Research on Privacy Protection and Utility Improvement of Recommendation Systems in Multi-Source Data Fusion Scenarios

Chouchak Chan^{1,a}, Ao Liu^{2,b,*}

[1]Faculty of Health Sciences, University of Macau (FHS), Avenida da Universidade, Taipa, Macau, China

[2]College of Science, Mathematics and Technology, Wenzhou-Kean University, Wenzhou, China

[a]charle02113@outlook.com, [b]1235945@wku.edu.cn

Keywords: multi-source data fusion; recommendation system; privacy protection; utility improvement

Abstract: In the scenario of multi-source data fusion, the issues of privacy protection and utility improvement of recommendation systems have increasingly attracted attention. Traditional recommendation algorithms often rely on users' personal information and behavioral data. However, in the big data environment, the risk of user privacy leakage has significantly increased. It is of great significance to study how to improve the performance of recommendation systems on the basis of ensuring user privacy. This paper discusses the recommendation algorithm based on differential privacy mechanism and homomorphic encryption technology. Through the fusion and analysis of multi-source data, a novel privacy protection framework is proposed. By integrating collaborative filtering with deep learning models, a balance between utility and privacy has been achieved. The experimental results show that the proposed method has good feasibility in terms of user privacy protection and is superior to traditional methods in recommendation effect, providing new ideas and directions for future research on recommendation systems.

1. Introduction

1.1. Research Background

With the rapid progress of scientific and technological information, users generate a large number of behavioral records in various activities such as social platforms, online shopping and content sharing. These records are widely integrated into the system architecture to optimize the user's interaction experience, which has caused widespread discussion with this privacy hazards, especially in the multi-source data integration environment, various data sources It may touch users' sensitive information, causing the risk of personal information leakage to rise sharply. Conventional systems rely more on users' past behavior and preference information, so privacy protection measures are particularly critical. Recently, privacy protection technology has made progress, and differential privacy and homomorphic encryption are gradually applied to system construction. This provides an innovative perspective for balancing privacy protection and functional realization. How to develop a high-efficiency system under the background of multi-source data integration to ensure privacy security and improve operation efficiency has become a hot topic focused on by the academic community and the industry. Against this background, improvement strategies are proposed for the system's theory in privacy protection and performance optimization. Exploration and practical application provide framework support and operating guidelines.

1.2. Research Significance

Discussing the issues of privacy protection and efficiency optimization in the system under the situation of multi-source data integration, the value is reflected in multiple dimensions. Users' immediate response needs for privacy protection are improved, and the data leakage risks prevalent in traditional systems urgently need to be efficiently resolved, which gives great significance to the theoretical and practical research of privacy protection. Relying on multiple Source data characteristics, by integrating privacy protection technologies and algorithms, can enhance the

DOI: 10.25236/gemmsd.2025.126

accuracy and personalization level of the system, maximize the security of user data, help strengthen users' trust in the system, and explore the balance mechanism of privacy protection and efficiency. Providing a practical framework and technical guarantee for the future development of the intelligent field plays a key role in promoting the extensive implementation of the system. The research conclusions can provide reference for relevant industries, help build a safe and efficient ecological environment, and lay a solid foundation for achieving the dual goals of business value and social responsibility.

2. Current Situation of Privacy Protection and Utility in Recommendation Systems under Multi-Source Data Fusion

2.1. Technical Principles and Features of Multi-Source Data Fusion Recommendation Systems

The core of the multi-source data integration system technical framework lies in achieving efficient integration and analysis of various data sources [1]. By leveraging users' behavioral information on different platforms, a more comprehensive user feature model is created. This system typically adopts a strategy that combines collaborative screening with content hybrid algorithms to evaluate user preferences from multiple perspectives and provide personalized services. At the technical level, Multi-source data integration relies on data preprocessing, feature extraction, and model training, etc. It enhances accuracy and correlation by leveraging data mining and machine learning technologies. Its characteristics are manifested in the following aspects: Diversity is reflected in the integration of different types of data sources. It includes both structured and unstructured data, significantly enhancing the system's coverage and accuracy. Real-time requirements demand that the system quickly process new data input, striving to ensure that the results are dynamically adjusted to reflect the latest behaviors and preferences of users. Accuracy and personalization are the core pursuits of the multi-source data integration system. Through cutting-edge technologies such as deep learning, it accurately grasping the potential needs of users. In response to privacy protection requirements, designing a reasonable privacy protection mechanism to ensure user information security has become an indispensable part of the system.

