Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6640667, 15 pages
https://doi.org/10.1155/2021/6640667

Research Article

SDRM-LDP: A Recommendation Model Based on Local Differential Privacy

Gesu Li,[1] **Guisheng Yin**,[1] **Jishen Yang**,[2] **and Fukun Chen**[1]

[1]*College of Computer Science and Technology, Harbin Engineering University, Heilongjiang, China*
[2]*Department of Computer Science, Georgia State University, Georgia, USA*

Correspondence should be addressed to Gesu Li; lgs7788@hrbeu.edu.cn

Received 4 December 2020; Revised 7 January 2021; Accepted 19 February 2021; Published 18 March 2021

Academic Editor: Jinbao Wang

The development of 5G technology has driven the rise of e-commerce, social networking, and the Internet of Things. Under the high-speed transmission, the data volume increases, and the user demand also changes. Personalized customization has become the mainstream trend of network development. However, as the speed of the Internet increases, a series of problems also arise. The increase in data volume results in a reduction of bandwidth, a growth of the central processor's pressure, and a higher risk of data leakage. A search system and a recommendation platform are the tools to improve people's search efficiency. However, providing personalized recommendations to different users according to their needs is still an urgent problem. Simultaneously, the big data volume means that attackers can also get more information. They can use background knowledge and various reasoning methods to deduce the user's private information using nonprivate items. In this paper, the solutions to safe and reliable recommendation services are the main problem explored. Based on this idea, this paper proposed short-term dynamic recommendation model based on local differential privacy (SDRM-LDP). This model uses a small amount of user information to construct short-term user preference behaviors and provides recommendations for users based on the similarity between items. We consider that an attacker uses nonprivate items to derive privacy items. Therefore, we randomly replace the original data in the same category. At the same time, the local differential privacy (LDP) is added to the privacy item query to make the private data available and protect the privacy information. In this paper, two real-world datasets, ML-100K and ML-10M, are used for experiments. Experimental results show that the results of SDRM-LDP are superior to other models.

1. Introduction

The fifth-generation mobile communication system (5G) supports enhanced mobile broadband (eMBB), mMTC, and uRLLC. In high-speed transmission, the user's personalized needs are enhanced [1]. Users want to be able to get their personal customization in the shortest time. At the same time, as the mainstream of the network platform, e-commerce, social networks, and especially the Internet of Things are inseparable from personalized recommendation services. Now, the development trend of the Internet of Things is the combination of e-commerce. For example, in the scenario of driving a car [2], the navigation recommends the relevant destination to the user according to the time and the user's past behavior and habits. Also, smart appliances are connected to the mobile phone terminal and set automatically according to the user's past behavior. Therefore, when users open the corresponding APP, relevant products are recommended to users according to their behavior habits and needs. Inevitably, the emergence of 5G accelerates the development speed of personalized recommendation. The traditional recommendation system is a mostly static recommendation. It gathers users' information and computes on a reliable device and finally outputs the recommended lists to users. There is a lag in the recommended results. As the speed of the network increases, users get more information. As a result, as the amount of information increases, users' needs change. Some regional businesses collect all the data in the cloud for processing, which commonly wastes bandwidth and increases transmission latency. After the user buys an item, the platform recommends more related items to the user, but at this time, the items are no longer needed by the user. Therefore,

transmission delay and data quantity determine the processing mode of 5G service. That is, computing is not all in the core network. Marginal computation allows the business to process a part of the data on the client side, such as data processing and encryption. By this method, the delay and load are reduced, and the efficient processing capacity is realized. This method reduces the delay and overload and improves the computing performance. Based on the background of 5G, this paper proposes a short-term dynamic personalized recommendation model on platforms such as the Internet of Things and e-commerce. The model uses marginal calculation to process a part of the user's data at the terminal to reduce bandwidth waste and latency and add the privacy protection mechanism. The privacy protection mechanism can protect the user's personal privacy and data by preventing attackers from using background knowledge and statistical knowledge to reversely deduce user privacy information.

Collaborative filtering is the main recommendation algorithm in the market. The following problems exist. (1) Due to the large data volume, the transmission process is easy to cause data leakage and loss. (2) The data is static, and there is hysteresis in the process of recommendation. (3) As the data is processed centrally, the computation speed is low for big data processing. Holbrook and Schindler [3] proposed in 1989 that users' preferences change over time. The user's preference is related to age, gender, etc. However, the traditional recommendation system mainly considers the similarity between users and items and ignores the users' preferences. With the development of the network, the promotion of intelligence, personalized service is the future mainstream. Using similarity to build a model is no longer enough to satisfy users. Therefore, we need to explore user preferences, to provide users with personalized customized services and drive the development of e-commerce and the Internet of Things.

Of course, there are always two sides to everything. 5G brings fast network experience and data transmission speed. More data are uploaded to the Internet. Attackers also can get all kinds of data more easily [4]. For example, in smart home appliances, users' daily behavior data are uploaded to the server center, which is likely to cause data leakage in the process of transmission [5]. Some attackers can use the existing background knowledge combined with statistical knowledge to infer the private information and find out the target user. For the network threats, how to provide users with safe and reliable recommendation services is the primary content of our discussion and research. We proposed SDRM-LDP. As mentioned earlier, to reduce transmission delay, the model puts some work of data processing at the user end. By filtering, the model reduces unnecessary data and provides users with more accurate recommendation services with the least user data. The model publishes the processed nonprivate data and the privacy item query results of privacy protection processing. This model uses the user's recent history to predict the user's future behavior and provides users with recommendations based on the predicted results. The retained data is protected by a random substitution combined with LDP to protect its private information and associated data to provide users with more safe and reliable fast recommendation services.

In this work, we focus on the following issues: (1) the recommendation in sparse data, (2) the prediction of short-term dynamic preferences, and (3) the privacy protection in data publication and query. The following is the summary of our contributions and improvements:

(i) This paper uses hidden Markov chains (HMM) to predict users' short-term preferences and uses graphs to explore the relationship between categories and items. Finally, recommendation lists are built based on the above information

(ii) The proposed model in this paper protects user privacy from two aspects: data publishing and data query. After deleting and randomly replacing nonprivate information, the recommendation system is constructed based on the published information. The above operations add local differential privacy to each user's privacy items and provide privacy protection for query results

(iii) The proposed model balances the data availability and the privacy security. The balance is illustrated in quantitative indicators

This paper is organized into six sections as follows. Section 2 summarizes the current research on relevant work, which is mainly divided into two subsections as the introduction of the state-of-the-art status of recommendation system and the introduction of the research status of recommendation system based on privacy protection. In Section 3, we introduce the preliminary algorithms, including HMM, graph, and local differential privacy. Section 4 gives the demonstration of the SDRM-LDP model. Section 5 introduces the overall framework and the specific algorithms for recommendation and privacy protection. Section 6 is about the experiments including the information of datasets, experiment settings, and results. Section 7 is the conclusion of this paper and the future work plans.

2. Related Works

2.1. Recommendation Systems Based on HMM. The traditional recommendation systems construct the similarity matrix according to historical records and then sort and recommend based on the calculation. Later, more scholars [6] consider preference and context information when designing a model. Most of the researches use static data, but the preferences change over time in a more realistic setting. Therefore, some scholars proposed dynamic recommendation. At present, most of dynamic recommendations are based on HMM. There are mainly the following reasons why HMM gains such popularity. (1) HMM is dynamic and based on time series. (2) Through the study of its hidden layer and observation layer, scholars can discuss the development of user preference. Aghdam and Mobasher [7] explore the utility of user preference and introduced hierarchical HMM to capture the changes in user preferences. According to the user's feedback sequence to the model, the hierarchical HMM uses the user's current context as the hidden variable.

