Design of Privacy-Preserving Personalized Recommender System Based on Federated Learning

Yikan Wang[1*], Chenwei Gong[2], Qiming Xu[3], Yingqiao Zheng[4]

[1]School of Systems and Enterprises, Stevens Institute of Technology, Babbio Center, 525 River St, Hoboken, NJ 07030, United States

[2]Henry Samueli School of Engineering, Department of Computer Science, University of California, Los Angeles, CA 90095, United States

[3]Khoury College of Computer Sciences, Northeastern University, West Village Residence Complex H, 440 Huntington Ave, Boston, MA 02115, United States

[4]College of Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States

[*]ywang463@Stevens.edu

Keywords: federated learning, privacy protection, personalized recommender systems, distributed training, data security

Abstract: The protection of user data privacy has become a key challenge in data-driven personalized recommendation systems. Traditional centralized recommendation methods often require uploading user data to a server for centralized training, which has a high risk of data leakage. For this reason, this paper designs a privacy-preserving personalized recommendation system based on federated learning. In this system, the user data is kept in the local device, and the distributed training of the model is realized by the federated learning algorithm, so as to complete the recommendation task under the premise of guaranteeing the user's privacy. This paper analyzes the privacy-preserving mechanism of federated learning and proposes a system architecture design for personalized recommendation; the performance of the system is verified based on different experimental scenarios, and the accuracy, response speed and privacy-preserving effect of the model are evaluated. The experimental results show that the system effectively reduces the risk of user data exposure while improving the recommendation effect, providing a new solution for building a safe and trustworthy personalized recommendation system.

1. Introduction

With the rapid development of the Internet and the diversification of user needs, personalized recommendation systems have been widely used in e-commerce, social media, online education and other fields to provide users with accurate content recommendations and personalized services[1]. However, traditional personalized recommendation systems often rely on centralized data processing, uploading user data to the server for centralized training to optimize the recommendation model[2]. While this model improves the recommendation effect, it also brings serious privacy issues, as users' personal information and behavioral data are highly susceptible to the risk of leakage or misuse during centralized storage and transmission[3]. Therefore, how to achieve personalized recommendation under the premise of ensuring user privacy has become an important problem to be solved by both academia and the industry.

Federated Learning, a new distributed machine learning technique, provides an effective privacy-preserving solution[4]. Through Federated Learning, user data is retained on local devices, while the training process of the model is accomplished through local computation and encrypted parameter aggregation, avoiding the direct upload of data. This approach not only improves data security, but also solves the problem of data silos to a certain extent, enabling data from different sources to be trained collaboratively, thus further improving the accuracy and personalization of the

  DOI: 10.25236/iwmecs.2024.001

recommendation system.

The research of this paper focuses on the design and implementation of privacy-preserving personalized recommender system based on federated learning[5]. The advantages of federated learning in privacy protection are analyzed, and the system architecture and model training process applicable to the recommender system are proposed; the detailed steps of data processing, model training and optimization are designed; the performance of the system in different scenarios is verified through experiments, focusing on evaluating its effectiveness in privacy protection, recommendation accuracy and response speed[6]. The research in this paper will provide reference and technical support for building an efficient and secure personalized recommendation system.

2. Foundations of Federated Learning for Personalized Recommendations

Federated learning is an emerging distributed machine learning technique that aims to enable joint training of models from multiple participants while keeping the data local, thus effectively protecting user privacy[7]. In this framework, the training task of the model is assigned to each participant (e.g., user devices), model updates are generated through local computation, and only the model parameters are uploaded to the server for aggregation so as to continuously optimize the global model without the need to share the raw data[8]. This training model gives Federated Learning a natural advantage in privacy protection and provides a secure solution for various data-sensitive application scenarios. Federated Averaging Formula:

$$w^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_k^t \tag{1}$$

Privacy protection is particularly important in personalized recommendation systems, which involve a large amount of users' personal information and behavioral data, including browsing history, purchase records, and preference data[9]. Traditional recommendation methods usually require centralized storage and processing of these data, which increases the risk of user privacy leakage[10]. Federated learning, on the other hand, avoids data transfer by training the model locally on the user's device, which effectively protects the user's privacy. In addition, federated learning reduces the burden of centralized data management through a decentralized model, bringing a safer and more efficient implementation of personalized recommendation systems,showed in Figure 1 :

