World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3013

TRAJECTORY DIFFERENTIAL PRIVACY PROTECTION MECHANISM BASED ON SEMANTIC LOCATION CLUSTERING

ShanLin Yu, Hui Wang*
School of Computerscience and Technology, Henan Polytechnic University, Jiaozuo 454003, Henan, China.
Corresponding Author: Hui Wang, Email: wanghui_jsj@hpu.edu.cn

Abstract: Aiming at the problems of being vulnerable to semantic attacks and having low data availability in the current trajectory data privacy protection schemes, a trajectory differential privacy protection scheme based on semantic location clustering is proposed. Firstly, the semantic distances between various positioning points in the trajectory are estimated by sorting out the logical relationships of different semantic concepts. Then, the clustering algorithm is used to generate clustering results with members having high semantic similarity for the trajectory data set as anonymous sets. Secondly, the differential privacy exponential mechanism is utilized to select representative positions with a lower possibility of privacy leakage from the clustering results to anonymize the sensitive points in the original trajectory, which achieves good privacy protection effects while avoiding large information losses.
Keywords: Semantic distance; Location clustering; Differential privacy; Privacy protection

1 INTRODUCTION

With the growing prevalence of mobile intelligent devices and the swift progress of the mobile Internet and GPS, Location Based Services (LBS) have become more and more popular in daily life. They now cover every aspect of the national economy and social life, enabling people to enjoy unprecedented convenience in the mobile Internet era. However, their geographical location data is also being massively collected, analyzed, and utilized to enhance service providers' operational quality. But if no effective protection measures[1] are taken when releasing and using users' trajectory data, it will result in severe privacy leaks and even endanger personal and property safety.

The academic investigation into trajectory privacy protection predominantly concentrates on privacy affairs in two application scenarios: real-time trajectory privacy protection in location services[2-4] and offline trajectory privacy protection in data publishing[5-7]. Differential privacy is the most widely used privacy protection technique in offline trajectory data publishing, but it has the problem of being difficult to balance privacy efficiency and data availability. Hua et al.[8] merged similar points at the same timestamp based on clustering to achieve location generalization and added Laplace noise to the generalized location domain to generate publishable privacy trajectories. Zhao et al.[9] were more concerned about privacy protection performance in cluster analysis. So they added Laplace noise subject to radius constraints to the trajectory locations, cluster centers, and location counts of each cluster to resist cluster location attacks and continuous query attacks. MA et al.[10], from the perspective of algorithm efficiency, proposed a differential privacy protection method based on random sampling. They added a random sampling process during trajectory clustering and used false locations close to the cluster center to replace other points in the cluster for synthesizing privacy trajectories, effectively improving the execution efficiency of the algorithm. Zhen et al. [11] considered that using false locations to generalize trajectories might lead to the published trajectories being recognized and filtered by adversaries. So they proposed using the differential privacy exponential mechanism to randomly select real locations from the generated clustering results as representatives of other points in the same cluster to form generalized trajectories for data publishing. Although the above methods can achieve good privacy protection performance, they still have problems such as being vulnerable to semantic attacks and having low data availability. Therefore, this article makes improvements for the above issues and proposes a trajectory differential privacy protection algorithm based on semantic location clustering.

2 PRELIMINARY KNOWLEDGE

Definition 1 (Semantic Trajectory). A sequence of semantic locations consisting of n elements arranged in chronological order according to timestamps is called a semantic trajectory: $l=\{p_1,p_2,\cdots,p_n\}$. Each location point p_i therein records several different attributes, which are user ID, latitude and longitude coordinates, semantic label, timestamp, dwell time, and so on. The set D composed of N semantic trajectories $l_1,l_2,...l_N$ is called a semantic trajectory data set.

Definition 2 (Differential Privacy)**.** For any two adjacent datasets D' and D, as well as a randomized algorithm M whose output space set is R, if they can satisfy the following condition:

$$P(M[D]\in S)\le \exp(\varepsilon)\times P(M[D']\in S)$$

(where S is an arbitrary subset of R, ε is the privacy budget, and $P[\cdot]$ represents the probability of the corresponding event occurring), then the randomized algorithm M is said to satisfy ε -differential privacy.

Differential privacy can provide strict privacy protection for sensitive information. Its core idea is to introduce a randomization mechanism to perturb the original data, so that third parties cannot determine the specific changes in the output content based on the modification, addition or deletion of a single record. Based on the above definitions, people have proposed multiple techniques for achieving differential privacy. Among them, the two most important ones are the Laplace mechanism and the exponential mechanism:

Laplace Mechanism: For a dataset D and an arbitrary function f, if there exists a randomized algorithm M that can satisfy ε -differential privacy, then we have $M[D]=f(D)+Y$. Y is the random noise that follows the Laplace distribution, denoted as $Y \sim Lap(\Delta f / \varepsilon)$, where $\Delta f = \max \| f(D) - f(D') \|_1$ represents the global sensitivity.