2.2. Analysis of Application Scenarios

2.2.1. Cross-Platform User Behavior Data Fusion

Users exchange behaviors in social media, electronic transactions, video platforms and other channels to provide rich nature dimensions for the system. After data mining and analysis, users can have potential interests and orientation [2]. During this period, it usually involves data collection, cleaning and standardization, and makes every effort to ensure the consistency and comparability of data from different sources. In terms of implementation, after using machine learning algorithms to carry out high-efficiency modeling of cross-platform data, identify user behavior patterns, improve accuracy, and combine users' browsing and shopping records on e-commerce web pages, social network likes and comment behaviors, users' consumption needs and interest changes can be comprehensively captured. In view of the analysis of situational information, to better understand the immediate needs of users in specific scenarios and improve the response rate and personalization of the system, how to efficiently handle data privacy issues during the integration of cross-platform user behavior data and make every effort to ensure user information security is still a matter that needs to be paid attention to.

2.2.2. Social Network and Content Data Fusion

Social networks provide the system with a large amount of user relationship and behavioral information, including social interactions such as friend contact, evaluation, likes, etc., while content information involves images, creations, audio-visual materials and other forms of information browsed by users, which are integrated from different angles to support the system to understand users' social background and content preferences [3]. In specific practicem, by analyzing the behavior of users on social platforms, it can identify which content is generally valued. With the help of content related to users' social circles, it enhances the attractiveness of content and users' sense of participation. Using deep learning technology can better grasp the nonlinear relationship between content and social behavior in feature acquisition. Form a personalized plan. In this process, when encountering privacy protection, it is necessary to use privacy protection technology to ensure the security of social information and content, and realize efficient personalized services under the premise of maintaining that user information is not leaked.

2.2.3. Internet of Things and Context-Aware Data Fusion

Object-connected network devices generally collect various daily behavior data of users, such as location, environmental temperature, frequency of equipment use, etc. This information becomes available data input [4]. The data reflects the user's preferences and provides detailed background information, so that the system can adapt to the needs of users in specific situations. At the technical level, by analyzing the data collected by the connected network equipment, the system can realize location and situation-based perception. When the user is in a specific geographical area, the system can automatically recommend the surrounding catering, goods or activities to enhance the user experience; according to the changes in environmental assumptions. The system can intelligently adjust the scheme to meet the immediate needs of users. The context perception system must be highly efficient, integrate diversified data sources, and make every effort to ensure the privacy and security of users. Use differential privacy and other technologies to prevent sensitive information leakage. In the process of integrating the object connection network and context perception data, how to cooperate is crucial to balance privacy protection and functional performance.

3. Challenges of Privacy Protection and Utility Improvement in Recommendation Systems under Multi-Source Data Fusion Scenarios

3.1. Technical Challenges

3.1.1. Data Heterogeneity and Data Island Effect

Data heterogeneity refers to the diversification and differentiation of each data source in terms of format, structure and meaning, which makes the process of data integration and integration more complicated. In the system, data from different sources covers a variety of categories such as user behavior, product characteristics, social interaction, etc [5]. How to efficiently identify standardized heterogeneous data to achieve Efficient integration is directly related to the performance and accuracy of algorithms. The data island phenomenon means that in the process of multi-source data integration, data cannot be effectively shared and integrated, resulting in a lack of correlation between various data sources, the potential value of data cannot be fully exploited, and the behavioral data of users on different platforms cannot be integrated. As a result, it is difficult for the system to fully grasp the real preferences and needs of users. The information isolation phenomenon reduces the accuracy, limits the algorithm optimization space, and affects the user experience. Solving the data heterogeneity and isolation phenomenon is a necessary condition for improving the privacy protection and efficiency of the system. It is necessary to realize various numbers through the innovation of data processing technology and algorithm design. Efficient integration and utilization according to the source.

