

Big Data Based on T-Closeness Processing Breaks the Information Cocoon

Jiahuan Lin

Shijiazhuang Tiedao University of Network Engineering, Shijiazhuang, Hebei 050000, China

ABSTRACT

Aiming at the phenomenon of information cocoon and the challenge of privacy protection in the era of big data, this paper deeply discusses the role of t-closeness and differential privacy technology in breaking the information cocoon, enhancing data privacy protection and promoting information diversity. As an optimized data anonymization technology, t-closeness prevents privacy leakage, reduces algorithm bias and information homogeneity by limiting the distribution of sensitive attributes in equivalence classes, and provides an effective way to break information cocooning. This paper reviews the current situation of t-closeness technology at home and abroad, especially the fuzzy t-closeness method proposed by Chen Xiaoyu et al., which reduces information loss and improves the intensity of privacy protection through fuzzy processing. As another important privacy protection technology, differential privacy can ensure the privacy of individual data by adding random noise to data query, especially in the aspect of high-dimensional data processing. The LDP algorithm based on union tree proposed by Cheng Siyuan et al., effectively improves the processing efficiency and data quality.

KEYWORDS

T-closeness; Differential privacy; Information cocoon; Privacy protection

1. INTRODUCTION

In the context of the information age, the rapid development of the Internet and social media has greatly promoted the process of information exchange and knowledge sharing, but this process is also accompanied by new challenges in privacy protection. With the wide application of big data algorithms, user data is deeply mined and efficiently utilized by algorithms, which also brings the problem of information cocoon. Information cocoon, as a phenomenon caused by the algorithm recommendation mechanism, traps users in the information closed loop constructed by their interests and preferences, limits the diversified access to information, intensifies cognitive bias, and may even induce trust crisis. Under this background, how to effectively protect personal privacy while ensuring data utilization efficiency and fairness of recommendation system has become an important issue to be solved urgently.

T-closeness technology, as an advanced data anonymity technology, is a deep optimization of traditional k-anonymity technology. Its core is to ensure that the distribution of data sensitive attributes in each equivalence class is close to the full set distribution, so as to reduce the potential threat of statistical analysis to personal privacy information disclosure. This technology not only significantly enhances the protection effect of private data, but also effectively alleviates the phenomenon of information cocoon. By realizing the diversification of information recommendation, it effectively broadens the user's vision and improves the transparency of data.

In China, with the promulgation and implementation of relevant laws and regulations such as the Personal Information Protection Law, t-closeness anonymization technology has attracted wide attention from all walks of life. Current research focuses on how to improve the efficiency of data use and the fairness of the recommendation system. For example, the fuzzy t-closeness privacy protection method proposed by Chen Xiaoyu et al can not only reduce the loss of effective information, but also significantly enhance the intensity [1] of privacy protection. Zhang Chenyan et al. improve the existing network security architecture by making t-closeness desensitization design of medical data, dividing the data permission boundaries, and combining with the network access security boundary technology, and propose a mechanism to enhance the desensitization and sharing control of medical archive data.

Internationally, the theoretical research and practical application of t-closeness are becoming more and more mature. Especially in Europe and the United States, the strict privacy protection legal framework, such as GDPR, provides a solid legal support for the wide application of this technology.

Therefore, this paper will focus on how to make use of t-closeness privacy protection technology and differential privacy protection technology to process big data so that the valid data can be retained as much as possible, and deeply explore its application potential in cracking the information cocoon. This paper will comprehensively balance the relationship between privacy protection and data availability, and discuss the optimization strategies of t-closeness technology in detail by means of case studies and comparative analysis, aiming at providing more privacy-friendly and information-rich recommendation strategies for big data algorithm design. In addition, this paper will also pay attention to the economic feasibility of the technology and user acceptance, to ensure that the research results meet the legal requirements and have a wide range of practicability. In summary, this paper aims to provide a comprehensive solution to solve the information cocoon problem that takes into account both privacy protection and information diversity, and makes assumptions about the development of information processing and recommendation systems towards higher standards of privacy protection technology.

