Theory and Practice of Science and Technology
2025, VOL. 6, NO. 8, 147-152
DOI: 10.47297/taposatWSP2633-456927.20250608

A Review of Triangular Core Problem of Recommendation System Based on Federated Learning

Xiao Liu

Hainan University, Haikou, Hainan, 570228, China

ABSTRACT
Traditional recommender systems face severe challenges of data privacy protection. Centralized data storage and processing mode can not meet the compliance requirements of General Data Protection Regulation (GDPR), and there is a significant risk of privacy leakage. Federated learning provides a new solution for recommender systems through the distributed paradigm of "data does not move, model moves", and federated recommender systems emerge. It is found that the federal recommendation system still faces many challenges in heterogeneity, privacy protection security and system efficiency, and the existing research has not fully revealed the internal relationship between these problems. This paper innovatively constructs a "privacy-effectiveness-efficiency" triangular equilibrium analysis framework, uses systematic literature analysis methods, focuses on analyzing the core problems and internal relations of privacy, effectiveness and efficiency of federated recommendation systems, systematically summarizes the progress and shortcomings of current research, and proposes optimization directions, providing theoretical references for subsequent research.

KEYWORDS
Federated learning; Recommender system; Triangular equilibrium; Heterogeneity; Privacy security; System efficiency

1 Introduction

Traditional recommender systems provide personalized recommendation services for users by analyzing their historical behaviors and personal characteristics, but the centralized storage of user data in the server for model training not only faces increasingly strict data protection compliance challenges, but also has a significant risk of user privacy disclosure. The combination of federated learning and recommendation system enables the federated recommendation algorithm to keep the user data in the local client all the time during the model building process, thus protecting the user privacy to a certain extent. In the aspect of privacy protection, although gradient encryption and differential privacy provide basic guarantee, the problems of reverse inference attack on model parameters and cross-domain privacy disclosure still need to be further solved; In the aspect of recommendation effect, Non-IID (Non-IID) caused by data island makes the efficiency of model aggregation decrease, and the contradiction between cross-domain knowledge fusion and personalized recommendation needs to be reconciled; Based on this framework, this paper analyzes the latest progress of federated recommendation from three key perspectives, which not only summarizes the latest progress of federated recommendation in the "privacy-effectiveness-efficiency" triangular balance, but also provides a new research idea to break through the existing technical bottleneck. Finally, based on this analytical framework, this paper points out the development directions of future research in dynamic adaptability, fine-grained privacy protection and decentralized incentive.

2 Recommendation system based on federated learning

2.1 Federated Learning

Federated Learning (FL) is a distributed machine learning framework proposed byGoogle in 2016, which allows multiple participants to jointly train shared models without sharing original data, thus protecting data privacy. Its core feature is that participants have absolute control over local data, and the server needs to meet different privacy protection requirements. Federated learning can be classified from four perspectives: model architecture, federated methods, optimization policies and privacy techniques, among which model federalization is the focus of research. Architecturally, it can be divided into client-server and decentralized: the former updates the global model through server-side aggregation parameters, but there is a single point of failure risk; the latter achieves anonymity and high availability through direct communication between clients. Privacy preserving techniques such as Homomorphic Encryption (HE), Differential Privacy (DP) and Secure Multiparty Computing (SMPC) are widely used, e. g. HE protects parameter transmission and DP adds noise to prevent data inference. Optimization methods include model compression, communication strategy improvement, incentive mechanism and client sampling to solve the problems of resource constraint, high communication cost and low participation rate.