3.1.2. Trade-off between Privacy Protection and Model Utility

At the level of practical application, to enhance personalization and accuracy, the system usually requires a large amount of user behavior information and private data. The process of information collection and utilization may infringe on personal privacy, thereby weakening users' trust in the system. In the system construction stage, how to maximize the model effectiveness of user privacy has become urgent [6]. The key issues to be solved, privacy protection technology, such as differential privacy, homomorphic encryption and data generalization, although can effectively prevent user data leakage, it often has an adverse impact on model performance. Technical means introduce interference factors in the computing process, which is easy to cause a decrease in accuracy. It even affects the user experience. Diversified privacy protection measures will increase the consumption of computing resources and reduce the response speed of the system. Researchers urgently need to explore innovative paths, seek an ideal balance between privacy protection and performance, and provide users with services with both security and high efficiency by optimizing algorithm architecture and data management strategies.

3.2. Challenges Brought by Multi-Source Data Characteristics

3.2.1. Data Quality and Noise Problems

During multi-source data integration, data sources often show inconsistent formats, missing information, and abnormal points [7]. These problems affect system performance and reduce accuracy. High-quality data is the basis for accurate personalized service. In actual use, how to clean and optimize data with high efficiency is a complex and important task. Models need a good data environment for training and prediction. Data noise affects model training. It influences feature extraction and scoring. Noise often comes from user input errors, abnormal operations, or system faults. Poor data causes bias in algorithm output. It cannot reflect real user preferences and needs. Strengthening data preprocessing, using noise filtering and data enhancement, and increasing data usability and stability are key strategies. Solving data quality and noise problems builds a solid foundation for better system performance.

3.2.2. Data Correlation and Privacy Leakage Risk

Information from different data sources is integrated, and multi-dimensional information such as user behavior records, geospatial distribution, social network data, etc. are interconnected, triggering the exploration of personal privacy. The information association mechanism can more accurately grasp the needs and preferences of users and exacerbate the risk of data abuse. When the information in various data sources combined analysis, the originally anonymous or non-sensitive data will expose personal privacy in specific situations, triggering privacy leakage [8].

User trust is affected by the risk of privacy leakage, causing legal and ethical disputes, which brings serious challenges to the reputation and business operation of the enterprise. When building a multi-source data integration system, it is necessary to pay attention to the setting of privacy protection measures, and strive to realize the value of data association under the premise of ensuring user privacy. Through the use of advanced encryption segment, differential privacy and other strategies can reduce the risk of privacy leakage to a certain extent, reasonably plan the collaborative design of privacy protection mechanisms and algorithms, and achieve a balance between user privacy and efficiency, which is of key significance for enhancing system security and optimizing user feelings.

3.3. Ethical and Legal Dilemmas

3.3.1. Definition of User Data Ownership and Usage Rights

User data is widely collected and analyzed, and individuals' awareness of control of information is increasing day by day. How to clearly define data ownership and use norms constitutes the cornerstone of compliance development. Generally speaking, users are recognized as the owners of data, but in practice, the generation, storage and application of data involve a variety of subjects, including service providers, data platforms and third-party partners make the interpretation of control ambiguous, which leads to legal disputes and ethical discussions.

To ensure the rights and welfare of users, the laws and regulations of various countries put forward different requirements for the processing of personal data. The General Data Protection Regulation of the European Union (GDPR) clearly gives users the right to dispose of personal information [9]. In the process of system construction and implementation, legal norms must be comprehensively considered and strive to ensure that users are fully informed before and independently decide how to use information. Organizations should formulate clear guidelines for the use of information, coordinate users' privacy rights and commercial interests with the help of appropriate information control and exchange means, and achieve the continuous consolidation of user trust and the continuous progress of the system. Properly explaining the right to privacy and use of user information helps to find an ideal balance between privacy protection and efficiency improvement.

3.3.2. Conflict between Privacy Regulations Compliance and Technical Implementation

Data protection laws in many countries become stricter. GDPR and CCPA state user rights to access, modify, and delete data. They require explicit user authorization for data processing [10]. They set higher standards for transparency and security. Systems need large amounts of user data for model training and optimization. They need data to improve accuracy and efficiency. This creates conflict with regulatory compliance. How to use data efficiently while protecting privacy becomes an important challenge. Differential Privacy and Homomorphic Encryption can protect information. But they still reduce computing performance and algorithm accuracy. Researchers and developers must explore new solutions. They must find a balance between privacy compliance and algorithm efficiency.