For known users, the model is used to infer the maximum likelihood sequence of the transformation between contexts and to predict the probability distribution of the next behavior based on this sequence. Tamayo et al. [8] study the changes of user preferences and propose a model which is a collaborative filtering recommendation system with temporary dynamics based on HMM. This model can trace the changes in user preference. Sahoo et al. [9] explain the user's item selection behavior based on HMM and provide personalized recommendations. A negative binomial-multinomial mixing model is proposed to simulate the user's choice of different items in each time period. Epure et al. [10] propose a personalized recommendation method to provide recommendations for specific users. It provides a basic degree of personalization while complying with the key characteristics of news recommendation including news popularity, recency, and the dynamics of reading behavior. All the above papers on dynamic recommendation are based on HMM. HMM illustrates good applicability in dynamic behavior prediction. This paper proposes SDRM based on HMM. This model is the same as the previous models. It is a dynamic model based on HMM. However, this model is not a dynamic prediction at the level of items, but a short-term dynamic prediction of users' internal preferences. This model pays more attention to user preferences rather than the correlation between items and users on the surface. However, the final purpose of this paper is to provide users with a list of recommendations. Therefore, on the basis of predicting the user's future behavior, we also explore the correlation between preferences and items, so as to build a personalized recommendation list in line with the user's personal preferences.

2.2. Recommendation Systems Based on Privacy Protection. The goal of a recommendation system is to recommend items or social content that users might be interested in. In recent years, many recommendation systems provide personalized recommendation services for users by collecting user's explicit or implicit characteristics, such as occupation, gender, age, address, and other information. Undoubtedly, the more comprehensive information, the better customized recommendations. However, improper collection, storage, and transmission of data lead to leakage of user sensitive information. Therefore, many scholars seek a more secure and reliable recommendation system, which prevents users' personal privacy from leaking while satisfying users' recommendation needs [11]. Chi et al. [12] use the location information as the research target. They proposed SRAmplified-LSH, which is amplifies LSH recommendation service based on location-sensitive information. This method can ensure the balance between accuracy, efficiency, and privacy. Garcia Clemente and He et al. [13, 14] study the privacy protection on mobile commerce recommendation. They used K-anonymity to protect user's id without the need for a trusted third party. The service is by the intelligent user selection algorithm based on grid map and Naive Bayes. They proposed a new framework to support private queries and evaluations. Al-Nazzawi et al.'s study [15] is a summary of the current research status based on LBRS. This paper explores the standards of privacy protection. Also, the paper demonstrates the existing privacy protection methods and concerns on privacy measurement and attack. He et al. [16] proposed a protection scheme for the potential privacy of users in social networks so that the potential privacy of users can be protected without affecting the utility of data.

3. Preliminaries

We propose a recommendation system based on HMM and graph, which is SDRM-LDP. We use HMM to predict the changes in user preferences and use the graph to build the relationship between categories and items. Based on this information, we constructed a short-term dynamic recommendation system. We also consider the personal privacy disclosure and the attackers' use of background knowledge to infer target users and other situations that may cause harm to users' privacy. We propose a privacy protection mechanism in the recommendation system, which could protect personal privacy while providing users services. In the recommendation part, this paper addresses the cold start problem and predicts the future behavior of users who have historical records. This part mainly involves HMM's Baum-Welch algorithm [17]. There is also a brief introduction to graph theory and LDP. The details are demonstrated in the following part.

3.1. HMM & Baum-Welch. HMM is used to describe a Markov process with an implied position parameter. It is derived from the Markov model and is a double stochastic process. It contains a finite state Markov chain that describes the probability distribution of a transition from one state to the next. Another stochastic process represents a probabilistic correspondence between observable values and hidden states. The formula is as follows.

$$\lambda = (\pi, A, B). \tag{1}$$

π is the initial state vector. A is the transfer matrix, and B is the emission matrix. π and A determine the sequence of states. B determines the sequence of observations. The process is shown in Figure 1.

The Baum-Welch (BW) algorithm is proposed by Welch in 1972 [18]. It is a special case of Expectation-Maximization (EM) algorithms. The algorithm is used to find the optimal HMM parameter $\widehat{\lambda}$, which makes $P(O \mid \lambda)$ maximum, in the case of a given observation sequence. In this paper, the BW algorithm is used to find the general optimal parameter $\widehat{\lambda}$ by combining user features and data features based on $\widehat{\lambda}$. The model adjusts $\widehat{\lambda}$ for each user and optimizes the user's personalized parameters become λ_i. The model uses the personalization parameter λ_i to predict future preferences. The particular algorithm of BW is not specified in this paper.

3.2. Graph. Graph theory is the applied mathematics of graphs. It is the most commonly used modeling language and analysis tool for research and implementation. In graph theory, the direct mapping of nodes and edges shows the relationships between nodes in a network. We use a graph to simplify the complexity of the network so the relevant

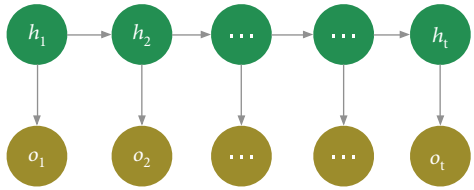


Figure 1: HMM structure.

information can be obtained more quickly and easily. We know that G consists of node V and edge E; namely, $G = (V, E), v_n \in V, e_m \in E$. The nodes and edges of a graph usually map directly to the relationships between nodes in a real network. The degree is the number of neighbors of a node as a basic index to evaluate the influence of a node. In this paper, the bipartite graph theory is used. Bipartite graphs divide nodes into two categories and represent the two related classes of nodes in a diagram. This makes it easy to analyze the interaction between nodes of different classes. In this paper, the user node and the item node are used as two types of graphs and the model uses the user as a medium to find relationships between items.

3.3. Local Differential Privacy. Privacy protection is a high concern. At present, scholars have put forward many methods to solve this problem, including anonymity, encryption, imnoise, and differential privacy. In this paper, local differential privacy is used to protect the recommendation system. This subsection gives a brief introduction to the basic concepts of local differential privacy. Traditional differential privacy, also known as central differential privacy, processes raw data centrally and publishes it to the public. The premise is that the data collector is a trusted third party. However, this is very difficult to do in reality because, since the birth of local differential privacy, it transfers the work to protect privacy to each user. This approach greatly reduces the likelihood of privacy leaks. The definition is set as follows.

Definition 1. (local differential privacy).

For any local differential privacy function f, its domain is $\mathrm{Dom}(f)$. Its range is $\mathrm{Ran}(f)$. For any input k, $k' \in \mathrm{Dom}(f)$, its output is $k^* \in \mathrm{Ran}(f)$. It satisfies the following formula.

$$P[f(k) = k^*] \leq e^{\varepsilon} \times P\left[f\left(k'\right) = k^*\right]. \tag{2}$$

4. Problem Statement

Before introducing the framework and algorithm of the paper, we give a clearer definition of concepts and formulas. This section is mainly composed of two parts. One is to give a clearer definition of some concepts that may cause confusion so the users can better understand this model. The other part is to give a specific definition of the formula involved in the algorithm.

4.1. Concept Definition. For a more explicit explanation of the research content, we give a clear definition of the concepts involved in the paper. For example, what is the preference in the text? What part of our privacy should we protect? How do users query the statistical results of private information? We will all give clear answers by definition.

Definition 2. (like and preference).

Like: the user's choice of an item.

Preference: the user's choice of a category.

Like is a moment. For example, the user buys a red ball today, which may be a temporary choice. But if the user bought other balls in the same style, in a different color, then this user might have a preference for balls. It can be seen that liking exists in preference, and preference has a wider range that is why we chose to research preference. In this paper, we consider categories as preferences and study users' future behaviors based on that preference.

Definition 3. (privacy query).

There are two output results of this paper. One is a list of recommendations and the other is a statistical query result based on user privacy items. Privacy queries are defined as follows.

