Personal Information Protection Based on Big Data

Yousheng Zheng¹,*

¹QuanZhou Preschool Education College, 362000, QuanZhou, FuJian, People R China

*Corresponding author e-mail: zhengyousheng1@qztc.edu.cn

Abstract. With the popularization and development of information technology, all walks of life have accumulated a large amount of rich data, which usually contains a lot of personal privacy information, and the direct release or analysis of it may cause the disclosure of privacy. Differential privacy, as a relatively new privacy protection model, can prevent the attack of the attacker with arbitrary background knowledge, and effectively solve the privacy threat problem in data release and analysis. This paper mainly studies the protection of personal information based on big data. In this paper, a position privacy protection scheme based on differential privacy in P2P mode is proposed, and the advantages of the scheme in performance are proved by experimental comparison and theoretical analysis.

Keywords: Big Data, Privacy Protection, Differential Privacy, K Anonymity

1. Introduction

In recent years, with the continuous expansion of the application fields of Internet, cloud computing and other technologies, massive data has been continuously produced in individuals, enterprises and research institutions. At present, the term "big data" has penetrated into all aspects of people's lives, and big data has comprehensively entered the explosive development stage. As an important strategic resource, big data is also being studied, formulated and implemented in China. Although big data to promote the information industry and cross-border integration, become the speed wisdom city construction, promote the information consumption, a new engine to promote the development of economic and social transformation, for each domain brings precious opportunities for development, but because of the complexity and uncertainty of big data in the future will face a huge challenge in the field of data security, privacy and security issues will is one of the bottlenecks restricting the development of large data [1]. Nowadays, more and more Internet companies and social organizations are interested in collecting and analyzing user data. For example, browsers and mobile application software train machine learning models and analyze user behavior patterns by collecting user terminal data all the time. Social service enterprises make corresponding personalized service plans by collecting users' life statistics data. Although these have brought practical help and convenience to our lives, it should not be ignored that, in this case, these data are likely to cause the disclosure of users' personal privacy information due to data analysis and release.
With the rapid development of information technology, it brings convenience and quickness to people's communication and data sharing, and at the same time, the risk of privacy leakage caused by it also increases. In recent years, there have been a number of high-level persistent threats to information systems, the main purpose of which is to enable the attacker to obtain recoverable data. Andor improved K-anonymity and proposed L-diversity, which is a group-based anonymization and protects privacy in data set by reducing the granularity of data representation. However, L-diversity cannot completely protect user privacy from disclosure, because in a real data set, attribute values are likely to be skewed or semantically similar, while L-diversity only ensures diversity and fails to recognize the situation that attribute values are semantically similar [2]. The T-Closeness scheme proposed by Ahmet is a further improvement of the L-Diversity. It makes up for the disadvantage of L-diversity by considering the distribution of attribute data values and treating them differently. However, with the increase of data size and types, the probability of re-recognized also increases [3].

In this paper, by studying the existing location privacy protection schemes based on P2P distributed system architecture, differential privacy is introduced into the location privacy protection scheme, and a differential privacy protection algorithm based on P2P is proposed. In this scheme, K anonymity technology and differential privacy are combined to resist background knowledge attack and regional center attack. Finally, the advantages of this scheme are demonstrated through experiments and safety analysis.

2. Overview of Relevant Technologies

2.1. Privacy Protection Models and Technologies

Sensitive information that data owners do not want to disclose is called private. Privacy protection is known as a measure to prevent malicious users from using their available information to identify physical existence or link information to physical existence, with the purpose of protecting personal information contained in data from leakage [4]. Traditional privacy protection technologies fall into three categories:

(1) Based on data distortion technology: by interfering with the original data to make sensitive data distorted, and keep some data or data attributes unchanged. The data after interference should maintain two characteristics: first, the original data cannot be re-recognized; second, distorted data can maintain some statistical information and support normal data mining and analysis.

(2) Based on data encryption technology: the original data is further processed in accordance with certain algorithms and laws to transform it into invisible ciphertext, and the receiver needs corresponding decryption means to obtain the ciphertext, thus realizing the privacy protection of the original data.

(3) Based on restricted release technology: restricted release technology is to restrict the sensitive attributes in the released data set to achieve the purpose of privacy protection. The technology achieves a dynamic balance between the risk of privacy disclosure and the sensitivity of data so as to achieve a dynamic balance between data privacy and data quality. The most widespread use of restricted publishing techniques is anonymity.