2.2 Federated recommendation system

Federated Recommendation System can improve the prediction performance of recommendation models while protecting user privacy data, which is an important application scenario of federated learning. Federated Recommendation System can be divided into two categories according to architecture design and data distribution characteristics. In the system architecture dimension, there are mainly two models: client-server architecture and decentralized architecture: The former coordinates global model updating through central server, while the latter realizes direct interaction and collaborative training among nodes. In the data distribution dimension, it can be divided into three paradigms: horizontal federated recommendation is applicable to scenarios with the same feature space but different user samples; vertical federated recommendation deals with situations where feature spaces are different but user groups overlap; and migration federated recommendation is for cross-domain scenarios where users and feature spaces do not overlap. These architectures and paradigms all adopt the basic mechanism of parameter or gradient exchange instead of original data sharing, and optimize the recommendation model while protecting data privacy through different collaborative training methods. Current research mainly focuses on horizontal and vertical federated recommendation, while migration federated recommendation is still in the exploration stage, and the choice of architecture depends on the requirements of centralized coordination and distributed autonomy for specific application scenarios.

2.2.1 Communication architecture and training process

In federated recommendation systems, client data is stored locally, and intermediate parameters are communicated between the server and client. There are two main communication architectures used in federated recommendation system research, including client-server architecture and decentralized architecture. The client-server architecture training process includes: 1) the central server screens clients participating in this round of training and distributes the latest global recommendation model to them; 2) Each client uses local private data to train and optimize the model based on the received model; 3) The client uses the trained model parameters 4) The server aggregates the collected client parameters (such as weighted average) to generate a new generation global model. Repeat the above steps until the model performance converges or reaches a preset termination condition. The decentralized architecture training process includes: 1) The server generates global model parameters and distributes them to all participating nodes;2) Each node calculates model parameter updates based on private data; 3) Each node randomly selects some online neighbor nodes and broadcasts their parameter updates through the P2P network; 4) The node aggregates the received neighbor parameters with the local model in real time. Repeat steps 2, 3, and 4 until the model performance converges or reaches the preset termination condition. Figure 1 compares the two architectures in detail in multiple dimensions:

Comparison Dimension	Client - Server Architecture	Decentralized Architecture
Privacy Protection Intensity	Medium: Relies on the server not to disclose gradient/parameter information [3]	Extremely High: Raw data and intermediate parameters never leave local nodes
Fault Tolerance Capability	Low: A single - point failure of the server will cause system collapse [2]	High: Has robust distributed fault - tolerance capability
Anti - Attack Capability	Needs to prevent server - side backdoor penetration and gradient inference attacks [3]	Naturally resists target - oriented attacks
Communication Overhead	Low: Only needs communication between client and server	High: Requires full - connectivity or random multi - hop communication between nodes [4]
Computation Overhead	Concentrated on the server side	Decentralized: Each node needs to perform aggregation computation [4]

Figure 1 Client-server architecture and decentralized architecture comparison

2.2.2 Classification

The Federal Recommendation System (FedRS) forms a classification system according to application scenarios and model characteristics. From the perspective of participant types, it is mainly divided into two categories: cross-device federated recommendation system takes Mobile device as the main participant, and realizes personalized recommendation while protecting user privacy through lightweight design, but it needs to solve the problems of device heterogeneity and unstable training; Cross-platform federated recommendation systems rely on collaboration between different data platforms to improve recommendation effectiveness while meeting regulatory requirements. The core challenge lies in incentive mechanism design and distributed coordination mechanism construction. In terms of model architecture, existing research has developed in three directions: matrix decomposition, deep learning and meta-learning. Matrix decomposition realizes basic recommendation through factorization, deep learning uses neural network to capture complex features, and meta-learning adapts to non-independent and identically distributed data through

parameter initialization and fine-tuning mechanism, so as to improve recommendation effect while protecting privacy.

2.2.3 Triangular balance problem

The core challenge of federated recommendation system lies in the essential trade-off relationship between recommendation effect, privacy protection and system efficiency, which constitutes a typical "triangular balance problem. " First, the non-IID characteristic caused by data island will cause client drift, which makes the model aggregation process face convergence challenge. Secondly, although privacy protection mechanism can effectively prevent security threats such as gradient leakage, it inevitably introduces extra computation and communication overhead; finally, resource constraints of mobile devices further limit the implementation space of the algorithm. Specifically, strict privacy protections (such as differential privacy and homomorphic encryption) will reduce the model accuracy and increase the system load by adding noise or parameter encryption, while pursuing higher recommendation results often requires more complex model architecture and more training rounds, which directly conflicts with the computing power and energy consumption limit of the mobile terminal. Theoretical analysis shows that privacy protection, recommendation effect and system efficiency have inherent trade-offs, which cannot be optimized at the same time. Pareto optimal solution can only be found by dynamic adjustment under specific scenarios.

