ATLANTIS PRESS
Advances in Engineering Research (AER), volume 61
International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2017)

A Generalization-Based POI Query Privacy Preserving Scheme

Yunxia Feng [1, a], Xu Li [2, b]

[1]College of Information Science & Technology, Qingdao University of Science & Technology, Qingdao 266061, China

[2]College of Electromechanical Engineering, Qingdao University of Science & Technology, Qingdao 266061, China

[a]cloudy_feng@163.com, [b]lixv117@163.com

Keywords: Location Privacy preserving, POI query, K-anonymity, Generalization.

Abstract. Point of interest (POI) is a special kind of point location that combines some information that people may find useful or interesting. Thus, POI query may disclose directly people's sensitive information. Current location anonymity schemes do not pay close attention to information added to locations, whereas cryptographic-based privacy information retrieval (PIR) protocol is too complex for the POI query. This paper studies the problem of privacy preserving in POI query system. Based on characteristics of POI category architecture, we present a generation based k-anonymity POI query scheme. Main idea of the proposed scheme is to decrease the ability of leaking privacy information by using the generalized characteristic of POI architecture. Other advantages of the proposed strategy are as follows: it is a personalized location preserving strategy, and it can be implemented easily on smart mobile devices without supporting of any third part.

1. Introduction

Benefited from both the wide availability of positioning systems and the diffusion of smart-phones, location based service (LBS) has become increasingly popular. As the foothold of a LBS system, POI (point of interest) query service gradually aroused attention. POI is a special point location that someone may find useful or interesting. POIs are mainly employed to describe locations of objects or events to improve service experience quality. POI usually includes three properties: name, classification (or description) and position information. The position information consists of the POI's latitude and longitude in most scenarios. In several scenarios, some other information such as altitude or a telephone number may also be attached to POIs. By utilizing POI query, users can retrieve detailed information of the places in their vicinity. Nevertheless, POI is mainly designed to construct high-quality location based services for the public. Therefore, data attached to POIs usually are public information or partial emergency information those are opened to government only. Theoretically, privacy data that may release state secrets or people's private information should not be included.

However, POIs may also disclose directly people's sensitive information when they are associated with the user's identity. For example, if a user, say Bob, sends several POI queries and the type of the POI is cancer hospital, then we can deduce that either Bob or his family may have gotten cancer or have shown similar symptoms. Cancer hospital itself is public information. Nevertheless, POIs connect Bob with the cancer hospital, and consequently, Bob's health condition information is leaked. Thus, POI query may lead to privacy leakage.

Widely discussed and researched location protection technologies include pseudonym, anonymity, encryption and perturbation. Private information retrieval (PIR) is a typical representative of cryptographic theory based privacy protecting protocol. PIR allows a user to retrieve elements of a database without the owner of that database being able to determine which element was selected. PIR theory has attracted considerable attention [6-8] in recent years due to its advantage in privacy protection. Ghinita G, and etc. [6] introduced private information retrieval (PIR) theory to avoid information leakages and provide strong privacy in shortest path computation. Whereas, the server cannot obtain any clues about which data was retrieved. While PIR ensures perfect privacy, the

communication complexity of PIR can be critical for large databases. Although several PIR protocols declare that they achieve practical response times (in the order of seconds over Gigabyte databases [7]), PIR is proved to be generally resource-intensive [8]. A great number of POR queries are executed by mobile users through smart mobile terminals. Additional overheads induced by PIR-based resolutions are inacceptable for mobile users.

On the other hand, anonymity can ensure both reality and privacy security, and it has attracted considerable attention. One of the most important anonymity-based privacy protecting technology is *k*-anonymity [2]. And location k-anonymity is an extension of the general *k*-anonymity model to provide location protection [7, 8]. Location *k*-anonymity can be implemented in many ways, among which, generalization and suppression are the two most important [9, 12-17]. Generation describes data in more general and abstract ways, whereas suppression hides (deletes) some sensitive data. By hiding (deleting) sensitive data, suppression guarantees users high privacy protecting degree. Nevertheless, too many sensitive data are suppressed may induce huge information loss. On the other hand, generation achieves desirable balance in privacy protection degree and data availability. Thus, generation based k-anonymity is a major method of privacy protection.