K-anonymity model can hide the identity or sensitive information of the individual by processing the data in the table. The disadvantage of K-anonymity algorithm is that when the sensitive information in the data table is the same, the attacker is easy to infer the user's privacy information. When the original data set is generalized to increase the data dimension through K-anonymity, the availability of the published data will be reduced, so it is difficult to control the privacy and practicability of the published data [5].
2.2. **Differential Privacy Concept**

Due to the limitation of background knowledge and attack hypothesis, the traditional privacy protection model fails to quantitatively analyze the strength of privacy protection, which affects its application effect to a certain extent. Differential privacy can well solve the above limitations, that is, when a record is arbitrarily added or deleted in the data set, it is impossible to determine whether a particular record exists in the data set by means of calculation and analysis when querying the data set [6]. Differential privacy is defined as follows:

In the interactive model, given any algorithm K, acting on any adjacent data set D and D ’all satisfy ε-difference privacy, its query function is Q, the Range of query result is Range(Q), and the mechanism of query result Q(D) is KQ(D), then the formula is:

\[
\Pr[K_Q(D_1) = r] \leq e^\varepsilon \times \Pr[K_Q(D_2) = r]
\]

(1)

In the interactive model, given any K, an algorithm applied to any adjacent data set D and D meet the epsilon - differential privacy, its output is D, have the formula:

\[
\Pr[K(D_1) = \hat{D}] \leq e^\varepsilon \times \Pr[K(D_2) = \hat{D}]
\]

(2)

The main data object of this scheme is numerical data, so the Laplace mechanism is mainly used to add noise to the data.

Laplace mechanism is used to add random noise subject to Laplace distribution to achieve ε-difference privacy protection [7]. X is a continuous random variable representing the Laplace distribution. The probability density function of this distribution is expressed as follows:

\[
p(x) = \frac{1}{2\lambda} \exp\left(-\frac{|x|}{\lambda}\right)
\]

(3)

Wherein, the noise parameter λ is proportional to the degree of privacy protection and inversely proportional to the noise sensitivity. It can be seen that the smaller the ε is, the larger the noise parameter ε is introduced [8]. Thus, the relation between noise parameters, query sensitivity X and differential privacy coefficient ε can be expressed as follows:

\[
\varepsilon = \frac{x}{\lambda}
\]

(4)

2.3. **K Anonymous Generation Algorithm Based on P2P**

The K anonymous region algorithm based on P2P consists of the following five steps:

(1) Obtain the user's location information, and obtain the longitude and latitude of the user's real location and the grid information of the location through the positioning function.

(2) Users initiate assistance broadcast to inquire neighbor users whose communication distance is 1 hop away from them, and collect grid information of neighbor users who agree to assist into their own set of neighbor users until there are K-1 neighbor users; When users direct communication with your neighbors (i.e., meet 1 jump distance) the number of neighbors with k - 1, requires the user to continue to launch assist radio again, make its direct communication neighbor asked about their immediate neighbours to get k - 1 location information data, the user can set the highest hop, but reached the highest when the hop neighbors set with k - 1, the anonymous failure [9].
(3) Check the anonymous area, will obtain k - 1 neighbor user information grid and its own grid information collection form a k anonymous group, use the anonymous group collection area of the grid position information of an anonymous, if the anonymous area area is less than the given minimum anonymous area, requires the user to launch assist request, will be k - 1 neighbor of users set a random user replacement and regenerate the anonymous group and the corresponding anonymous area until conform to the requirements of the anonymous; anonymity fails if the composition of the anonymous region is greater than the maximum anonymous region area.

(4) Add the centroid points formed by all the grid center points in the noise anonymous region, and add the noise disturbance of Laplace distribution to the generated centroid points to generate the disturbed position.

(5) Send request information. A proxy user is randomly selected from the set of K-1 neighbor users, and the perturbed centroid point and request information are sent through the proxy user; when the LBS server receives the query request, it finds out the corresponding result set according to the location information of the centroid point and the query information, and returns the result set to the agent user. Finally, when the agent user receives the result set, it sends the result set to the user who really requests it [10].

3. Experiment of Privacy Protection Scheme Based on Differential Privacy

3.1. Experimental Environment

This experiment is a simulation experiment. All the mobile data points in this experiment were generated through the road network data generator. Meanwhile, in order to better simulate the spatial geographic position information, the experiment adopted the traffic network map of Oldenburg, on which a total of 10,000 mobile objects were simulated. The map area of the experiment is 16x16km2. The algorithm is implemented by Java language, and the test environment is Windows 10 operating system, CPU Intel Core i5 main frequency is 2.8GHz, and the memory is 8GB.

3.2. Experimental Parameters

The experimental parameters in this paper are set as follows: the default value of anonymity k is 10, and the value range is 1-30; number of users n Default value 10000, value range 1-10000; the minimum anonymous area is 10000m2, and the value range is 10,000-1000000m2; the maximum anonymous area is 1000000m2; default value of fixed mesh side length is 100m, with a value range of 10-100m; noise parameter 0.8, the value range is 0-2.0; the disturbance distance is 200m, and the value range is 0-1000m. The maximum communication distance is 200m.