Privacy query: given the user information table D, the query function $f(D^s)$ is the statistics of the frequency or mean or probability of some privacy S. The specific formula is as follows.

$$f(D^s)\colon D^s \longrightarrow R. \tag{3}$$

Privacy query has some security risks. Attackers can target users based on background knowledge and privacy query results. Therefore, privacy protection is needed for this part.

4.2. Basic Definition

Definition 4. (G).

$G = U \times L$, where L is the link between users and items. U is defined as follows.

$$\begin{cases} u_i \in U, i = |U|, \\ C \times \mathrm{Pr} = u_i, c_n \in C, n = |C|, \\ g_m \times R = c_n, \{1, 2, 3, 4, 5\} = R, \\ g_m \in G, m = |G|, \\ \{\mathrm{id}, \mathrm{sex}, \mathrm{age}, \mathrm{code}\} = \mathrm{Pr}, \end{cases} \tag{4}$$

where u_i consists of historical records C and personal privacy items Pr. c_n is composed of its corresponding categories and ratings. The number of history records of users is (i, n). The privacy item consists of the user's name, gender, age, code, and other privacy information.

Definition 5. (Λ & $\widehat{\lambda}$).

$\lambda_i \in \Lambda$, λ_i represented the model of u_i. Λ is a model set for all users. $\widehat{\lambda}$ is the optimal model based on all user data. This model is used to solve the cold start problem. The item category is taken as the hidden state. $g_m \in G, m \in |G|$. G is genre

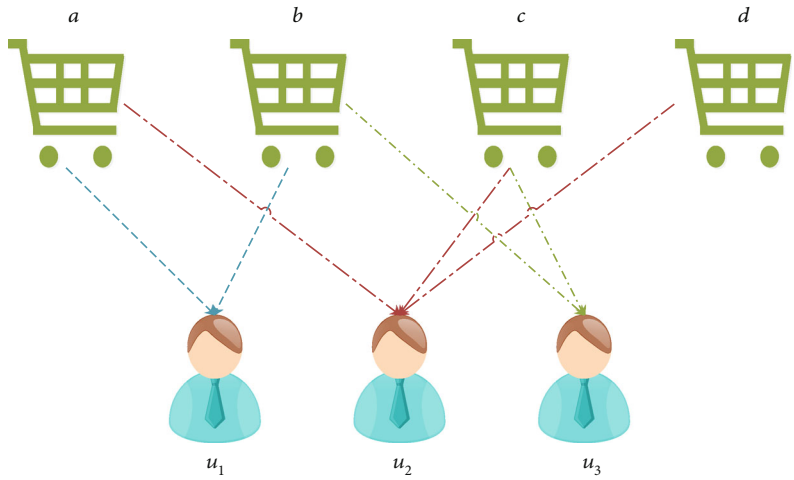


Figure 2: User-item relationship path diagram.

set. m is the number of genres. Each genre g_m contains n specific items. That is, $c_n \in g_m, n \in |g_m|, c_{n\times m} \in C, |C| = n \times m$, C is item set. Rating as observed status $R = (1, 2, 3, 4, 5)$.

HMM consists of two sequences. One is a sequence of states also known as hidden sequence H, $h_t \in H$. The other one is the observation sequence O, $o_t \in O$. The genre sequence of items in the user's history is H. The sequence of corresponding ratings is O. A time threshold is set during preprocessing. The recent T items can be retained, while the others are deleted. Therefore, T can be regarded as the time threshold for filtering the history and can also be viewed as a sequence number for HMM. $t \in T$.

Definition 6. (λ_i).

We present a general HMM model in Definition 5, which is specific for cold star users. This model is not suitable for the personalized recommendation of users with certain historical records. Therefore, we adjust $\widehat{\lambda}$, according to the user's personal historical records. We add weights to the transfer matrix of the generic model to make it fit the user's personal preferences. The formula is as follows.

$$\widehat{a}_{i,j} = a_{i,j} \times \sigma, \tag{5}$$

where $a_{i,j}$ represents the transfer probability of u's item i to j. i and j are items in u's historical records. σ represents the weight added to an item in u's history.

Definition 7. (relationship matrix E).

Given an undirected bigraph G, the vertices of G are composed of users and items. We build the relationship matrix E between user and item based on G. Connect users to nonhistorical records using multilevel hops. $E = U \times C, u_i \in U, c_n \in C$. e is the number of paths between u_i and c_n. Examples are shown in Figure 2.

u_1 watches the movies a and b. u_2 watches the movies a, c, and d. u_3 watchees the movies b and c. Then, we have $(a \longrightarrow b) = 2$, $(a \longrightarrow c) = 2$, $(a \longrightarrow d) = 2$, $(b \longrightarrow c) = 2$, $(b \longrightarrow d) = 2$, $(c \longrightarrow d) = 1$. If we recommend a list for u_1 because u_1's historical records have a and b, $(\{a, b\} \longrightarrow c) = 4$, $(\{a, b\} \longrightarrow d) = 4$. Hence, we recommend c and d for u_1.

Definition 8. (privacy).

The LDP in this paper is designed to address the situation where an attacker uses nonprivate items to reversely deduce user privacy information. For example, in a data list, 40% of users who have bought A are female, 60% of users who have bought B are female, and 20% of users who have bought C are female. Now, u has bought A and C, and then, there is a high probability that U is male. On the basis of formula (2) and in combination with the requirements of this paper, formula (7) is given. Meanwhile, the privacy budget is formula (12). If a user randomly changes an item in his or her history and randomly responses to his or her privacy items, the change in accuracy is no bigger than ε, and the privacy is protected.

Formula (6) is a random exchange formula for user history. u_i has t records. If one is randomly picked due to the range of time threshold setting is small, there will be few records. If there are too many random substitutions, the data utility will become too low. Our purpose is to confuse the attacker, so replacing one of them is enough. The substituted item is other nonhistorical record in the same genre. The probability of that is $1/(n_u - q)$. n_u is the number of history records for the user. n_g is the number of items in this category. q is an item which has appeared in the historical records. Local differential privacy processes for privacy item query in this replace item. Through derivation, the privacy protection process satisfies the random disturbance mechanism and conforms to the definition of local differential privacy.