Due to limitation of space, we just introduce several generation based *k*-anonymity resolutions here. Must current generation based location *k*-anonymity resolutions adopt either spatial [9, 12, 13, 15, 16] cloaking technology or temporal [14, 17] cloaking technology. Temporal *k*-anonymity cloaking, such as CliqueCloak [9], will wait until at least *k* different queries have been sent from a particular region. This allows the k-anonymous area to be smaller in space, but it may extend the waiting time of some users by forcing them to wait until there are *k* users, thus anonymity can take place. Spatial *k*-anonymity cloaking can be achieved by sending a sufficiently large *k*-anonymous region that encloses *k* users in space [14], instead of reporting a single. However, an adversary may manage to identify that a spatial region has been visited by *k* different people, but it won't know who was there at time of the service request. Intuitively, creating a region around multiple users significantly decreases spatial accuracy.

However, above location *k*-anonymity implementations reduce the quality of user's localization in space or time and may prevent meaningful use of various LBSs, especially when user density is low. In these scenarios, current resolutions need additional delay or large region. Then, CacheCloak [15] was developed to enable real-time anonymity of location data. According to CacheCloak, a trusted anonymous server generates mobility predictions for each mobile user from historical data, and submits intersecting predicted paths with the users' locations simultaneously to the LBS. Each new predicted path is made to intersect with other users' paths, ensuring that no individual user's path can be reliably tracked over time. Mobile users retrieve cached query responses for successive new locations from the trusted server, triggering new prediction only when no cached response is available for their current locations.

To the best of our knowledge, few researches have considered characteristics of POI query. Most of current location *k*-anonymous resolutions concentrate on location information itself, and do not consider additional information attached to the location in POI query. If most of the *k* users in the anonymity region query the same type of POI, an adversary may get useful information in this scenario.

In this paper, we investigate the problem of privacy preserving in POI query and present a generation based k-anonymity POI query scheme. Main idea of the proposed scheme is to decrease the ability of leaking privacy information by using the generalized characteristic of POI architecture. We first study the privacy leaking ability of different POI categories and present a new metric, called PPSD (POI privacy sensitivity degree), to denote the privacy leaking probability of a given POI category. We then describe the PPSD based *k*-anonymity POI query scheme, and discuss its implementation technology, performance and availability using a simple example. The proposed scheme assumes that users may access POIs of a given region, and there is no third part location anonymous servers between mobile users and LBS servers.

The rest of this paper is organized as follows. In Section II, we give the problem description, assumptions and illustrate details of PPSD. In Section III, we illustrate the proposed *k*-anonymity POI query scheme using a simple example. At last, we provide a short discuss and conclude this paper.

2. The Privacy Sensitivity Grading Scheme (PSGS)

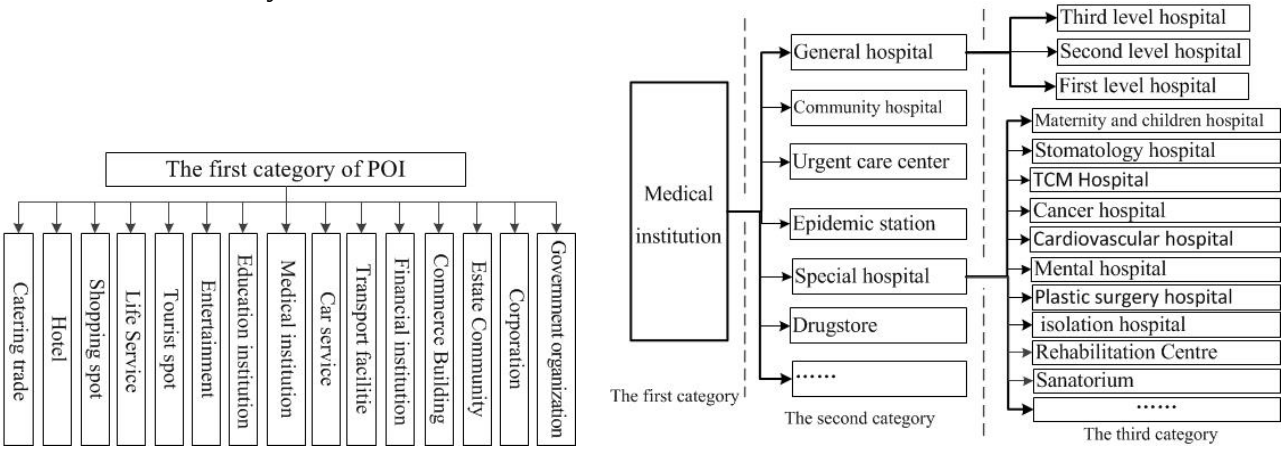
In this section, we first study characteristics of POIs provided by LBS providers and then present several essential definitions and assumptions. We then describe main idea of the proposed scheme and illustrate it using a POI category example from Baidu maps.