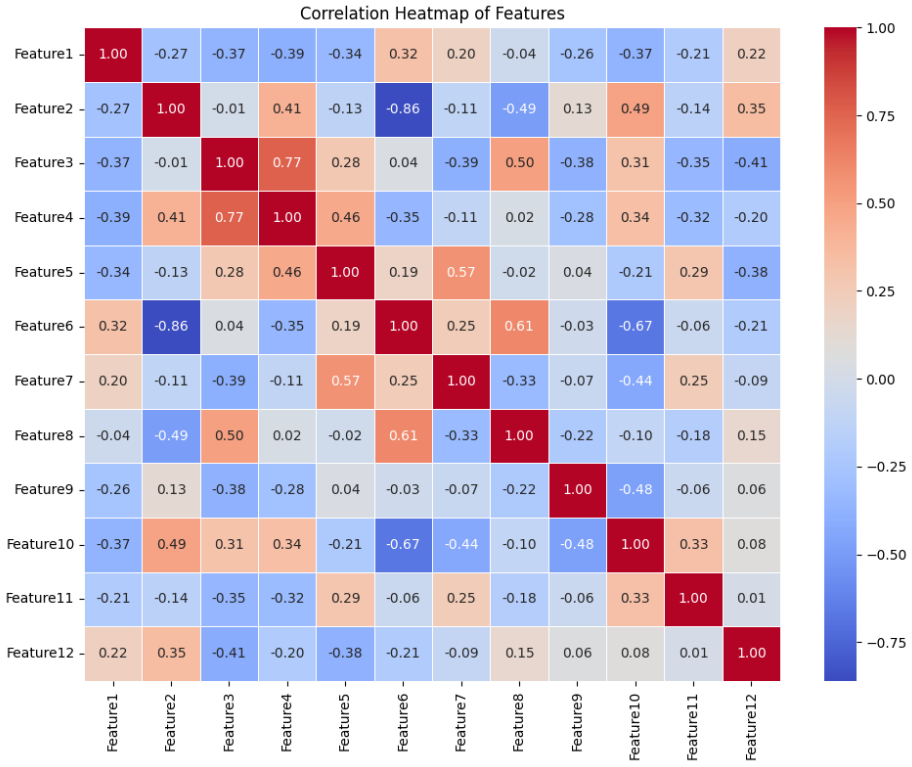

Figure 1 Correlation Heatmap of Features

Federated learning has a wide range of application scenarios in personalized recommendation, especially for e-commerce, social media, and news push. In these applications, federated learning allows recommender systems to train models on different user devices to capture individual behavioral characteristics and generate personalized recommendations. For example, in e-commerce platforms, the system is able to recommend appropriate products for users by analyzing their shopping preferences without uploading their personal data centrally to the server, ensuring the privacy and security of user data. Since the computing power and network environment of each user's device are different, how to efficiently perform model training locally is a difficult problem; there are significant differences in the data distribution of different users, and this Non-Independent Identical Distribution (Non-IID) characteristic of the data may lead to difficulties in convergence of the global model, which affects the recommendation effect. In the process of model parameter aggregation, how to balance recommendation accuracy and privacy protection is also an important direction in federal learning research.

3. Design of personalized recommendation system based on federated learning

In order to realize the design of personalized recommendation system based on federated learning, we discuss the design of a reasonable system architecture from three aspects: the system architecture, the model training process, and the privacy protection mechanism to ensure that each user device can efficiently participate in the model training; the model training process is formulated to guarantee that the model can be continuously optimized and updated while the data is kept locally; and through the privacy protection mechanism, to further enhance the security of the system to ensure that user data is not leaked during the whole recommendation process.

3.1. System Architecture Design

In a personalized recommendation system based on federated learning, the core of the system architecture design is how to achieve effective training of models and parameter aggregation while user data remains local. The system consists of multiple user devices and a central server, and each user device contains a local model for personalized local training without uploading raw data. The server is responsible for coordinating the training process among user devices and aggregating the model parameters uploaded by each device to generate the global model. The architecture enables access to shared recommendation models while ensuring user privacy. Privacy Protection with Differential Privacy (Noise Addition):

$$w_k^t \leftarrow w_k^t + \mathcal{N}(0, \sigma^{2I}) \tag{2}$$

On the user's device, the local training process is optimized based on the user's historical data Each device calculates the model gradient or parameter updates based on the local data and completes one or more iterations locally to generate the updated model. Since these computations are performed on the user's device, the data is always kept locally, thus ensuring user privacy. At the end of each round of federated learning iterations, each device uploads the model parameter updates obtained from local training to the server without transmitting the actual data content, further reducing the risk of privacy leakage.