$$P(c_i \to c_j) = \frac{1}{n_u} \times \frac{1}{n_g - q}, \tag{6}$$

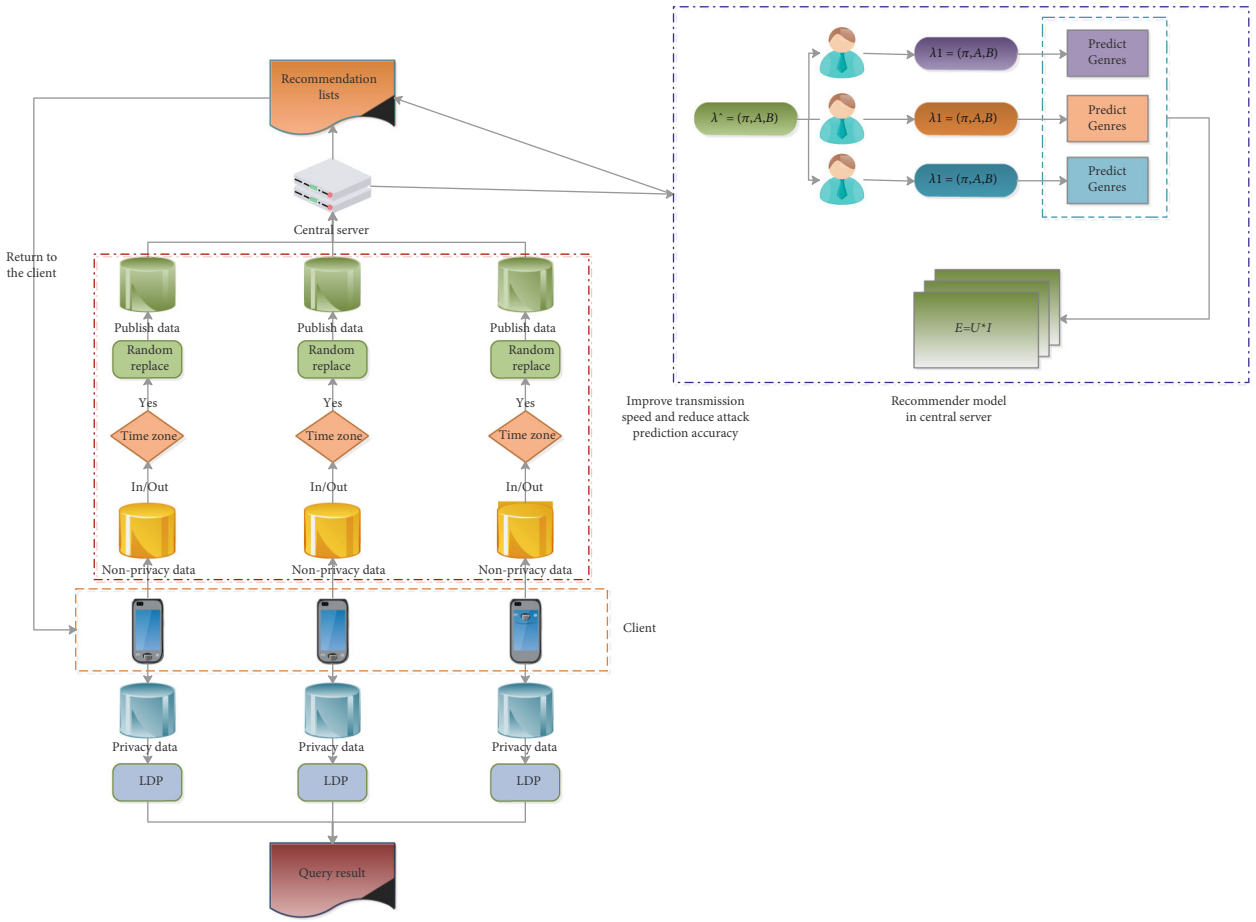


Figure 3: Overall framework of SDRM-LDP.

Input: $U = C \times \Pr; \alpha ; \gamma$
Output: The Query Results & New History Records
1: **Initialization**
2: **for** i in $| U |$ **do**
3: Sort $c_{i,n}$ based on time
4: **if** $n \leq |T|$ **then**
5: Deleting $c_{i,n-|T|}$.
6: Get new records $c_{i,T}$.
7: **Perturbing History Records**
8: $P(c_i \longrightarrow c_j) = (1/t) \times (1/(n-q))$
9: Get $c_{i,T}$.
10: **end if**
11: **end for**
12: **Local Differential Privacy**
13: Probability of true query result is γ, add noise α for each user.
14: $(P[f(k) = k^*])/(P[P_{c_i \longrightarrow c_j} \times f(k') = k^*]) \leq e^{\varepsilon}$
15: $\varepsilon = \ln(\alpha/(1-\alpha))$
16: Get the privacy item query result k^*.

Algorithm 1: Privacy protection.

Input: π ; $O = \{o_1, o_2, \cdots, o_t\}$; $H = \{h_1, h_2, \cdots, h_t\}$; $G = (V, L)$, $V = U \times C$
Output: Recommendation Lists Rel.
1: **Get** $\widehat{\lambda}$ & λ_i
2: **Initialization**
3: Get $\widehat{\lambda}$ based on all dataset
4: **for** i in $|U|$ **do**
5: $A_i = A_{c_{i,n}} \times \beta, 1 \le \beta$
6: $B_i = B_{c_{i,n}} \times \beta$
7: Get λ_i
8: Predicting the next state item genre $g_{i,m}$
9: **end for**
10: **Get Recommendation list Rel**
11: Initialization $E = U \times C$
12: **for** i in $|U|$ **do**
13: **for** t in $|c_i|$ **do**
14: Find all the users that satisfy condition $c_{i,t} \neq 0$. Extra and get a new matrix.
15: Sum each column separately. Get submatrix $SE_{i,t}$
16 **end for**
17: Sum SE's each column separately.
18: Get the relationship matrix E_i between the history records of user i and other items.
19: Find all items in $g_{i,m}$. Sorting items. Get u_i's recommendation list $\mathrm{Re\,l}_i$
20: **end for**
21: Get all users' recommendation list Rel.

Algorithm 2: Short-term dynamic recommendation model.

Table 1: General statistics about the two datasets.

Network property	MovieLens-100K	MovieLens-10M
Size of dataset	5 M	10 M
Number of users	943	2113
Number of movies	1682	10197
Number of genres	19	20
Range of rating	1-5	1-5
Number of privacy items	4	0

$$\frac{P[f(k) = k^*]}{P\left[P_{c_i \to c_j} \times f\left(k'\right) = k^*\right]} \le e^{\varepsilon}. \tag{7}$$

The proof and derivation of local differential privacy after adding random replacement are as follows.

$$\begin{cases} b[\gamma\alpha + (1-\gamma)(1-\alpha)] = \dfrac{n_1}{n}, \\ b[(1-\gamma)\alpha + \gamma(1-\alpha)] = \dfrac{n_2}{n}, \\ n_1 + n_2 = n, \\ b = \dfrac{1}{n_u}\dfrac{1}{n_g - 1}. \end{cases} \tag{8}$$

γ is the true proportion. α is the added perturbation, which is the probability of random perturbation. n_1 is the number of males. n_2 is the number of females. n is the total. Construct the likelihood function.

$$L(\gamma) = [b\gamma\alpha + b(1-\gamma)(1-\alpha)]^{n_1}[b\gamma(1-\alpha) + b(1-\gamma)\alpha]^{n_2}. \tag{9}$$

Get logarithmic likelihood function and derivate.

$$\begin{gathered} \ln L(\gamma) = n_1 \ln [b\gamma(2\alpha - 1) + b(1-\alpha)] + (n - n_1) \ln [b\alpha - b\gamma(2\alpha - 1)], \\ \frac{d}{d\gamma} \ln L(\gamma) = \frac{(2\alpha - 1)[-n_1 + n\alpha(2\gamma - 1) - n\gamma + n]}{[\alpha(2\gamma - 1) - \gamma][\alpha(2\gamma - 1) + (1 - \gamma)]} = 0, \\ \widehat{\gamma} = \frac{\alpha - 1}{2\alpha - 1} + \frac{n_1}{(2\alpha - 1)n}, \\ E(\widehat{\gamma}) = \frac{1}{2(\alpha - 1)}\left[\alpha - 1 + \frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{2(\alpha - 1)}[\alpha - 1 + \gamma\alpha + (1-\alpha)(1-\gamma)] = \gamma. \end{gathered} \tag{10}$$

Suppose that N represents the estimated number of men selected by a certain item statistically:

$$N = \widehat{\gamma} \times n = \frac{\alpha - 1}{2\alpha - 1} n + \frac{n_1}{2\alpha - 1}. \tag{11}$$