3 The core dimension of federated recommendation system

Federal recommendation systems face a complex balance of privacy protection, effectiveness enhancement and efficiency optimization. To address this core challenge, existing research focuses on three key dimensions: first, improving recommendation effectiveness through cross-domain knowledge fusion to solve the problem of information fragmentation caused by data silos; second, designing differentiated privacy protection mechanisms to meet the security requirements of different scenarios; and finally, optimizing system efficiency to reduce resource consumption. These three dimensions restrict each other and complement each other, which constitute the core framework of the research on federated recommendation system.

3.1 Effect enhancement and cross-domain knowledge fusion

In federal recommendation systems, data privacy regulations (e. g. GDPR) and commercial competition lead to the dispersion of recommendation data stored in different institutions or user devices, forming a typical data island phenomenon. Federated recommendation systems face the triple challenge of system heterogeneity, statistical heterogeneity and privacy heterogeneity caused by data islands. The heterogeneity of the system is reflected in the low efficiency of training and the difficulty of model convergence caused by the difference of equipment performance; the statistical heterogeneity is reflected in the problems of label offset, feature offset and quantity offset caused by uneven data distribution; and the privacy heterogeneity is reflected in the difficulty of meeting the differentiation requirements and compliance risks of the unified privacy scheme. These challenges affect each other, uneven data distribution will amplify the impact of device differences, and privacy protection mechanisms may aggravate model fluctuations and form complex synergies. Relevant scholars have proposed systematic solutions to the three heterogeneous problems of systems, statistics and privacy.

In terms of system heterogeneity, considering the problem of low training efficiency caused by waiting for all participating devices to complete the calculation based on synchronous communication mechanism, several asynchronous communication optimization schemes are proposed. FedAsync proposed by Xie et al. uses time-varying gradient descent and exponential decay weighting mechanism to systematically reduce the influence of old gradient update on global model. The theory is complete but sensitive to hyperparameters. Client selection strategy optimizes the subset of clients participating in training through intelligent scheduling strategy. Nishio et al. [6] construct a set of resource-aware dynamic client selection framework FedCS. First, collect the computing and network resource information of clients, establish a time estimation model, and then select clients that can complete the whole training process within a specified time. This not only ensures aggregation efficiency, but also avoids training stagnation caused by device performance differences. But it adds extra detection overhead; in terms of statistical heterogeneity, Chen et al. proposed meta-learning Meta-Learning enables models to quickly adapt to new tasks based on a small number of training samples through a mechanism of "learning how to learn." The FedMeta framework proposed by Luo et al. learns globally shared model initialization parameters through model independent element learning (MAML), so that each client can quickly adapt personalized recommendation tasks with only a small number of local training samples, but there are limitations such as high computational complexity of second-order gradient and forced data partition. Reptile algorithm proposed by Wang et al. adopts first-order approximation optimization strategy, which significantly reduces the computational overhead while maintaining the fast adaptation ability of meta-learning, and does not need data segmentation process, which is more suitable for client-side scenarios with limited sample number. Clustering method improves the

personalization effect through parameter clustering and federated graph neural network, but significantly increases the computational burden; In terms of privacy heterogeneity, the probability factor mechanism of FedeRank proposed by Anelli et al. provides flexible privacy control but may affect the performance of the model, and the HPFL hierarchical aggregation strategy proposed by Wu et al. achieves accurate protection but requires higher component partition. These methods have their own characteristics and complement each other in terms of computational efficiency, convergence, personalization degree and privacy protection strength, which need to be weighed and selected according to specific application scenarios.