Exponential Mechanism: For a dataset D and a randomized algorithm M whose output space set is R, if it can select and output a result r from R with a probability $P(r)$ that is proportional to $\exp\left[\frac{\varepsilon\mu(r)}{2\Delta\mu}\right]$, then the randomized algorithm M can satisfy ε -differential privacy, where $u(r)$ represents the utility score of the output result r.

Definition 3 (Semantic Classification Tree)**.** A tree structure formed by classifying and organizing all the location points in the trajectory dataset according to their semantic concepts is called a semantic classification tree, denoted as $Tr=\{C,h\}$. C represents the semantic classification concept on each layer, h represents the level of semantic concept classification, and each leaf node represents a certain real location on the map.

Definition 4 (Semantic Similarity)**.** The degree of similarity in the semantic concept classification of different location points in a trajectory is called semantic similarity, denoted as $s(x,y)\in[0,1]$, where x and y represent the semantic concepts of any two location points.

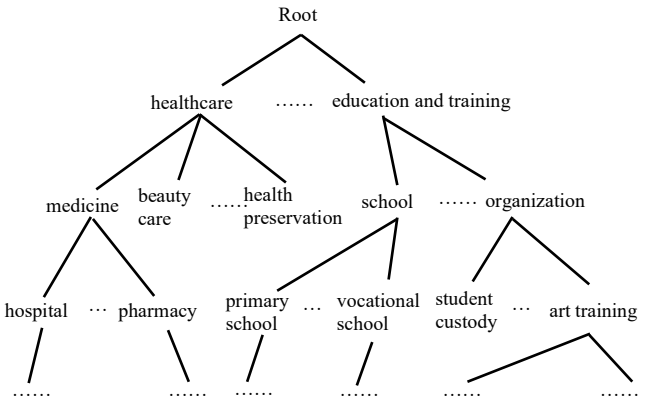

Figure 1 Semantic Classification Tree

As shown in Figure 1, through relevant information such as the level and branches where the semantic concepts corresponding to geographical locations are located in the semantic classification tree, the degree of closeness of the semantic relationships between different location points can be defined, so as to estimate their semantic similarity[12]. Generally speaking, the semantic similarity can be comprehensively evaluated from multiple aspects of factors, such as the structural information of the classification tree like its depth and width, the path distance between different nodes, the hypernym-hyponym semantic relationships between concepts, and the whole-part semantic relationships.

In the scenario of semantic location clustering[13], there are usually some location points that are relatively close in spatial distance but have very large differences in semantic information among them. If at this time, the Euclidean distance is still used as the measurement standard according to the conventional clustering algorithm without correlating the location semantic features, it is very likely to lead to the clustering results being inconsistent with the actual situation, thus reducing the availability of the finally generated anonymous trajectory data set. Therefore, this article chooses to use the Euclidean distance fused with semantic features, that is, the semantic distance, to conduct clustering operations. Its formula is as follows:

$$d_s(p_i,p_j)=Log_\beta[\alpha\times s(x,y)]\times d_e(p_i,p_j)$$

Among them, $d_s(p_i,p_j)$ and $d_e(p_i,p_j)$ respectively represent the semantic distance and the Euclidean distance between different location points. The parameters α and β are arbitrary real numbers in the interval (0,1). α is used to control the scaling degree of the similarity to the spatial distance, while β can adjust the magnitude of the output semantic distance value. It can be seen from the formula that the essence of the semantic distance is to use the semantic similarity to scale the spatial distance between location points, so that those points with more similar semantic features are closer to each other, in order to generate more accurate clustering results.

It should be noted that when calculating the semantic distance, attention also needs to be paid to and the following two difficult problems need to be solved: Firstly, there is the issue of parameter values. Since semantic similarity $s(x,y)$ has a significant impact on the final output value of the semantic distance, if it is set unreasonably, it may lead to the final result deviating from the original data scale and not conforming to the actual geographical scale. Secondly, there is the problem of computational efficiency. When performing semantic location clustering, if the spatial distance between two

points is far enough, it can be considered that the possibility of them being divided into the same cluster is low. If the semantic distance is still used for measurement at this time, it will result in more algorithm running time being occupied.