2. OVERVIEW OF T-CLOSENESS TECHNOLOGY

2.1. Basic Concepts and Definitions

T-closeness is a data anonymization technique that plays an important role in enhancing privacy protections, especially when working with large data sets containing sensitive information. This technology is optimized on the basis of k-anonymity and l-diversity, hoping to solve the shortcomings of k-anonymity and l-diversity in privacy protection. In order to protect the privacy of individuals, K-anonymity requires that each equivalence class in the data set contain at least k records. While l-diversity technology improves the level of privacy protection by increasing the diversity of sensitive attributes.

T-closeness goes a step further by requiring that the distribution of sensitive attributes in each equivalence class is close to the distribution of sensitive attributes in the entire dataset, and the difference does not exceed the threshold t [2]. In this way, even if an attacker has access to information about the equivalence class, it is difficult for them to infer the sensitive attributes of a particular individual. This method effectively prevents attackers from inferencing information about individuals through statistical analysis, thus improving the security of data.

After the data is released, the technique can ensure that it is difficult for the attacker to statistically infer the sensitive attributes of a specific individual, and effectively combat link attacks and homogenous attacks. In general, t-closeness is an efficient data anonymization technique. By ensuring that the distribution of sensitive attributes is close to the whole data set, t-closeness improves the level of data privacy protection and effectively prevents privacy leakage [3].

2.2. Case Study - Fuzzy T-Closeness Privacy Protection Method

In view of the loss of effective information caused by the generalization mechanism of the t-closeness method, scholars such as Chen Xiaoyu innovatively propose the fuzzy t-closeness privacy protection method. This method uses the fuzzy theory to design a set of data blurring process instead of the traditional generalization process, which effectively reduces the loss of effective information for users and enhances the intensity of privacy data protection. The method divides equivalence classes by fuzzy clustering, calculates the membership degree of elements, and obtains the alternative value of fuzzy elements based on these membership degrees. For classification attributes, the fuzzy semantic structure tree is also constructed to deal with sensitive information more flexibly, so that even the same attribute value may correspond to different fuzzy values in different records, which greatly increases the difficulty of the attacker to infer the original information. The experimental results show that the fuzzy t-closeness method can not only improve the validity retention of data, but also effectively prevent homogeneity and background knowledge attacks, which indicates that the method has high practical value and privacy protection effect in practical applications.

3. DIFFERENTIAL PRIVACY PROTECTION TECHNOLOGY

3.1. Theoretical Basis

Differential Privacy (DP) is a strictly defined privacy protection framework, the core of which is to ensure that the statistical query results do not vary significantly due to the participation of specific individuals in the process of data release [4]. This concept was originally proposed by Cynthia Dwork and others to provide a solid mathematical privacy guarantee for data publication [5]. The basic principle of differential privacy is to incorporate random noise into the query results, so that an attacker can have all the data except one record, but still be unable to determine the existence of the record and its specific value. Privacy budget (ϵ) is a key parameter in differential privacy, which is used to accurately balance the strength of privacy protection with the accuracy of queries. The smaller the ϵ value, the more significant the privacy protection effect, but the utility of the data decreases accordingly.

3.2. Technological Progress

The application field of Differential Privacy has gradually evolved from centralized differential privacy (CDP) to Local Differential Privacy (LDP) [6]. [6]LDP technology allows users to randomize data locally, and then send the processed data to the central server, which effectively avoids the risk of untrustworthy central server. However, as the data dimensions increase, LDP technology also faces the challenge of decreasing efficiency while ensuring privacy, especially when processing high-dimensional data.

In order to solve the privacy protection problem of high-dimensional data, Cheng Siyuan et al. proposed a high-dimensional data localization differential privacy protection algorithm (JT-LDP) based on Junction Tree. The algorithm uses the random response technique to protect the privacy of the data at the client side, and then uses the junction tree model to reduce the dimension of the data at the server side, and divides the high-dimensional data attribute set into multiple low-dimensional attribute sets. This process reduces the computational complexity, while preserving the attribute dimension relationship of the original data, so that the synthetic data set and the original data set maintain highly similar features. By constructing the improved algorithm of union tree, the processing efficiency is further improved, and an efficient and reliable solution is provided for the privacy protection of high dimensional data.