In order to satisfy $\varepsilon -$ local differential privacy, the privacy budget ε is set as

$$\varepsilon = \ln \frac{\alpha}{1 - \alpha}. \tag{12}$$

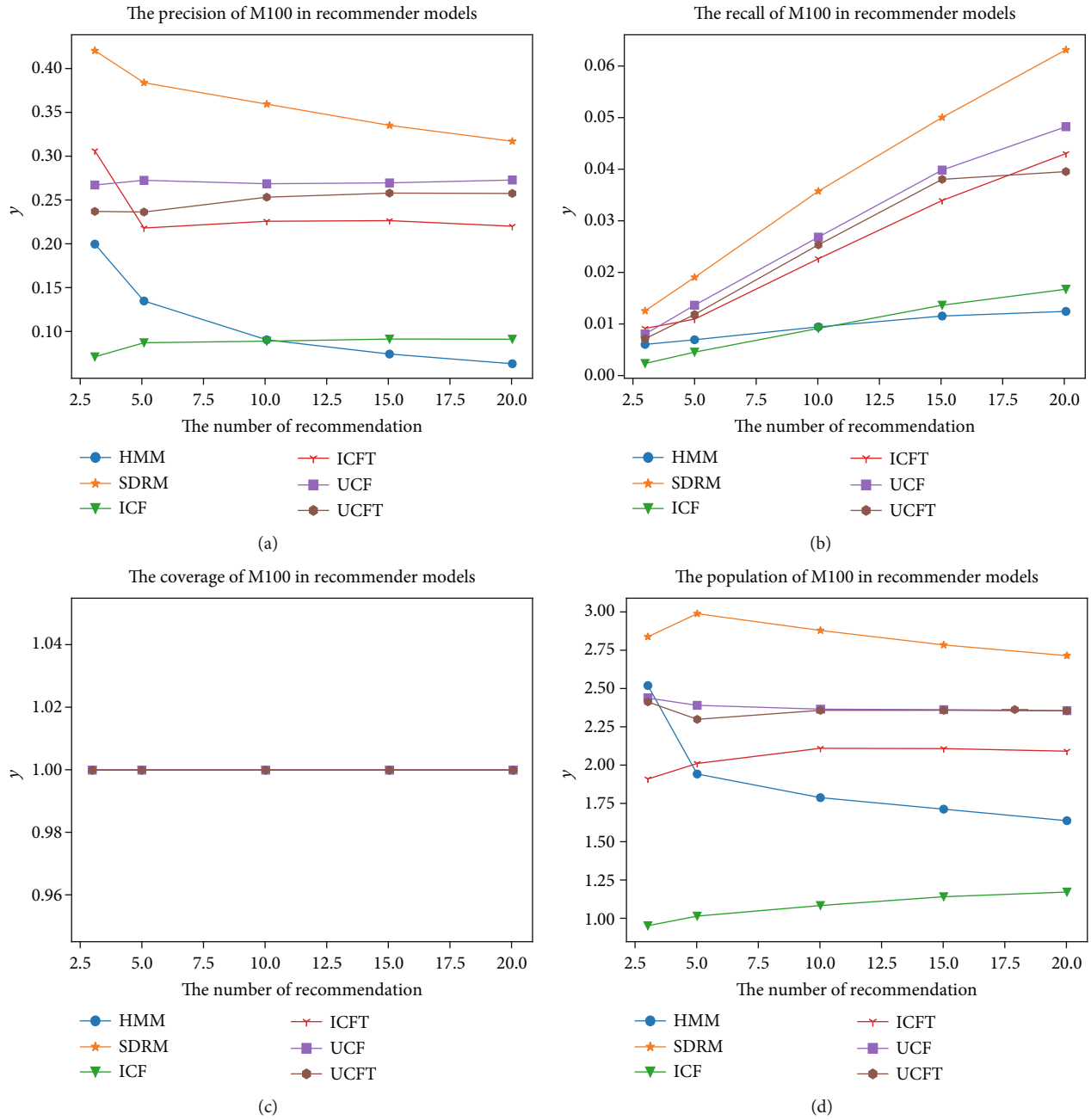


Figure 4: Continued.

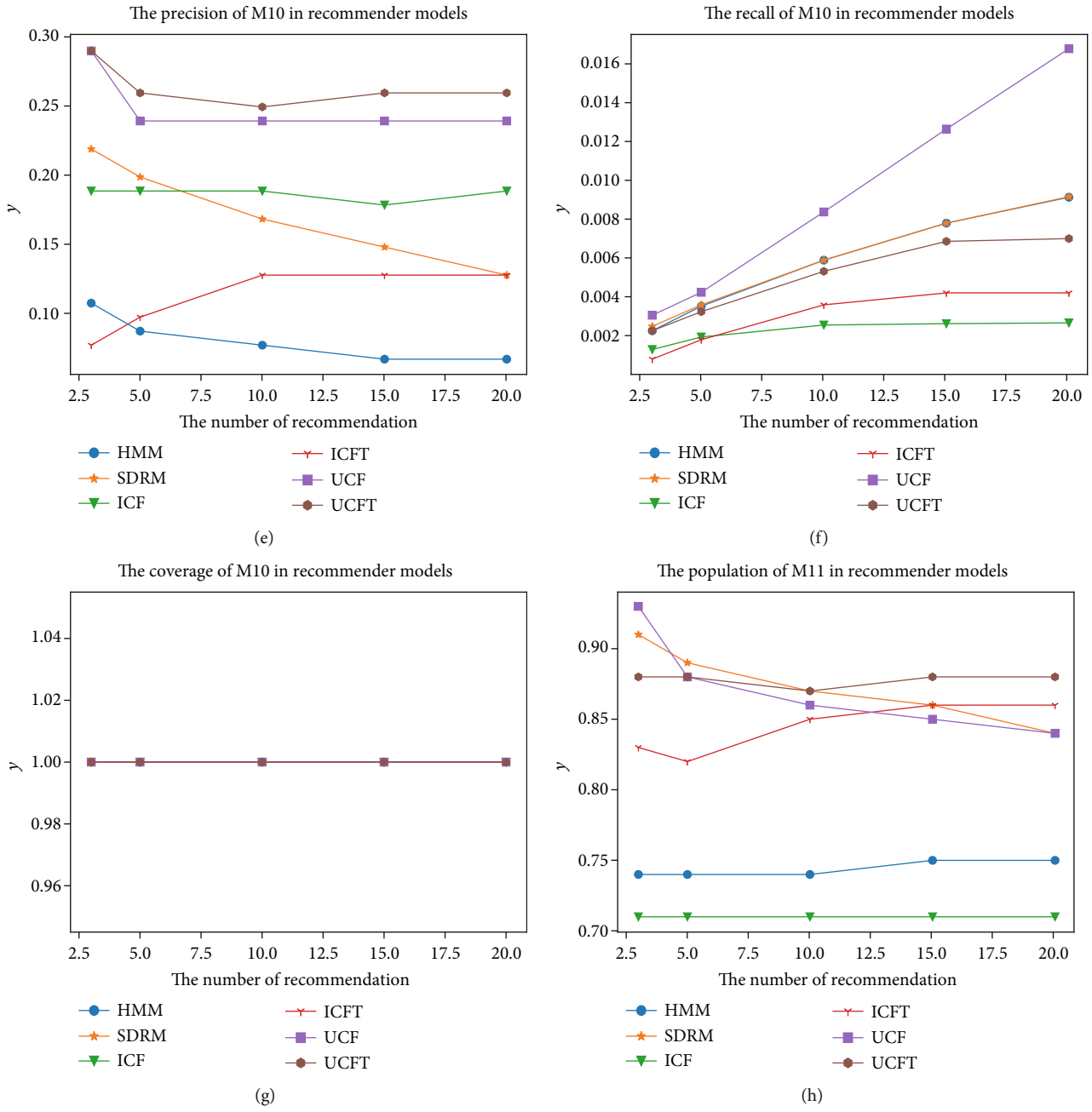
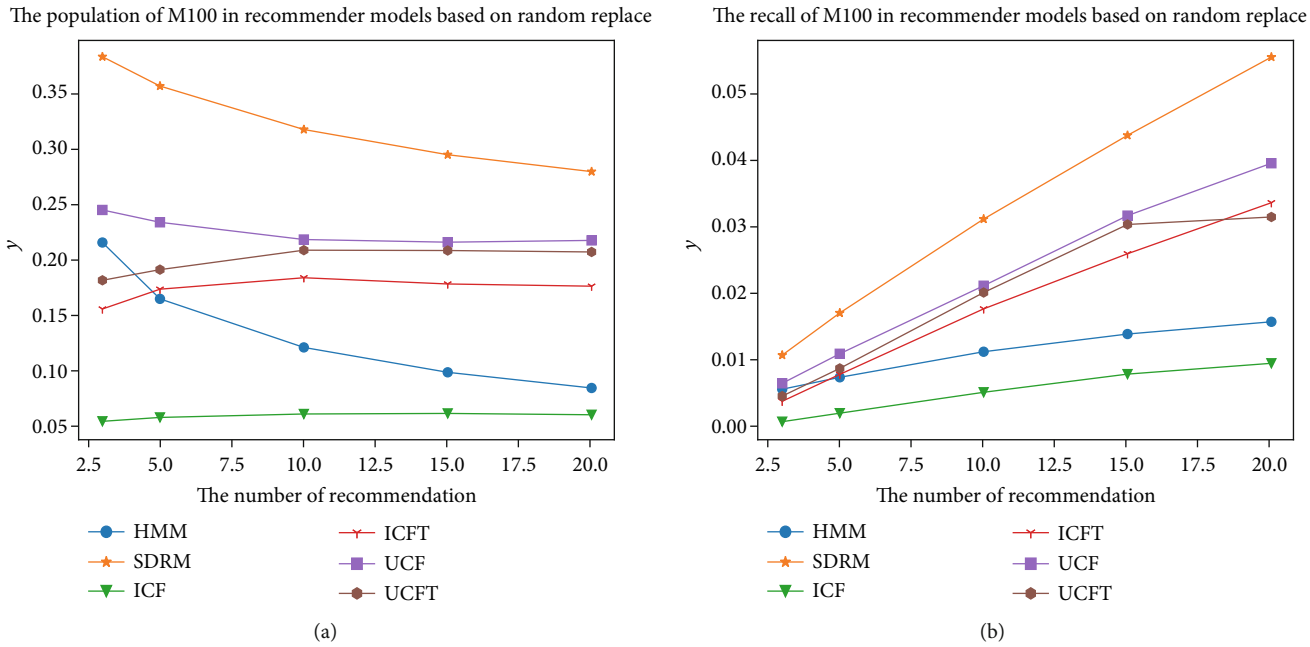