3 TRAJECTORY PRIVACY PROTECTION ALGORITHMS

3.1 Steps of the Algorithm

The privacy protection scheme in this paper mainly consists of three steps (as shown in Figure 2):
(1) Divide the trajectory data set into multiple different location subsets according to the specified time stamp, so that the time records of each location point in the same set are the same or similar.
(2) Use the semantic distance to perform clustering operations on each divided location subset, and generate multiple different clusters as anonymous sets.
(3) Randomly select the real records that meet the privacy requirements from the clustering through the exponential mechanism as the representative positions to conduct privacy processing on the sensitive points in the trajectory data.

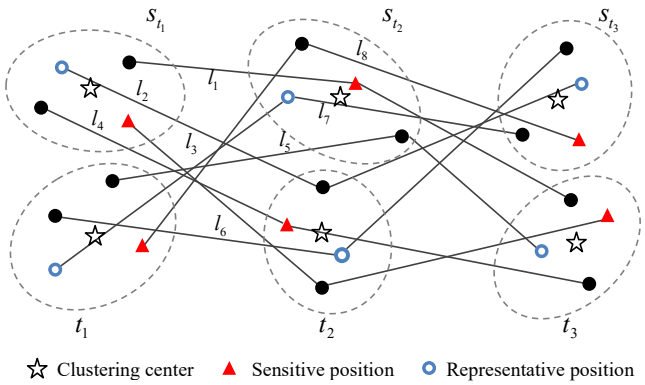

Figure 2 Schematic Diagram of the Privacy Protection Scheme

The specific algorithm steps are shown as follows:
Algorithms 1. Differential Privacy Protection Algorithm Based on Semantic Location Clustering
Input: Semantic trajectory data set D.
Output: Publishable trajectory data set D'.

1 $D' = \varnothing$;
2 *Divided SD into t subsets* $S_T = \{S_1, S_2, ..., S_t\}$;
3 **For each** S_t **in** S_T
4 *Calculate the semantic distance SD*[ ] *among the points of* S_t;
5 *Using k-means++ to cluster locations at* S_t *by SD*[ ] *into k classes* $C_{S_t} = \{c_t^1, c_t^2, ..., c_t^k\}$;
6 $S_t' = \varnothing$;
7 **For each** c_t^k **in** C_{s_t}
8 *Obtain the probability a ray Pr*[ ] *of selecting representative positions for* c_t^k;
9 **For** $i = 0$ *to* $|c_t^k|$
10 *Screen out the sensitive positions from* c_t^k;
11 *Select representative location from* c_t^k *by array Pr*[ ];
12 *Add Laplacian noise to representative location and Replace the sensitive position*;
13 $S_t' = S_t' \cup c_t^k$;
14 $D' = D' \cup S_t'$;
15 **Return** D'

In the above algorithm, in lines 4-5, first, the semantic distance generation model is used to obtain the semantic distances between different location points in the set. Then, the partitioning clustering method[14] k-means++ is employed to conduct clustering operations on each set S_t according to the generated semantic distances, thus generating k anonymous sets with high semantic similarity for it. In line 8 of the algorithm, the representative position selection model is utilized to obtain the probability that each point in the cluster c_t^k may be output as a representative position. Since the privacy performance of the differential privacy exponential mechanism mainly depends on two factors, namely the scoring function and the privacy budget, in the scheme of this paper, the distance between the location point and the cluster center is used to design the scoring function, so that the average distance between the selected point and all other points in the same cluster is relatively short. Meanwhile, considering that the possibility of privacy leakage

usually has a strong correlation with the length of the stay time of mobile users at a certain location point, the allocation of the privacy budget parameter is defined by the stay time attribute of the semantic location. Line 10 of the algorithm is used to screen the sensitive locations in the trajectory that are prone to causing privacy leakage. They usually refer to those places that users have frequently visited or stayed at for a long time, that is, stop points[15], which contain abundant personal sensitive information and are the key objects for privacy protection in this paper. Line 12 shows the privacy processing method. The representative positions selected are randomized by adding Laplace noise, and then relevant attributes are extracted to replace the privacy information in the sensitive points, thus completing the anonymization operation. Finally, in lines 13-14 of the algorithm, the *k* clusters that have undergone privacy processing are aggregated into privacy location subsets $S_t^{'}$, and the sets $S_1^{'},S_2^{'},\cdots,S_t^{'}$ under *t* different time stamps are linked to generate a publishable privacy trajectory data set D' .

3.1 Algorithm Analysis

If the trajectory data set is divided into *t* location subsets, and each set S_t contains *n* location points, then the time complexity of the scheme in this paper can be expressed as $O(Ctn^2)$, that is, the time cost of the algorithm is mainly reflected in the location clustering of semantic trajectories. The higher the clustering efficiency is, the better the time performance of the algorithm will be.