2.1 Characteristics Analysis

In order to study characteristics of POIs, we take the simplified version of POI category scheme adopted by baidu map as the example to illustrate characteristics of POIs. The data structure of every POI category is represented in a tree-like structure, with a root POI at the top of the tree and children under it. A parent POI in a tree is the generalization result of all of his children in the POI category. Every POI category in the higher layer may have multiple sub-classifications as children. Nevertheless, each POI category can only belong to at most one specific classification as its parent. Furthermore, the height of a POI category tree is decided by specific condition of the corresponding POI category. Let H be the height of a POI tree. The level of the root POI category is set to 1, and the level of each POI category in the i-th ($1<i\leq H$) layer of the tree is set to i. For ease of explanation, we named the POI categories which level is i ($1<i\leq H$) by the i-th category.

In this example, the height of the medical institution category tree is 3. Nevertheless, the architecture is extensible. For example, the third level hospital (in the third category) can be further classified into three sub-categories: the third-level 1st Class (grade-A) hospital, the third-level 2nd Class hospital, and the third-level 3rd Class hospital. In this scenario, the medical institution POI tree consists of four layers.

The first (1-th) category of baidu map POI category consists of 15 POI categories, and every POI classification in the first category includes multiple children categories. Due to space limitation, we just take the medical institutions as example to denote partial contents included in the second category, and general hospital and special hospital as example of the third category. Contents of the first category and the architecture of the medical institution category are shown in Figure 1(a) and (b), respectively. Interested readers may refer to LBS related websites for more details.



(a) Example of the first category (b) The tree architecture of the medical institution category

Fig. 1 An example of the POI category architecture (from baidu maps)

As shown in Figure 1, the ability of leaking privacy information of a POI category is proportional to its level. That is, the higher the level of a POI category is, the more sensitive information it contains.

Studies have shown that privacy is a relative concept, and different people may have different concerns. Thus, the ability of leaking privacy information for a specific POI category is affected by the users. Ackerman and etc. [1] conducted a survey of 381 Internet users from the U.S., detailing a range of commerce scenarios and examining the participants' concerns and preferences about privacy. They attempted to look beyond the fact that people are concerned in order to understand what aspects of the problem they are most concerned about. Their results also show that the attention people paid to different sensitive information is different. What current studies show us is that POI's privacy leaking ability is affected by its category type. For example, the POI of shopping mall usually denotes less sensitive information than that of a hospital.

2.2 Location Privacy Sensitivity Degree

Characteristics of POIs show that the ability of leaking privacy information for a specific POI category is affected by both his category type and level. Based on this observation, we introduce a new term, called PPSD (location privacy sensitivity degree) to denote the ability of leaking privacy information of POI. PPSD is a nonnegative integer assigned to every POI. The value of PPSD denotes reflects strong or weak of the privacy leaking probability. The larger the PPSD assigned to a POI is, the stronger the privacy leakage probability of the POI will be.

Assume that each POI category is assigned a PPSD. Let S_{hspp} be the set of all candidate PPSDs, and let h be the maximum candidate PPSD. According to the definition of PPSD, $S_{hspp}=\{0,1,2,\ldots\ldots, h\}$. Theoretically, h can be any positive integer. However, the implementation complexity of the privacy protection scheme increases as h grows. Hence, h should be the trade-off result between the user's privacy protection requirement and the equipment of the user's mobile device.

The next question is how to select a proper PPSD for POIs. Characteristics of POIs tell us that the PPSD assigned to a specific POI category should consider his category type and level. We further noticed that every POI tree is the result of multiple generalizations starting from leaf nodes. The higher the height of a node in the tree, the larger the generalization degree is. Hence, the privacy information leaking ability of children is greater than their parents. Therefore, the corresponding PPSD of a child should be no smaller than that of its parent.

In order to reflect impacts of type and level, the PPSD assignment procedure is divided into 2 steps: In the first step, every first POI category is assigned a PPSD. We take the research results achieved by Ackerman and etc. [1] as criteria in assigning PPSDs to different POI categories. That is, the medical institution category is assigned the highest PPSD, followed by the financial institution, and etc. Taken the first POI category shown in Figure 1(a) as example, the PPSD assignment results are shown in Figure 2:

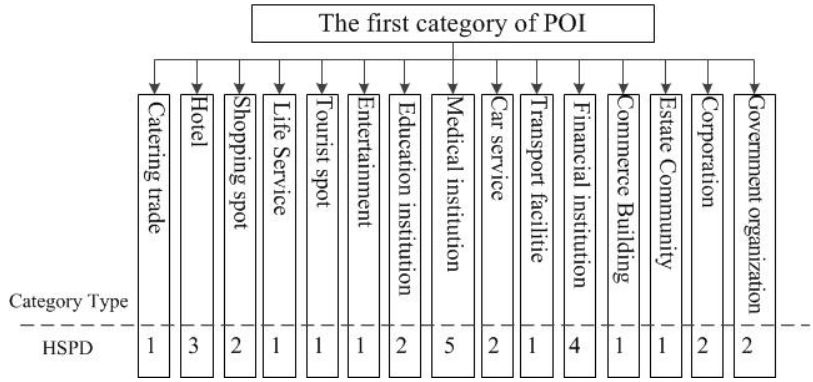


Fig. 2. An example of assigning PPSD to the first POI category

It should be noted that, in practice, the real PPSD assigning results are decided by the specific user. And for different people the results may be different with each other since they may have different privacy preserving requirements.

In the second step, starting from the 2-th (second) category, POIs of every POI category tree is assigned a PPSD in sequence obeying following rules to meet requirements described above:

A POI can be assigned PPSD only if its parent node has been allocated PPSD.

The PPSD of a child node should be bigger than that of its parent.

Different children of the same POI can be assigned different PPSDs.

We then take the medical institution shown in Figure 1(b) as example to illustrate the PPSD assignment for sub-categories. As shown in Figure 2, the PPSD of the medical institution category is 5. PPSDs of all its subcategories should be no less than 5 according to the PPSD assignment rules provided above. One possible assignment result is shown in Figure 3:

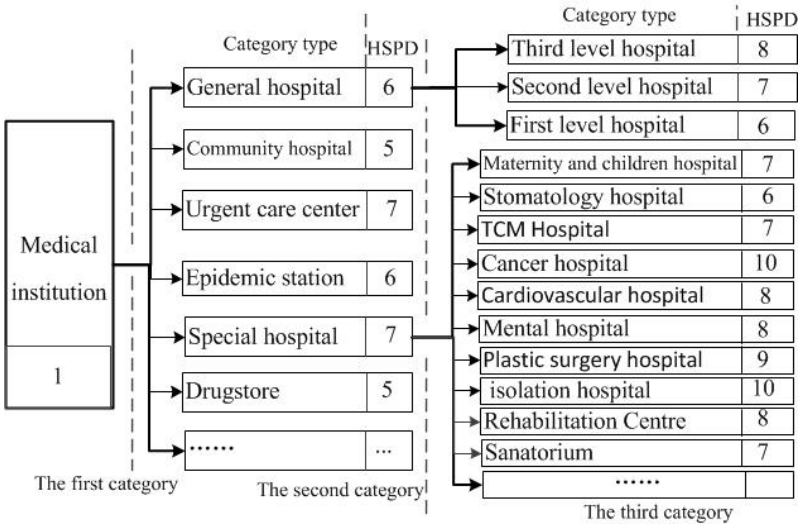


Fig. 3. One PPSD assigning example of Medical institution category tree

3. PPSD Based k-anonymity POI Query Scheme

Main idea of the proposed *k*-anonymity POI query scheme is to control its ability of leaking privacy information in an acceptable range by generalized. In order to achieve this goal, the proposed scheme replaces the real query location with a high PPSD by a POI which PPSD satisfies the user's specific privacy preserving requirement. In practice, the proposed scheme requires the user to provide two parameters to describe his specific privacy requirement: k and l_e. Where, k is the number of POIs in the sub tree rooted at the selected POI, whereas l_e denotes his specific PPSD requirement for the selected POIs, and l_e is selected from the set S_{hspp}.

Let x be the location included in the true POI query, and l_x be the PPSD of x. Assume that both k and l_e are known. Main procedure to generate the generalized POI is as following. If $l_e \leq l_x$, then, starting from the parent POI of x, it searches up the corresponding POI category tree and compares the PPSD of every node on it with l_e. The searching process continues until it finds a POI, say y, which PPSD is no bigger than l_e. Let c be the number of POI categories included in the sub-tree which root is y and x is its leaf. If $k \leq c$, then, it replaces x by y in the query message. Otherwise, the searching process continues.

To facilitate the understanding of the proposed anonymity procedure and scheme, we take another example. In this example, assume that the value of k and l_e is 7 and 2, respectively. Assume that the POI corresponding to the true location is Stomatology hospital. According to Figure 4, the PPSD of the Stomatology hospital category is 6. Apparently, 6 is smaller than 7. The first condition is satisfied. However, the Stomatology hospital is the leaf node. Hence, it does not have any child. If we select the POI, the number of POI categories included in the query is 1. Apparently, it does not satisfy the second condition. Then, the searching process continues. Fortunately, its parent, the special hospital has 11 sub-categories, which is much bigger than 2. At this time, the second condition is satisfied. In this example, the Special hospital is selected and is included in the query message sent to the lbs server.