4. Experimental Result

4.1. Nearness
As shown in Figure 1, when ε = 0.8, it can be seen from the experimental data that the distance between the coordinate points of the perturbation position and the real position is less than 1000m, among which the distance between the coordinate points of the perturbation position and the real position is less than 500m for more than 70%. It can be seen from the experimental data that when the value of ε gradually increases, the generated perturbation location is closer to the real location. It can be seen that the proximity increases as the value of ε increases. However, it makes no sense to use a higher ε value in the real world, because the higher the proximity, the lower the degree of privacy protection.

4.2. Similarity

Similarity represents the ratio of the intersection of the points of interest queried by the perturbed location and the center of mass of the user's anonymous region to the set of points of interest queried by the center of mass of the user's anonymous region. The quality of service provided by the LBS server and the disturbed location have a direct relationship with the location distance of the center of mass of the anonymous region. Therefore, in this experiment, the perturbation radius of the center of mass position from the anonymous region is 200 meters, 400 meters, 600 meters, 800 meters, and 1,000 meters respectively, and the similarity of the queried points of interest is counted.

| Table1. Query similarity at different distances |
|---------|---|---|---|---|---|---|---|
|        | 3 | 6 | 9 | 12 | 15 | 18 | 21 |
| 200m    | 89 | 87 | 86 | 87 | 86 | 88 | 86 |
| 400m    | 75 | 78 | 74 | 79 | 72 | 73 | 77 |
| 600m    | 67 | 63 | 69 | 66 | 64 | 63 | 70 |
| 800m    | 56 | 56 | 53 | 54 | 61 | 58 | 69 |
| 1000m   | 51 | 53 | 52 | 53 | 51 | 51 | 50 |
As shown in Table 1 and Figure 2, it can be seen that the similarity will gradually decrease with the increase of disturbance radius, but the privacy security of users will be higher. Moreover, within the same disturbance range, the similarity of this scheme will not change with the change of anonymity k, and it has high stability.

5. Conclusion

With the development of information technology, the risk of personal privacy disclosure is gradually increasing, and citizens' awareness of privacy protection is getting stronger and stronger. This paper improves the location privacy protection scheme based on third-party trusted server and multi-user collaboration in P2P mode respectively, and proposes a location privacy protection scheme based on user grid and two-level cache and a K anonymous region privacy protection scheme based on differential privacy in P2P mode. Many single k anonymous methods cannot protect location privacy, but perturbation techniques with differential privacy have been proved to be effective against attackers with arbitrary background knowledge. The scheme proposed in this paper aims at the location privacy protection of single point query, but does not consider the privacy protection of users' continuous query. Therefore, how to protect the privacy of users' location route, query content and other information when mobile users conduct continuous query service is the focus of the next step of research.

References
[1] Hong J , Wen T , Guo Q , et al. Privacy Protection and Integrity Verification of Aggregate Queries in Cloud Computing[J]. Cluster Computing, 2019, 22(1):1-11.

[2] Andor A . An Intelligence Perspective on Privacy and Data Protection Risks in Social Media[J]. International Conference Knowledge-Based Organization, 2020, 26(1):151-156.

[3] Ahmet Ali Süzen, Imek M A. A Novel Approach to Machine Learning Application to Protection Privacy Data in Healthcare: Federated Learning[J]. Namk Kemal Tp Dergisi, 2020, 8(1):22-30.

[4] Sun J , Hu S , Nie X , et al. Efficient Ranked Multi-Keyword Retrieval With Privacy Protection for Multiple Data Owners in Cloud Computing[J]. IEEE Systems Journal, 2020, 14(2):1728-1739.

[5] Duan Y , Gao G J , Xiao M J , et al. Cloaking Region Based Passenger Privacy Protection in Ride-Hailing Systems[J]. Journal of Computer Science and Technology, 2020, 35(3):629-646.

[6] Baradaran F , Hamzenjad M . Explaining The Acceptable Privacy Differences of the Habitants in the Iranian Quadruple Climates Vernacular Houses[J]. Human Geography, 2020, 52(3):1031-1053.

[7] Pedic Z. Interconnectivity and Differences of the (Information) Privacy Right and Personal Data Protection Right N the European Union[J]. Review of European and Comparative Law, 2017, 30(3):125-135.

[8] Tifferet S . Gender Differences in Privacy Tendencies on Social Network Sites: A Meta-Analysis[J]. Computers in Human Behavior, 2018, 93(APR.):1-12.

[9] Dargahi T , Ambrosin M , Conti M , et al. ABAKA: A Novel Attribute-Based k-Anonymous Collaborative Solution for LBSs[J]. Computer Communications, 2016, 85(jul.1):1-13.

[10] Zhang Y B , Zhang Q Y , Li Z Y , et al. A k-anonymous Location Privacy Protection Method of Dummy Based on Geographical Semantics[J]. International Journal of Network Security, 2019, 21(6):937-946.