3.3. Application Scenario Analysis

At present, the application of privacy protection technology in the field of big data is becoming more and more common, which significantly alleviates the phenomenon of information cocooning. Taking a certain health information exchange system as an example, the combination of t-closeness and differential privacy protection technology ensures the privacy and security of patient medical record data, and also allows researchers to conduct in-depth analysis on the premise of data authenticity, which not only protects personal privacy, but also realizes the effective use of data.

In the social network, e-commerce and other scenarios, privacy protection technology also plays a key role. With LDP technology, users can enjoy personalized recommendation services and reduce the cognitive limitations caused by over-customization of algorithms, thus broadening the vision of information acquisition and effectively weakening the information cocoon effect.

Privacy protection technology, such as differential privacy technology, with its high computational efficiency and outstanding flexibility, such as LDP and JT-LDP algorithm, not only set an example for data privacy protection technology on the theoretical basis, but also promote the open sharing of data in practice. It provides strong technical support for solving the problem of information cocoon, and lays a solid foundation for realizing a more transparent and secure data environment. With the continuous progress of technology, privacy protection technology will play an important role in more fields in the future, and continue to contribute to building an ideal data environment.

4. INNOVATION AND IMPROVEMENT

4.1. Limitations of the Integration of T-Closeness Technology and Differential Privacy Technology

As for the combination of t-closeness technology [7] and differential privacy technology to break the information cocoon phenomenon, both technologies show their unique advantages in data processing. t-closeness technology is good at maintaining the global distribution consistency of big data to protect sensitive information. The differential privacy technology ensures the privacy of statistical results by introducing random noise. The core challenge of this method is how to maintain the consistency of the data distribution, and how to strengthen the privacy protection through the random noise of differential privacy, and precisely control the degree of data distortion.

4.2. Innovation Strategy

After the equivalence class division of t-closeness technology, differential privacy protection is implemented within each equivalence class. By adjusting the noise level, the similarity of sensitive attribute distribution among equivalence classes is ensured to achieve harmony and unity with the protection intensity of differential privacy, so as to achieve the purpose of blending the two.

In terms of technology, we need to learn from the LDP algorithm based on joint tree proposed by Cheng Siyuan, and study how to use the joint tree structure to accurately identify the attribute correlation of data in the framework of t-closeness. In the process of data dimensionality reduction, local differential privacy protection is introduced to realize the dual goals of data privacy protection and data structure optimization.

A hybrid privacy protection model is assumed in this paper. This model allows us to balance the balance between privacy protection and data utility [8] quickly and flexibly adjust the weighting ratio between t-closeness technology and differential privacy technology according to the sensitivity of data and specific application scenarios before data release. This flexible way can meet the requirements of different scenarios and security requirements, so as to break the phenomenon of information cocoon more effectively and protect data privacy information.

Aiming at the computational complexity problem, this paper seeks a solution by exploring efficient data structure and algorithm design. For example, index technology is used to accelerate the calculation of equivalence class division and sensitive attribute distribution, and a distributed computing framework is developed to accelerate t-closeness's data processing speed.

Optimize the existing privacy protection model and integrate the t-closeness variant of fuzzy theory to reduce the amount of information loss. For example, the fuzzy t-closeness method proposed by Chen Xiaoyu et al. For differential privacy, a mechanism can be developed to dynamically adjust the privacy budget to meet the needs of different query sensitivities and data dimensions. When combining the two, we can refer to Domingo Ferrer J's proposal to obtain stochastic t-closeness by k-anonymizing the quasi-identifiers of data sets and ensuring e-differential privacy for sensitive attributes. In this case, the T-value is a function of ϵ , the size of the data set, and the size of the equivalence class [9].

Finally, based on the LDP algorithm based on union tree proposed by Cheng Siyuan and other scholars, combined with other scholars' privacy protection model ideas, machine learning and natural language processing technology, the attribute correlation identification and data blurring process are optimized. While realizing t-closeness principle, the definition of ϵ -difference privacy is combined. Effectively limit the knowledge gain gained by intruders, achieve the purpose of protecting data privacy [10] and ensure the maximum retention of data validity, so as to break the phenomenon of information cocoon.