3.2 Data privacy protection mechanisms

Federated recommendation systems achieve privacy protection by sharing model parameters rather than raw data, but they still face core challenges such as gradient leakage, privacy-performance trade-offs, and data heterogeneity. Research shows that servers can still infer user behavior from intermediate parameters, which can be classified into encryption classes, noise classes, and data control classes based on different privacy protection techniques. Encryption classes include homomorphic encryption, secure multiparty computation (secret sharing); Noise class includes differential privacy, local differential privacy, pseudo-items; data control class includes selective data sharing.

3.2.1 Noise class

Noise-based techniques protect user privacy by adding controllable noise to data or gradients, mainly including pseudo-items, differential privacy (DP) and local differential privacy (LDP). Differential privacy and local differential privacy make it difficult for attackers to infer individual data through mathematical theory, where differential privacy is suitable for global or local noise injection, while local differential privacy is added by the client to avoid relying on trusted servers. The advantages of these methods are low computational overhead and easy to deploy, but noise may reduce model accuracy, and privacy strength and recommendation performance need to be weighed.

3.2.2 Encryption class

Encryption technology mainly relies on cryptographic methods such as homomorphic encryption and secure multiparty computation to directly protect data or gradients. Homomorphic encryption supports computation operations in ciphertext state, ensuring that the server always cannot obtain the original data, but its computational resources are extremely expensive. Secure multiparty computation enables secure collaborative computation among multiple participants through secret sharing mechanisms, especially suitable for cross-agency joint modeling scenarios. Such technologies can provide the highest level of privacy security and are of great value in fields such as medical health and financial services that require extremely high data privacy. However, they generally have technical bottlenecks such as excessive computing and communication costs and complex key management systems, which leads to greater challenges in deployment and implementation in practical applications.

3.2.3 Data control class

Data control class technology achieves privacy protection by giving users the right to choose to share data independently, achieving a balance between privacy protection and model utility. Users can upload only non-sensitive data or partial gradient updates. This selective sharing mechanism effectively limits the information range that servers can obtain. The advantage of this method is that it can flexibly balance privacy protection and model utility, especially suitable for application scenarios with different data sensitivities. However, incomplete sharing of information may result in limited global representation of the model, and noise addition or encryption techniques are often needed to enhance the protection effect.

3.3 Efficiency optimization and resource synergy mechanism

In the training of federated recommendation system, the user end is mainly composed of a large number of mobile devices. Mobile devices generally have limited computing, storage and communication capabilities. Therefore, optimization of the federated recommendation system model reduces various resource overhead during training at the user end.(such as computing resources, storage resources and communication resources), which can lay the foundation for a wide range of practical applications of federated recommendation systems.

3.3.1 communication optimization

Communication optimization technology mainly improves system efficiency from data transmission level. In terms of model update based on importance, Qin et al. proposed PPRSF framework adopts four-layer screening mechanism, and the client only needs to update a small number of candidate items embedded. Efficient-FedRec of Yi et al. decomposes the model into user model and news model, greatly reducing the amount of transmitted data; Among the model compression techniques, Konen et al. proposed two methods of structural update and sketch update, and Sijing et al. proposed JointRec framework combined with low-rank decomposition and 8-bit quantization to achieve 12.83

compression ratio; In terms of client sampling, FedFast framework proposed by Muhammad et al. selects representative users through K-means clustering, reducing communication rounds by 94%, one-time learning such as FedSPLIT proposed by Eren et al. realizes single-round communication by knowledge distillation. Although these methods effectively reduce communication overhead, they generally face the problem of accuracy loss, such as Khan et al. FCF-BTS, which leads to a 4%-8% reduction in recommendation accuracy. Overall, existing communication optimization techniques have achieved remarkable results in reducing transmission costs, but there is still room for improvement in accuracy maintenance, dynamic adaptability and computation-communication trade-offs.