4. Optimization Strategies for Privacy Protection and Utility Improvement in Recommendation Systems under Multi-Source Data Fusion

4.1. Technical Optimization Paths

4.1.1. Improving Data Fusion and Representation Learning Methods

With the help of cutting-edge means such as Graph Neural Networks (GNN), it can efficiently capture the intricate connection between users and goods, integrate multi-attribute information, and build more explanatory user descriptions [11]. At the level of characterization learning, deep learning algorithms can be used, including Including the self-encoder and the variable self-encoder, in order to ensure the confidentiality of data, the distillation of advanced abstract features, through the implementation of unsupervised learning of user interaction information, the system obtains valuable characterization forms, optimizes the efficiency, and enhances the confidentiality level of the algorithm under the premise of not disclosing the original information. It can integrate differential privacy technology to strengthen the feature learning process, restrain the risk of information leakage to maintain user information security, and strengthen the coordination of data integration and characterization learning technology, which can not only efficiently improve system performance, but also strive to ensure sustainable progress under confidentiality conditions.

4.1.2. Strengthening the Use of Privacy Computing Technologies

Differential Privacy, Homomorphic Encryption, and Secure Multi-Party Computation provide multi-layer protection for user data [12]. Homomorphic Encryption allows computation on encrypted data. The original information remains encrypted at all times. It ensures privacy protection through the entire computing procedure. It is suitable for complex algorithm tasks. Differential Privacy adds random noise to query results. It prevents reverse inference of individual information. It reduces privacy leakage risk. It is suitable for large-scale data analysis. Secure Multi-Party Computation ensures that multiple data holders can only obtain the final computed result. They cannot access each other's raw data. It protects privacy and maintains the value of data sharing.

4.2. Strategies for Improving Model Utility

4.2.1. Designing Privacy-Protected Collaborative Filtering and Deep Learning Models

Privacy-protected collaborative filtering can introduce Differential Privacy. By adding noise to user ratings or behavior data, it reduces the risk of leaking individual information. It protects user privacy. When detecting similar users or items, the system can still obtain useful similarity signals, even with noisy data. Accuracy and relevance can be maintained. Deep learning models, deep collaborative filtering networks, and Long Short-Term Memory networks (LSTM) can perform deep data mining under privacy protection [13]. With Homomorphic Encryption, deep learning models can train and infer on encrypted data. The system protects original data. The system can also perform online learning through user feedback. It can adjust privacy measures based on user needs.

4.2.2. Introducing Federated Learning and Differential Privacy

Federated Learning is a distributed machine learning framework. It trains models locally on user devices. It only uploads model parameters to the central server. This reduces the risk of centralized storage of sensitive data. User information stays on local devices. It reduces leakage and misuse. It enhances system security. Adding Differential Privacy to Federated Learning adds noise to uploaded model parameters. It protects user privacy. Differential Privacy controls the model's sensitivity to individual data. Even during aggregation, attackers cannot infer original user information. It provides strong privacy protection.

4.3. Ethical and Legal Safeguards

4.3.1. Improving User Data Authorization and Usage Rules

In order to protect user privacy and improve efficiency, enterprises need to formulate clear data collection and use guidelines, especially before obtaining user data, and must obtain the explicit consent of users to ensure that users fully understand the use of data and processing processes. In the process, they should explain to users with the help of easy-to-understand terms of service and user agreements to reduce information deviation and enhance openness. In the stage of data use, it is necessary to clearly define the data retention period and scope of use. For data beyond the purpose of use, it can only be processed after obtaining the user's permission. The enterprise should give the user the right to manage personal information, including the permission to access, modify and delete. In order to enable individuals to control their own information independently, a data use supervision system should be built, data operation behavior should be regularly reviewed, and violations should be found in a timely manner that meet the authorization standards.