Figure 4: The recommendation result based on ML-100K dataset and ML-10M dataset. (a–d) Results of ML-100K dataset. (e–h) Results of ML-10M dataset.

5. SDRM-LDP Model

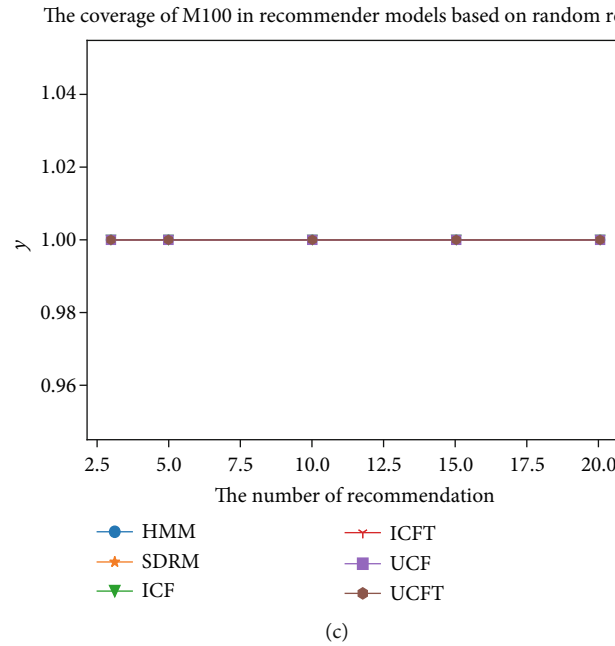
This part is the core of the thesis. We introduce the framework of the model and the specific algorithm. SDRM-LDP model is mainly composed of two parts. One part is the SDRM recommendation system, and the other is the SDRM-LDP privacy protection mechanism. The two parts are presented separately. The following figure shows the overall frame of this paper (see Figure 3).

In Figure 3, we take into account the problems of user data transmission speed and user data leakage under 5G, so we delete the data based on the time threshold before uploading the data. At the same time, in order to strengthen the security of user data and prevent attackers from deducing private information by using nonprivate data, we randomly replaced the nonprivate data before localized differential privacy. The processed data is uploaded to the central processor, and the prediction of the user's future behavior and the construction of matrix E are completed in the central processor. Finally, a list of recommendations for each user is given. The other side is the bottom half of the terminal. With the addition of LDP to user privacy items, we offer statistical queries to third parties. The protection of this part is combined with the deletion and disturbance of



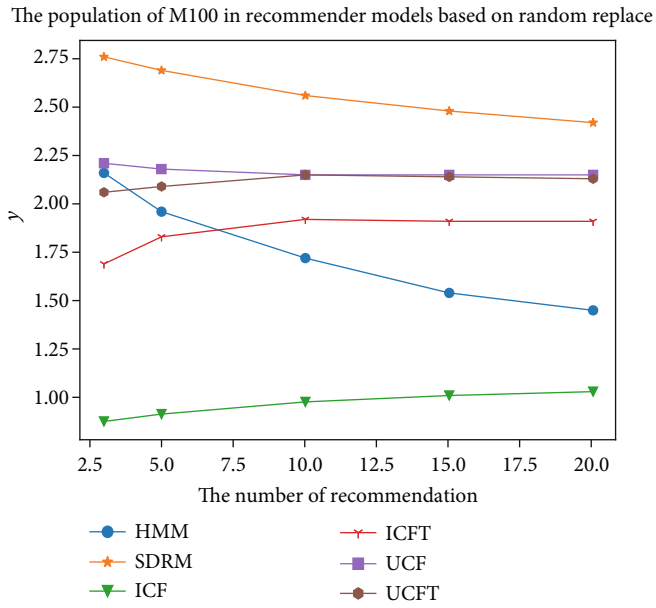
(a)

(b)

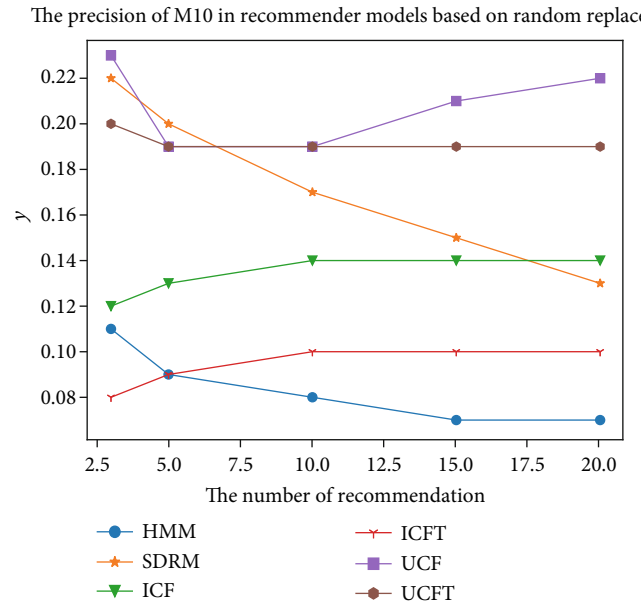


(c)

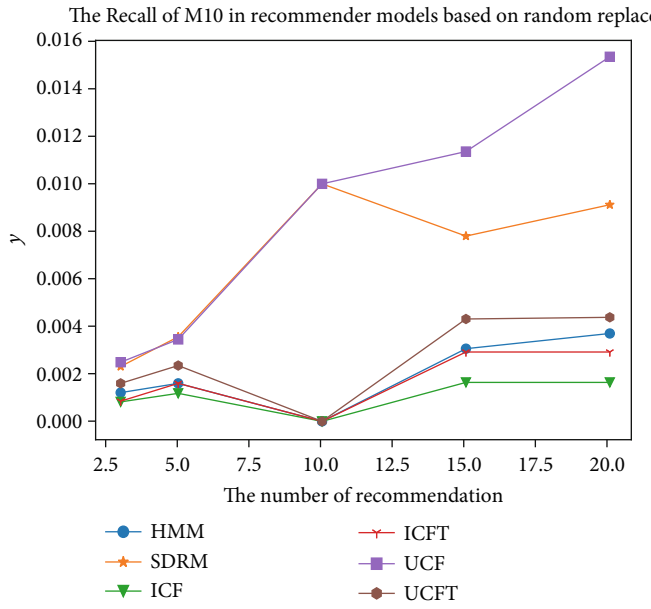
Figure 5: Continued.