3.3.2 Algorithm efficiency improvement

Algorithm optimization focuses on improving computational performance and convergence speed. FedFast proposed by Muhammad et al. accelerates training through user clustering and parameter sharing within clusters. MetaMF proposed by Lin et al. deploys collaborative memory and meta-recommendation modules on the server side to reduce the computational burden on the client side. Wu et al. uses multi-attention mechanism to shorten training time by 45%. FedMeta learns well-initialized models based on MAML algorithm. Zhou et al.] use CMAB framework to implement context-aware recommendation. These innovative methods improve efficiency from different angles: FedFast reduces redundant computation, MetaMF optimizes computation allocation, attention mechanism improves model architecture, FedMeta enhances generalization ability, and CMAB implements dynamic adjustment. However, each has limitations, such as high implementation complexity or sensitivity to data distribution. Overall, significant progress has been made in improving training efficiency, but how to balance computational efficiency with model performance and improve universality and robustness of algorithms are still key issues to be solved.

4 Summary

In this paper, we systematically study the key technologies and development status of federated recommender systems, and analyze the core progress and main challenges of current research by constructing a "privacy-effectiveness-efficiency" triangular balance framework.

In terms of privacy protection, existing technologies such as homomorphic encryption and differential privacy provide basic guarantees, but they still face high computational costs and model performance degradation. Emerging technologies such as secure multiparty computing show better balance. In terms of recommendation effect, cross-domain knowledge fusion technology achieves data collaboration through multiple federated learning methods, but there is still room for improvement when dealing with heterogeneous data and cold start scenarios. In terms of system efficiency, model compression and selective updating methods effectively reduce communication overhead, but need to further improve the adaptability in dynamic environment.

It is hoped that this study can provide valuable reference for relevant researchers in industry and academia, promote the collaborative development of federal recommendation system in privacy protection, recommendation effect and system efficiency, and contribute to the construction of a more secure, efficient and intelligent recommendation system.

5 Prospects

5.1 Dynamic adaptability

Future research will focus on developing an intelligent adaptive federated recommendation system that enables dynamic optimization of scenario awareness through reinforcement learning frameworks. The system can automatically adjust technical solutions according to industry characteristics: prioritize homomorphic encryption and multi-party security computing in the financial sector to ensure transaction security; optimize lightweight models and fast algorithms for e-commerce platforms to improve response speed; balance data privacy and recommendation accuracy in medical scenarios, and adopt fine-grained access control. This intelligent solution dynamically allocates computing resources by monitoring device status, data characteristics and business requirements in real time, realizing the optimal combination of privacy protection, recommendation effect and system efficiency, and promoting the large-scale application of federal recommendation technology in various industries.

5.2 Fine-grained privacy protection

Differentiated protection for different data sensitivities will become a trend, including: strong encryption for core privacy features and lightweight desensitization for common features; dynamically adjusting privacy budgets at different stages of training; and combining verifiable computing techniques to ensure that privacy measures are truly effective. This layered protection strategy provides precise protection with minimal loss of effect.

5.3 Decentralized incentives

Future research will explore a decentralized incentive mechanism based on blockchain to achieve accurate quantification and fair return of node contributions through smart contracts. The mechanism includes three key designs: first, establish a multi-dimensional contribution evaluation model that comprehensively considers factors such as computing resources, data quality, and participation time; second, design a dynamic reward and punishment algorithm to give token incentives to honest nodes and impose economic penalties on malicious behavior; Finally, a resource-aware elastic participation framework is developed, allowing devices with different capabilities to choose appropriate participation intensity according to their own conditions. This mechanism can not only protect the privacy rights of participants, but also effectively enhance the enthusiasm of collaboration, thus optimizing the overall system efficiency.

References

[1] P. Voigt and A. Von dem Bussche. The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing, 2017.

[2] C. Xie, S. Koyejo, and I. Gupta, "Asynchronous federated optimization," arXiv preprint arXiv:1903.03934, 2019.

[3] D. Chai, L. Wang, K. Chen, and Q. Yang, "Secure federated matrix factorization," Intelligent Systems, IEEE, vol. PP, no. 99, pp. 1–1, 2020.