The main task of the server side is to receive the parameter updates uploaded by each device and perform aggregation to generate the global model. Commonly used aggregation methods include Federated Averaging, in which the parameter updates from each user device are weighted and averaged to obtain an update of the global model. The server sends the updated global model to each device to complete an iterative round of federated learning. This architectural design can ensure that the recommendation model gradually becomes stable after multiple rounds of training, while taking into account the improvement of personalization and recommendation effect. The system architecture also needs to consider issues such as communication optimization and device management to improve the efficiency and reliability of the system. Since federated learning involves the transmission of a large number of parameters, communication overhead is an important factor, so bandwidth consumption can be reduced by compressing the transmitted data and reducing model complexity.

The system needs to be able to dynamically manage the participation of user devices and consider the processing strategy when the devices are offline or withdrawn to ensure the robustness of the system and the continuity of the recommendation service.

3.2. Model Training Process

In federated learning-based personalized recommendation systems, the model training process aims to achieve efficient model optimization in distributed environments to meet personalized recommendation requirements and protect user privacy. The training process starts with server-side initialization. The server generates an initial version of the global model and distributes it to each user device. This initial model can be a randomly initialized model or a pre-trained model from a small public dataset. With a unified initial model, the system ensures that all devices are at the same starting point for subsequent local training.

Each round of federated learning iteration consists of a local training process for the user device. After receiving the global model from the server, user devices train the model based on locally stored data. Each device performs several local gradient updates to the model using its unique user data to adapt to individual preferences. Since the training data is retained locally, all model optimization processes do not involve data uploading, which ensures the effectiveness of personalized recommendations while effectively protecting user privacy,showed in Figure 2 :