(d)

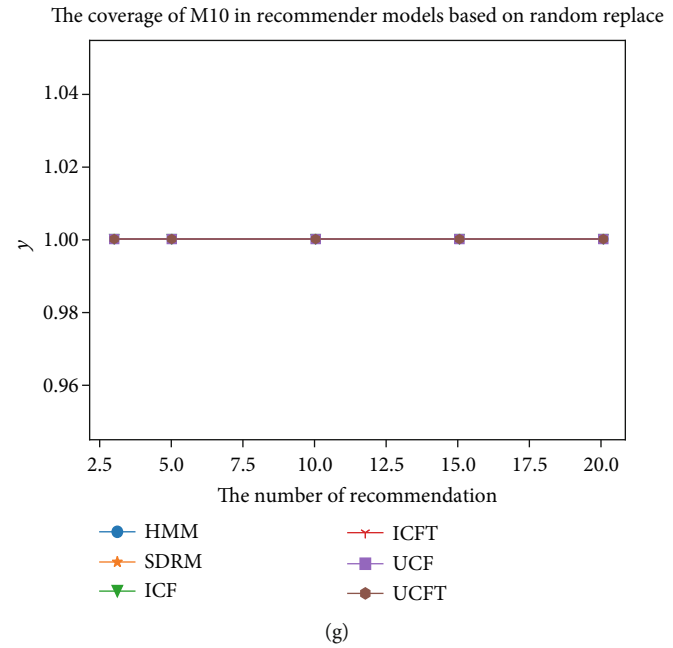


(e)

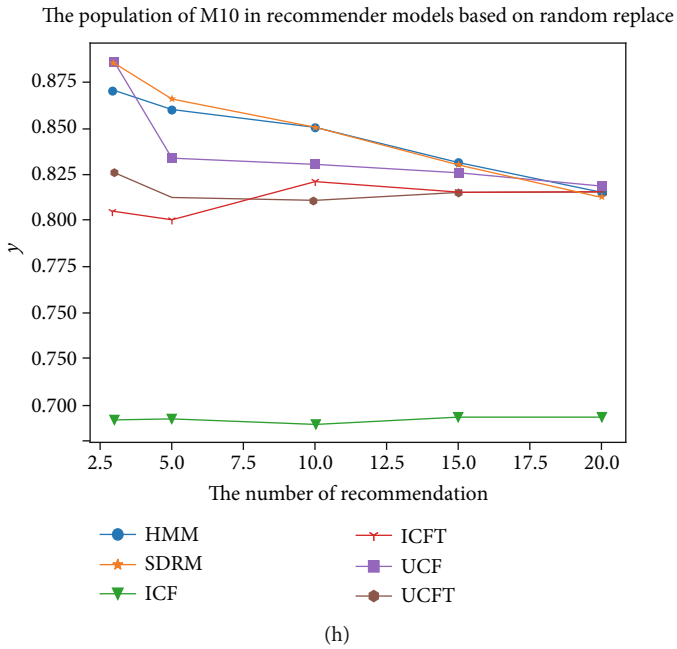


(f)

Figure 5: Continued.



(g)



(h)

Figure 5: The replaced recommendation result based on ML-100K dataset and ML-10M dataset. (a–d) Results of ML-100K dataset. (e–h) Results of ML-10M dataset.

the upper part. See Definition 8 for a detailed explanation of why this is considered.

5.1. Privacy Protection Mechanism. The data are processed from three perspectives: data source, data publication, and privacy query. The process includes data cleaning, filtering, replacement, and local differential privacy to protect the security of user data from all aspects. First, we process the raw data according to the time threshold and delete the data beyond the time threshold. There are two reasons. (1) Reduce user information from the raw data and reduce the risk of data leakage. (2) This model is the short-term behavior prediction of users. Too much history can lead to overfitting the results. The total number of dataset after processing becomes $t \times i$, which is the number of users times the number of retained historical records. The second is to randomly replace items in the user's history with random nonhistorical items of the same genre. The purpose of this operation is to interfere with the attacker to accurately obtain nonsensitive information. The third step is to query the statistical results of privacy for the third party. Based on the above two steps, we add random disturbance to the user's privacy. It satisfies local differential privacy, which has been proved in the previous section. This step is intended to protect users'

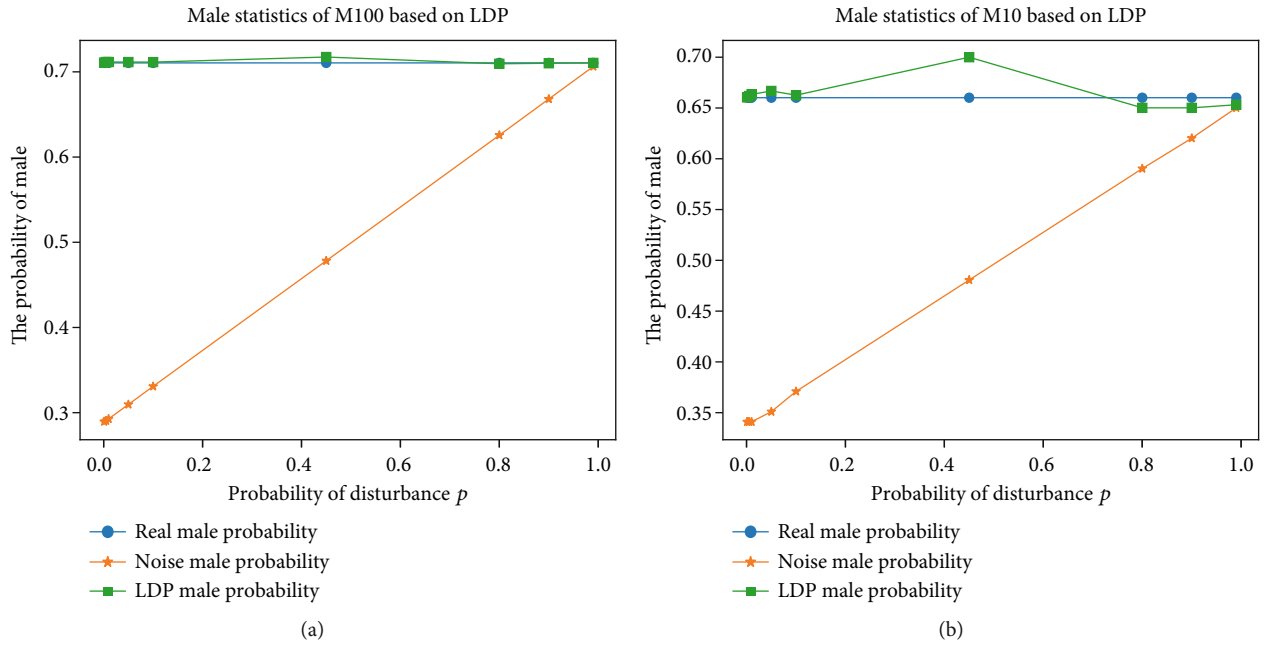


Figure 6: The results based on LDP replace on ML-100K and ML-10M.

information during private queries and prevent reverse inferences by attackers. The algorithm is as follows (see Algorithm 1).