[4] J. Yi, F. Wu, C. Wu, R. Liu, G. Sun, and X. Xie, "Ef fi cient-fedrec: Ef fi cient federated learning framework for privacy-preserving news recommendation," arXiv preprint arXiv:2109.05446, 2021.

[5] Q. Ma, Y. Xu, H. Xu, Z. Jiang, L. Huang, and H. Huang, "Fedsa: A semi-asynchronous federated learning mechanism in heterogeneous edge computing," IEEE Journal on Selected Areas in Communications, vol. 39, no. 12, pp. 3654–3672, 2021.

[6] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in ICC 2019-2019 IEEE international conference on communications (ICC). IEEE, 2019, pp. 1–7.

[7] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," arXiv preprint arXiv:1904.04232, 2019.

[8] F. Chen, M. Luo, Z. Dong, Z. Li, and X. He, "Federated meta-learning with fast convergence and efficient communication," arXiv preprint arXiv: 1802.07876, 2018.

[9] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in International conference on machine learning. PMLR, 2017, pp. 1126–1135.

[10] Q. Wang, H. Yin, T. Chen, J. Yu, A. Zhou, and X. Zhang, "Fast-adapting and privacy-preserving federated recommender system," The VLDB Journal, vol. 31, no. 5, pp. 877–896, 2022.

[11] V. W. Anelli, Y. Deldjoo, T. D. Noia, A. Ferrara, and F. Narducci, "Federank: User controlled feedback with federated recommender systems," in European Conference on Information Retrieval. Springer, 2021, pp. 32–47.

[12] J. Wu, Q. Liu, Z. Huang, Y. Ning, H. Wang, E. Chen, J. Yi, and B. Zhou, "Hierarchical personalized federated learning for user modeling," in Proceedings of the Web Conference 2021, 2021, pp. 957–968.

[13] J. Qin, B. Liu, and J. Qian, "A novel privacy-preserved recommender system framework based on federated learning," in 2021 The 4th International Conference on Software Engineering and Information Management, 2021, pp. 82–88.

[14] J. Konecnˇ y, H. B. McMahan, F. X. Yu, P. Richt ` arik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," arXiv preprint arXiv:1610.05492, 2016.

[15] Sijing, Duan, Deyu, Zhang, Yanbo, Wang, Lingxiang, Li, Yaoxue, and Zhang, "Jointrec: A deep-learning-based joint cloud video recommendation framework for mobile iot," IEEE Internet of Things Journal, vol. PP, no. 99, pp. 1–1, 2019.

[16] K. Muhammad, Q. Wang, D. O'Reilly-Morgan, E. Tragos, B. Smyth, N. Hurley, J. Geraci, and A. Lawlor, "Fedfast: Going beyond average for faster training of federated recommender systems," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1234–1242.

[17] M. E. Eren, L. E. Richards, M. Bhattarai, R. Yus, C. Nicholas, and B. S. Alexandrov, "Fedsplit: One-shot federated recommendation system based on non-negative joint matrix factorization and knowledge distillation," arXiv preprint arXiv:2205.02359, 2022.

[18] F. K. Khan, A. Flanagan, K. E. Tan, Z. Alamgir, and M. AmmadUd-Din, "A payload optimization method for federated recommender systems," in Fifteenth ACM Conference on Recommender Systems, 2021, pp. 432–442.

[19] Lin Y, Ren P, Chen Z, et al. Meta matrix factorization for federated rating predictions[C]. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 981-990.

[20] Wu A, Kwon Y W. Enhancing Recommendation Capabilities Using Multi-Head Attention-based Federated Knowledge Distillation[J]. IEEE Access, 2023: 45850–45861.

[21] Zhou P, Wang K, Guo L, et al. A privacy-preserving distributed contextual federated online learning framework with big data support in social recommender systems[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(3): 824-838.

[22] Lu T, Pál D, Pál M. Contextual multi-armed bandits[C]. Proceedings of the Thirteenth international conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, 2010: 485-492.