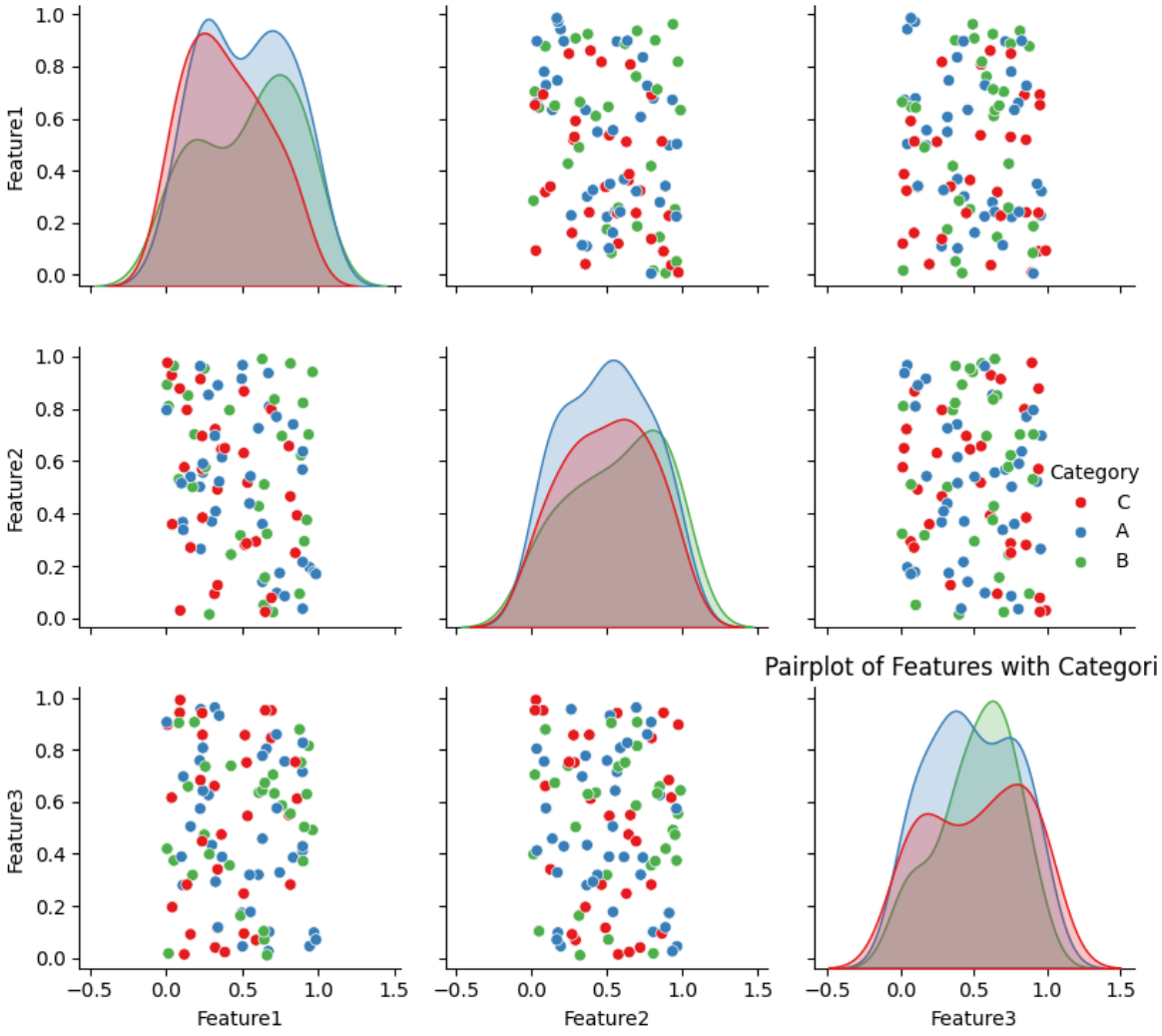

Figure 2 Pairplot of Features with Categories

After the local training is completed, each user device passes the model parameter updates from the training back to the server instead of passing back the raw data. The server side receives the parameter updates from multiple devices and performs aggregation to generate a new global model. The aggregation method usually uses Federated Averaging, which is a weighted average of the

parameter updates from each device to ensure that the impact of different devices on the global model reasonably reflects the amount of their data or training effect. Through this aggregation step, the server is able to integrate the local learning results of individual users into the global model, making it more generalizable. After the server-side aggregation is completed, the new global model is again sent down to the user devices to start the next round of federated training iterations. The process is repeated until the performance of the global model reaches the preset criteria or meets the convergence conditions. Through multiple rounds of iterations, the model gradually improves the recommendation effect while protecting privacy and achieves more accurate personalized recommendation.

3.3. Privacy protection mechanisms

In federated learning-based personalized recommender systems, the design of privacy-preserving mechanisms is crucial to ensure that user data is adequately protected throughout the training and recommendation process. The underlying privacy-preserving mechanism of federated learning relies on localized computation, i.e., user data is always kept on the local device and only processed model parameters are shared. With this distributed computing approach, the system avoids the risk of data leakage in traditional centralized approaches and provides higher privacy protection for users.

To further enhance the security during data transmission, federated learning is often combined with encryption techniques such as Secure Multi-Party Computation (SMCC) or Homomorphic Encryption (HEC). By encrypting the model parameters during the parameter upload process, the server does not have direct access to the user's update details, thus avoiding the potential risk of privacy leakage. Encryption ensures data security during the transmission of parameters from each device, so that even if the transmitted data is intercepted, it is difficult to obtain meaningful information from it.

Another commonly used privacy protection mechanism is Differential Privacy. Differential privacy adds noise to the parameter updates generated by the user's device so that the server cannot infer the specific behavior or data characteristics of individual users when aggregating these updates. By reasonably setting the size of the noise, differential privacy not only protects the privacy of user data, but also ensures that the recommendation effect of the global model will not be significantly degraded by noise interference. The differential privacy mechanism provides a stronger privacy guarantee for the application of federated learning in personalized recommendation, which is especially suitable for highly privacy-sensitive scenarios.

Privacy-preserving mechanisms also need to be integrated with device management policies to enhance the robustness of the system. Due to the varying computing power and network conditions of user devices, the system needs to design an effective participation strategy so that only device updates that satisfy specific privacy criteria can participate in the global model aggregation. This strategy not only protects data privacy, but also improves the quality of model training, thus achieving the dual optimization of privacy protection and recommendation effectiveness.

4. Experimental validation and result analysis

In personalized recommendation systems based on federated learning, experimental validation is a key link to test the effectiveness of the design and the performance of the system. In this experiment, a typical personalized recommendation dataset is selected to simulate the collaborative training process of multiple user devices in different network environments. The main purpose of the experiment is to evaluate the performance of the system in terms of privacy protection, training efficiency, and personalized recommendation effect. Several different scenarios are set up in the experiments to verify the feasibility and advantages of federated learning in practical applications. All user devices use only their respective behavioral data during local training to ensure that data privacy is fully protected.

The experimental results show that the recommendation system based on federated learning can effectively improve the accuracy of personalized recommendation. Compared with the traditional centralized recommender system, federated learning not only provides similar recommendation

accuracy, but also can deal with a larger user group under the premise of ensuring data privacy. With the collaboration of different user devices, the system gradually converges and optimizes the global model, and the effect of personalized recommendation improves round by round. In addition, the experiments also found that the training efficiency of the system is significantly improved and the accuracy of the recommendation results is also greatly improved when the computing power of the user devices is high.