From the example, we can see that, the query after anonymity consists of c (c is no smaller than k) different POI categories. PSGS based *k*-anonymity POI query can be easily achieved without third part. Nevertheless, PSGS needs that the users may access POIs of a given region. This condition can be easily satisfied since there are multiple POI providers and most of them provide free POI packages on the web. With the popularity of smartphones and high-capacity removable storage card, users can download and store POI packages on their smartphones without affecting normal operation.

4. Conclusion

In this paper, we studied the problem of privacy preserving in POI query. Taking advantages of the layered data structure of POI category, we proposed PPSD, a new criteria to reflect privacy leaking ability differences between POI categories. Based on PPSD, we then present a personalized POI query

privacy preserving scheme. The proposed scheme is a generation based extension of location *k*-anonymity model and it can be implemented easily on smart mobile devices without supporting of any third part. Users may change the privacy preservation degree dynamically and control the granularity of the location they send to LBS servers. Hence, users may play a part in the generalization process and control the granularity of the location they send to LBS servers.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China under Grant 61303193, 61572268 and 61402246, Qingdao indigenous innovation program (No. 15-9-1-47-jch.

References

[1] MS Ackerman, LF Cranor, J Reagle, Privacy in e-commerce: Examining user scenarios and privacy preferences[C], ACM Conference on Electronic Commerce, 2000: 1- 8.
[2] L. Sweeney, k-anonymity: A model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), 2002. pp. 557-570
[3] Qiuyu Xiao, Jiayi Chen, Le Yu, Huaxin Li, Haojin Zhu, Muyuan Li, Kui Ren, LocMask: A Location Privacy Protection Framework in Android System, ACM CCS. 2014.
[4] J-H Song , V. W. Wong , V. C. Leung, Wireless Location Privacy Protection in Vehicular Ad-Hoc Networks, Mobile Networks and Applications, 15(1), 2010, pp.160-171
[5] A. R. Beresford and F. Stajano, Location Privacy in Pervasive Computing, IEEE Pervasive Computing, 2(1), 2003, pp. 46-55.
[6] Ghinita G, Kalnis P, Khoshgozaran A, Shahabi C, Tan KL, Private queries in location based services: Anonymizers are not necessary, ACM SIGMOD, 2008: 121-132.
[7] P. Williams and R. Sion, Usable PIR, NDSS, 2008.
[8] S. Papadopoulos, S. Bakiras, and D. Papadias, Nearest neighbor search with strong location privacy. PVLDB, 2010, 3(1): 619-629.
[9] B. Gedik, L. Liu, and G. Tech, Location Privacy in Mobile Systems: A Personalized Anonymization Model, IEEE International Conference on Distributed Computing Systems (ICDCS), 2005.
[10] B. Gedik, L. Liu, Protecting location privacy with personalized k-anonymity: architecture and algorithm. IEEE Trans Mobile Comput, 7(1), 2008. pp. 1-18
[11] Y. Feng, P. Liu and J. Zhang, A Mobile Terminal Based Trajectory Preserving Strategy for Continuous Querying LBS Users, IEEE International Conference on Distributed Computing in Sensor Systems, Hangzhou, 2012, pp. 92-98.
[12] T. Xu, and Y. Cai, Exploring Historical Location Data for Anonymity Preservation in Location-based Services, IEEE Infocom, 2008. pp. 547-555.
[13] 10] G. Ghinita, K. Zhao, D. Papadias, P. Kalnis, A reciprocal framework for spatial K-anonymity, Information Systems, 35(3), 2010. pp. 299-314.
[14] T. Xu, and Y. Cai, Exploring Historical Location Data for Anonymity Preservation in Location-based Services. IEEE Infocom, 2008.
[15] J. Meyerowitz, and R. R. Choudhury, Hiding stars with fireworks: location privacy through camouflage, ACM MobiCom, 2009.
[16] G. Ghinita, K. Zhao, D. Papadias, P. Kalnis, A reciprocal framework for spatial K-anonymity, Information Systems, 35(3), 2010. pp. 299-314.
[17] M. Gruteser, and D. Grunwald, Anonymous Usage of Location-based Services Through Spatial and Temporal Cloaking, MobiSys, 2003.