In addition, since both the exponential mechanism and the Laplace mechanism are used in Algorithm 1, if their privacy budgets are set as ε_1 and ε_2 respectively, then according to the serial composition property of differential privacy, it can be known that the privacy transformation executed on the location points in each cluster by the algorithm will satisfy $(\varepsilon_1+\varepsilon_2)$ -differential privacy. Meanwhile, since the algorithm divides the original trajectory data set into t location subsets which are all independent of each other, it can be known according to the parallel composition property that the finally output privacy trajectory will also meet the requirements of differential privacy, and its privacy budget is $\max[(\varepsilon_1+\varepsilon_2)_1,\cdots,(\varepsilon_1+\varepsilon_2)_t]$.

4 CONCLUSION

This paper mainly studies the relevant issues in the privacy release scenario of semantic trajectory data and makes adjustments and improvements to the existing algorithms. Since the scheme in this paper takes into account both the spatial and semantic characteristics of location points, it has the advantages of low information loss and good privacy protection effect. In the following research work, the allocation of the privacy budget will be adjusted in combination with users' personalized needs, and the feature dimensions in semantic clustering will be enriched, so as to further improve the privacy protection model for trajectory data.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Zhang Qingyun, Zhang Xing, Li Wanjie, et al. Overview of location trajectory privacy protection technology based on LBS system. Application Research of Computers, 2020, 37(12): 3534-3544.

[2] Cunningham T, Cormode G, Ferhatosmanoglu H, et al. Real-world trajectory sharing with local differential privacy[J]. Proceedings of the VLDB Endowment, 2021, 14(11): 2283-2295.

[3] Zheng Zhirun, Li Zhetao, Jiang Hongbo, et al. Semantic-Aware Privacy-Preserving Online Location Trajectory Data Sharing. IEEE Trans on information forensics and security, 2022, 17: 2256-2271.

[4] Liu Peiqian, Jia Qinglin, Wang Hui, et al. Differential privacy trajectory privacy protection scheme based on user correlation. Application Research of Computers, 2024, 41(7): 2189-2194.

[5] Kim J W, Jang B. Deep learning-based privacy-preserving framework for synthetic trajectory generation. Journal of network and computer applications, 2022, 206: 103459.

[6] Wen Ruxue, Cheng Wenqing, Huang Haojun, et al. Privacy Preserving Trajectory Data Publishing with Personalized Differential Privacy\\ Proc of IEEE Intl Conf on ISPA/BDCloud/SocialCom/SustainCom. Exeter, United Kingdom: IEEE Press, 2020: 313-320. DOI: 10.1109/ISPA-BDCloud-SocialCom-SustainCom51426.2020.00065.

[7] Zhang Jing, Li Yanzi, Ding Qian, et al. Successive Trajectory Privacy Protection with Semantics Prediction Differential Privacy. Entropy, 2022, 24(9): 1172.

[8] Hua Jingyu, Gao Yue, Zhong Sheng. Differentially private publication of general time-serial trajectory data// Proc of IEEE INFOCOM. Hong Kong, China: IEEE Press, 2015: 549-557. DOI: 10.1109/INFOCOM.2015.7218422.

[9] Zhao Xiaodong, Pi Dechang, Chen Junfu. Novel trajectory privacy-preserving method based on clustering using differential privacy. Expert Systems with Applications, 2020, 149: 113241.

[10] Ma Tinghuai, Song Fagen. A Trajectory Privacy Protection Method Based on Random Sampling Differential Privacy. ISPRS International Journal of Geo-Information, 2021, 10(7): 454.

[11] Gu Zhen, Zhang Guoyin. Trajectory Data Publication Based on Differential Privacy. International Journal of Information Security and Privacy, 2023,17(1): 1-15.
[12] Guo Kun, Wang Dongbin, Zhi Hui, et al. Privacy-preserving Trajectory Generation Algorithm Considering Utility based on Semantic Similarity Awareness// Proc of IEEE International Conference on Communications. Seoul, Korea: IEEE Press, 2022: 992-997. DOI: 10.1109/ICC45855.2022.9838628.
[13] Cheng Wenqing, Wen Ruxue, Huang Haojun, et al. OPTDP: Towards optimal personalized trajectory differential privacy for trajectory data publishing. Neurocomputing, 2022, 472: 201-211.
[14] Li Meng, Zhu Liehuang, Zhang Zijian, et al. Achieving differential privacy of trajectory data publishing in participatory sensing. Information sciences, 2017, 400-401: 1-13.
[15] Gao Zhigang, Huang Yucai, Zheng Leilei, et al. Protecting Location Privacy of Users Based on Trajectory Obfuscation in Mobile Crowdsensing. IEEE Trans on industrial informatics, 2022, 18(9): 1.