In terms of privacy protection, the experiment verifies that the recommender system after adopting differential privacy and encryption can effectively prevent the leakage of user data. In the case of adding noise, the recommendation accuracy of the model only slightly decreases, indicating that the differential privacy mechanism has less impact on the recommendation effect while ensuring data privacy. The encryption technology ensures the security of user model update, and even if there is a security problem during data transmission, it can still effectively protect user privacy. The experimental results demonstrate that federated learning can maintain efficient personalized recommendation performance while guaranteeing user data privacy.

The experiments also analyzed the performance of federated learning under different device engagement and data distribution. In the case of low device participation or uneven data distribution, the stability of the system training is reduced, resulting in large fluctuations in the recommendation effect. In order to improve the robustness and stability of the system, future research should focus on optimizing the device management strategy and model aggregation method. For the hardware differences of different devices, how to improve the training efficiency and recommendation accuracy while ensuring privacy protection is still a problem worth exploring in depth.

5. Conclusion

The design of privacy-preserving personalized recommendation system based on federated learning can effectively improve the accuracy of personalized recommendation and the overall performance of the system while ensuring the privacy of user data. Through localized training and the application of encryption technology, the system not only protects user data, but also enhances the layer of privacy protection through mechanisms such as differential privacy to ensure that users do not disclose personal information while enjoying personalized recommendation services. The experimental results show that federated learning is able to optimize the global recommendation model in a collaborative multi-device environment and gradually improve the recommendation effect while maintaining the efficiency of the training process.

The experiments also found that the stability and robustness of the system decreased in the case of low device participation or uneven data distribution. This indicates that there is still room for improvement in the future in terms of device management, model aggregation, and optimized training strategies, especially how to deal with hardware differences between devices and how to improve recommendation accuracy and training efficiency. The personalized recommendation system based on federated learning has great application prospects, and in the future, with the further development and optimization of the technology, the privacy protection ability and recommendation accuracy of the system will be further improved, which will promote the wide application of personalized recommendation services in various industries.

References

[1] Ma X , Li H , Ma J ,et al.APPLET: a privacy-preserving framework for location-aware recommender system[J].Science China Information Sciences, 2017, 60(9):1-16. DOI:10.1007/s11432-015-0981-4.

[2] Yao Y , Liu J .On Privacy-preserving Context-aware Recommender System[J].International Journal of Hybrid Information Technology, 2015, 8(10):27-40.DOI:10.14257/ijhit.2015.8.10.04.

[3] Zhu T , Li G , Ren Y ,et al.Privacy preserving data release for tagging recommender systems[J]. Web Intelligence & Agent Systems, 2015, 13(4):229-246.DOI:10.3233/WEB-150323.

[4] Ma X , Zhang H , Zeng J ,et al.FedKGRec: privacy-preserving federated knowledge graph aware recommender system[J].Applied Intelligence, 2024, 54(19):9028-9044.DOI:10.1007/s10489-024-05634-4.

[5] Ameur E , Brassard G ,José M. Fernandez,et al.Alambic: a privacy-preserving recommender system for electronic commerce[J].International Journal of Information Security, 2008, 7(5):307-334. DOI:10.1007/s10207-007-0049-3.

[6] Ravi L , Subramaniyaswamy V , Devarajan M ,et al.SECRECSY: A Secure Framework for Enhanced Privacy-Preserving Location Recommendations in Cloud Environment[J].Wireless Personal Communications, 2019, 108(3):1869-1907.DOI:10.1007/s11277-019-06500-0.

[7] Badis L , Amad M , Assani D ,et al.P2PCF: A collaborative filtering based recommender system for peer to peer social networks[J].Journal of High Speed Networks, 2021, 27(5):1-19. DOI:10.3233/JHS-210649.

[8] W. Nejdl and P. Brusilovsky, "Editorial" in IEEE Transactions on Learning Technologies, vol. 2, no. 04, pp. 259, October-December 2009, doi: 10.1109/TLT.2009.52.

[9] Boutet A , Frey D , Guerraoui R ,et al.Privacy-Preserving Distributed Collaborative Filtering[J]. Computing, 2016, 98(8):827-846.DOI:10.1007/s00607-015-0451-z.

[10] Pu P , Chen L , Hu R .Evaluating recommender systems from the user's perspective: survey of the state of the art[J].User Modeling and User-Adapted Interaction, 2012, 22(4-5):317-355. DOI:10.1007/s11257-011-9115-7.