4.3.2. Establishing Privacy Protection Assessment and Audit Mechanisms

Privacy assessment should be carried out regularly to identify and evaluate the privacy risks that may exist in all aspects of data collection, preservation and processing. In the evaluation process, it is necessary to consider data flexibility, user privacy needs and legal requirements. By building a privacy impact assessment framework, various data processing activities can systematically analyze the privacy of users. The influence of, strive to implement efficient privacy protection strategies under the background of big data. The supervision mechanism should track and record data processing behavior in accordance with specific guidelines and procedures, and strive to ensure that all operations comply with the established privacy norms and laws and regulations, in the supervision process. It is necessary to regularly check the access and use of user data, identify any unauthorized operations, and take timely remedial measures to prevent privacy leakage.

5. Conclusion

Under the background of multi-source data integration, system privacy protection and performance optimization are the key to achieving the two goals of personalized service and user trust. By exploring the integration of privacy protection technology and algorithm efficiency, under the premise of ensuring the security of user data, the accuracy and relevance of the system can be enhanced, and differential privacy and the same State encryption and other cutting-edge technologies can not only fully protect user privacy, but also allow the system to efficiently analyze dynamic information, deepen data integration strategies and characterization learning innovation, and play a decisive role in improving results. The introduction of federal learning helps to maintain user privacy and realize the efficient use of multi-source data. Improve the application guidelines of user data authorization, build a privacy protection evaluation and supervision system, and provide a solid guarantee for the compliant operation of the system. In the future, research needs to focus on the dynamic balance between privacy protection and effectiveness, and explore more intelligent security data processing schemes to cater to users' increasing privacy demands and personalized services. Looking forward to the continuous evolution of the system in the data era.

References

[1] TAN Z H. Design and implementation of a metadata service management platform for multi-source heterogeneous big data[D]. Beijing: Beijing University of Posts and Telecommunications, 2021.

[2] LI Y Q. Analysis of marketing thinking on social media platforms under the "Internet+" context[J]. Network Security Technology and Application, 2017(4): 3. DOI:10.3969/j.issn.1009-6833.2017.04.105.

[3] ZHAO Y C. User browsing content analysis and user interest mining[D]. Chongqing: Chongqing University, 2024. DOI:10.7666/d.y704131.

[4] WU D, LIANG S B. A review of online information search behavior in multi-device environments[J]. Journal of Library Science in China, 2015, 41(6): 19. DOI:10.13530/j.cnki.jlis.156009.

[5] LIU H N, ZHANG Y, LV P W. Application of heterogeneous databases in the two-ticket system of thermal power plants[J]. Computer Simulation, 2014, 31(11): 4. DOI:10.3969/j.issn.1006-9348.2014.11.031.

[6] LI M Z. Research on key technologies of privacy protection in location-based social networks[D]. Beijing: Beijing University of Posts and Telecommunications, 2023.

[7] TONG L, WANG Z M, YI D Y. Research on multi-source information evaluation methods for data processing[J]. Systems Engineering and Electronics, 2006, 28(6): 5. DOI:10.3321/j.issn.1001-506X.2006.06.007.

[8] CUI G Y, WANG P. Risk challenges and agile governance: Personal information protection in the context of digital economy[J]. Library Forum, 2024(3).

[9] TIAN X P. Extraterritorial effect of the EU GDPR: jurisdiction, implementation path, institutional effects and implications[J]. International Journal of Economic Law, 2023(1): 20-36.

[10] WANG M, CAO F. Data protection standards in Europe and America and Chinese strategies in the GDPR era[J]. Journalism University, 2022(7): 53-66.

[11] SCARSELLI F, YONG S L, GORI M, et al. Graph neural networks for ranking web pages[J]. ACM, 2005. DOI:10.1109/WI.2005.67.

[12] AHAMED S I, RAVI V. Privacy-preserving chaotic extreme learning machine with fully homomorphic encryption[C]//International Conference on Data Management, Analytics & Innovation. Singapore: Springer, 2024. DOI:10.1007/978-981-97-3242-5_40.

[13] PATNAIK S K, BABU C N, BHAVE M. Intelligent and adaptive web data extraction system using convolutional and long short-term memory deep learning networks[J]. Big Data Mining and Analytics, 2021, 4(4): 19. DOI:10.26599/BDMA.2021.9020012.