A, Ratnasamy S, Papadimitriou C, et al. Geographic routing without location information[C]// Proceedings of the 9th Annual International Conference on Mobile Computing and Networking. San Diego, CA, USA, 2003:96-108

[12] Lederer S, Wang Y, Gao J. Connectivity-based localization of large scale sensor networks with complex shape[C] // Proceedings of the IEEE Conference on Computer Communications. Phoenix AZ, USA, 2008:789-797

[13] Li X, Hunter D K, Yang K. Distributed coordinate-free hole detection & recovery[C]// Proceeding of the IEEE GlobeCom. San Francisco, USA, 2006:1-5

[14] Ashraf H, Radhika T, Chakrabarti S, et al. Approach to increase the lifetime of a linear array of wireless sensor nodes[J]. Wireless Information Networks, 2008(15):72-81

[15] Zeng Z W, Chen Z G, Liu A F. Energy-hole avoidance for WSN based on adjust transmission power[J]. Chinese journal of computers, 2010(33):12-22

[16] Chen Z G, Liu A F, Yang G J. Energy hole avoid by alternately working with different cluster-radius for wireless sensor networks[J]. Journal on Communications, 2010(31):1-8

[17] Jarry A, Leone P, Powell O, et al. An optimal data propagation algorithm for maximizing the lifespan of sensor networks[C]// Gibbons P, ed. Proc. of the Distributed Computing in Sensor Systems(DCOSS). Berlin, Heidelberg: Spring-Verlag, 2006:405-421

[18] 宋超, 刘明, 龚海刚, 等. 基于蚁群优化解决传感器网络中的能量洞问题[J]. 软件学报, 2008, 5(5):8-11

[19] Lian J, Naik K, Agnew G. Data capacity improvement of wireless sensor networks using non-uniform sensor distribution[J]. International Journal of Distributed Sensor Networks, 2006, 2(2): 121-145

[20] BejeranoY. Simple and efficient k-coverage verification without location information[C]// Proceeding of the IEEE Conference on Computer Communications. Phoenix A Z, USA, 2008:291-295