5. CONCLUSION

As an optimized data anonymization method, t-closeness technology provides a strong support for privacy protection in big data environment. This technology is further upgraded on the basis of k-anonymity and l-diversity. By limiting the distribution of sensitive attributes in equivalence classes and the differences between all data sets, it effectively prevents the risk of personal privacy disclosure in the analysis process, especially in preventing link attacks and homogeneity attacks. The implementation of t-closeness technology ensures that sensitive personal information can maintain a high degree of concealment even after the data is released, laying a solid foundation for breaking the information cocoon phenomenon.

On the other hand, differential privacy technology has become a core technology in the field of privacy protection due to its strict protection of individual privacy in the process of data release. Its principle is to ensure that any legitimate query to the data set, the result of the change will not fluctuate significantly due to the addition or deletion of a single data record, thus protecting privacy while maintaining the statistical utility of the data. In particular, driven by laws and regulations such as GDPR, differential privacy has been widely used in the field of big data in Europe and the United States.

The combination of t-closeness technology and differential privacy technology opens up a new perspective for breaking the problem of information cocooning. Information cocoon phenomenon, that is, users are trapped in the same information due to algorithm recommendation and preference limitation, resulting in limited cognition. t-closeness technology is used to strictly control the equivalence class distribution of large data sets. Combined with differential privacy technology, noise is introduced in data query to protect individual privacy, so as to build a new data processing mechanism. This mechanism can not only effectively maintain the user's personal privacy information, but also ensure that the validity of the user data can be preserved to the maximum extent, and help the user to break the phenomenon of information cocoon and contact more extensive and diversified information resources. This fusion strategy provides users with a safe and broad information environment, ensuring the fair dissemination of information and the reasonable protection of user privacy data.

In the future, with the continuous development of privacy protection technology, it is reasonable to believe that the integrated application of t-closeness technology and differential privacy technology will play a more significant role in the fields related to cracking the information cocoon phenomenon.

REFERENCES

- [1] Chen Xiaoyu, Han Bin, Huang Shucheng, et al. Journal of Computer Application and Software, 2018, 35(09):317-322+333.
- [2] Wu S H. Research on t-closeness Privacy Protection Model supported by rough sets and clustering [D]. Shanxi Normal University, 2015.
- [3] Gong Qiyuan. Research on Data anonymity technology for Data publishing [D]. Southeast University, 2016.
- [4] [Lu Guoqing, Zhang Xiaojian, Ding Liping, et al.] A Frequent Sequence Pattern Mining Method based on Differential Privacy [J]. Journal of Computer Research and Development, 2015, 52(12):2789-2801.
- [5] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity: Privacy-preserving data publishing [J]. ACM Transactions on Database Systems, 2006, 31(3): 1-41.
- [6] Niu Ben, Li Fenghua, Chen Yahong, et al. Privacy protection method, device and electronic equipment for image recognition [P]. Beijing: CN202010346054.9, 2020-08-25.
- [7] N. Li, T. Li and S. Venkatasubramanian, "t-closeness: Privacy Beyond k-Anonymity and l-Diversity," 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 2007, pp. 106-115, doi: 10.1109 / ICDE. 2007.367856.
- [8] Gangarde R, Sharma A, Pawar A. Enhanced Clustering Based OSN Privacy Preservation to Ensure k-Anonymity, t-closeness, l-Diversity, and Balanced Privacy Utility [J]. Computers, Materials & Continua, 2023, 75(1):2171-2190.
- [9] Domingo-Ferrer J, Soria-Comas J. From t -closeness to differential privacy and vice versa in data anonymization [J]. Knowledge-Based Systems, 2015, 74:151-158.
- [10] Gao Ling, Jiang Yang, Ren Zhe, et al. A t-closeness privacy Protection Method satisfying ϵ -differential privacy [P]. Shaanxi Province: CN201910875809.1, 2023-07-11.