5.2. SDRM. This paper explores users' future preferences based on their short-term behaviors. The relationship matrix between items and users is constructed based on user history to solve the relationship between items and genres in the future. First, according to Definition 5, we construct the general $\widehat{\lambda}$. $\widehat{\lambda}$ could solve cold start for the new users. Based on $\widehat{\lambda}$, we adjust each user based on the retained historical records and get λ_i for u_i. λ_i is used to predict the user's future preferences, which are the genres the user's most likely to choose in the next state and find all items in these genres in E and sort them to get the recommendation list for u_i. The recommendation algorithm is as follows (see Algorithm 2).

6. Evaluation

6.1. Datasets. The experimental datasets in this paper are all from the real world. One is the Movielens-100K dataset and the other is the MovieLens-10M dataset. Because the experiment involves user privacy information, but MovieLens-10M lacks this part, we fill the part with random data. The information involved in the two datasets is collated in Table 1.

6.1.1. MovieLens-100K. The dataset is a virtual community site that recommends movies to users based on their preferences. It is a project of Grouplens Research LABS. The MovieLens dataset in this paper is the MovieLens-100K dataset from Grouplens [19]. The dataset is 5 MB in size and contains 100,000 comments, 943 users, 1682 movies, and at least 20 movies per user. The rating range is 1-5. Only integer rating is allowed. The dataset also includes some privacy of users, such as age, sex, occupation, and zip code. The data are collected from September 19, 1997, to April 22, 1998. The dataset is divided into 5 cross-validation sets on a scale of 8 : 2.

6.1.2. MovieLens-10M [20]. In this dataset, users can comment on items they have purchased or used in the past. Other users can rate comments based on how useful the comments are. This part of the data is not required for this paper. The crawling period of this dataset is from May 2011. The dataset size is 10 M and has 2113 users and 10,197 films. The features include user ID, movie ID, genre, rating, comments, and time. The dataset lacks user privacy information. We use gender as the target information for privacy queries. Therefore, gender is randomly generated for users to complete the privacy inquiry part of the experiment. The experimental results in this part are for reference only. In fact, the randomly generated gender data has little effect on the local differential privacy experiment because this part of local differential privacy in the derivation is consistent with its original definition. If the ML-100K's results are consistent with our definition of privacy protection, we can assume that the ML-10M results are consistent in principle. Privacy information such as gender has no effect on other parts of the experiment except for the statistical query.

6.2. Experiment Settings. In this part, relevant settings are given for the experiment. The setting includes parameters, division ratio of training set and test set, evaluation metrics of recommendation system, and privacy protection.

6.2.1. Training Dataset and Test Dataset. The Movielens-100K has 943 users and the Movielens-10M has 2113 users. Each dataset is based on the real users. The users are randomly selected in a ratio of 8 : 2 to constitute the training set and the test set. The data is sorted by each user's history

time. Because each user has at least 20 records, the chronological order is separated by the number of time thresholds. For example, if the time threshold is set to 5, 3-5 historical records are kept to evaluate the final prediction results. So, we select the latest 5 as the training dataset of the model.

6.2.2. Basic Parameter Settings. Because each movie in the dataset corresponds to multiple categories, to reduce the computational complexity, we integrate multiple categories into one and rearrange them to form new categories. ML-100K is sorted into 218 new categories and ML-10M into 718 new categories. t_0 is the latest date, which is the last date of data sorted by times as a reference data. The number of CF neighbors is 5. The kept states in SDRM are 5. β is 2. σ is 1.6. The number of recommended result is 3, 5, 10, 15, and 20. α in the privacy query based on LDP is 0.001, 0.005, 0.01, 0.05, 0.1, 0.45, 0.8, 0.9, and 0.99.

6.2.3. Evaluation Metrics. The overall structure of this paper is divided into two parts: one is the recommendation system, and the other is the privacy protection algorithm. Therefore, corresponding evaluation indexes are given for different parts of the evaluation. The evaluation of a recommendation system is mainly divided into online evaluation and offline evaluation. Online evaluation is generally to evaluate the problems involved in the actual scenario, such as click rate and conversion rate. Obviously, this is not appropriate for this paper, so we choose to use offline evaluation. A total of 4 offline evaluation indexes are selected: accuracy rate, recall rate, coverage rate, and novelty. Accuracy is the percentage of the predicted results that are correct. The recall rate measures how many of all the correct outcomes are predicted. Accuracy and recall rates measure the accuracy of a recommendation algorithm. Coverage is a description of a recommendation system's ability to find the long tail of an item. The higher the coverage rate, the more long-tail items are recommended to users. Novelty refers to the proportion of items recommended to users that they have not heard of. A new method for measuring glutenness is to use the average popularity of the recommended results, because the less popular the item, the more novel the user thinks it is. The specific formulas for accuracy, recall rate, and coverage are as follows:

$$\begin{aligned} \text{Precision} &= \frac{\sum_{u \in U} | R(u) \cap T(u) |}{\sum_{u \in U} | R(u) |}, \\ \text{Recall} &= \frac{\sum_{u \in U} | R(u) \cap T(u) |}{\sum_{u \in U} | T(u) |}, \\ \text{Coverage} &= \frac{U_{u \in U} R(u)}{I}. \end{aligned} \tag{13}$$

6.3. SDRM Results. We compared SDRM with general HMM-RS, item-based collaborative filtering, user-based collaborative filtering, item-based collaborative filtering with time factor, and user-based collaborative filtering with time context. Item-based collaborative filtering and user-based collaborative filtering can be regarded as the baseline. SDRM was compared to two models with added time context, because both of them and our algorithms take time into account. In relevant work, we mentioned that some scholars also use HMM to make a recommendation system, but our HMM is an improvement, and we compare the two methods. We use four indicators of accuracy, recall rate, coverage rate, and population to evaluate the recommendation models. We use two real network datasets ML-100K and ML-10M for the experiment. The results are shown in Figure 4. Figures 4(a)–4(d) are the comparison results on the ML-100K dataset. The results show that SDRM is superior to other algorithms in the precision rate. The recall rates are similar except for the ICF and ICFT. The coverage of each algorithm is basically the same. SDRM is superior to other algorithms in terms of popularity, which is more important to the recommended. SDRM is superior to other algorithms in terms of overall performance.

Figures 4(e)–4(h) are the recommended results of ML-10M. The results show that UCF results are relatively high among these and SDRM is only inferior to UCF. The main reason for this is that ML-100K and ML-10M use the same parameter setting but ML-100M is much larger than ML-100K in both the number of movies and the number of movie genres. With no more parameters, the results of SDRM are slightly lower than UCF but higher than HMM and ICF.

The results shown in Figure 5 are random substitutions of some items in the original data within the same genre. Figures 5(a)–5(d) are the prediction results of randomly replacing one item in the training data of each user in ML-100K. In the figure below, SDRM is far superior to other recommended models in accuracy, recall rate, and popularity. Figures 5(e)–5(h) are the results of random replacement based on dataset ML-10M. In Figures 5(e) and 5(f), when the number of recommendations is relatively small, the accuracy and recall rate of SDRM are higher than other recommended models. SDRM's popularity is better than other models. The interpretation for this result is basically the same as the original data recommendation that the parameters are not adjusted according to the dataset to be optimal for SDRM. However, SDRM is more stable when the data part is changed, which also indicates that the model is indeed a prediction based on user preferences.

6.4. SDRM-LDP Results. Figure 6 is the comparison result of LDP privacy query based on random replacement under ML-100K and ML-10M datasets, respectively. From the results, the LDP-based query results are pretty much the same as the real result. This evidence shows that while users' privacy is protected, the utility availability of the data is still relatively high, with little loss. The start line is the result of uncorrected, only after adding noise. The LDP-based privacy protection works better according to the experiments.

7. Conclusions

In this paper, we discuss the following issues: (1) user preference prediction and recommendation in dynamic conditions and (2) how to provide recommendations while protecting user privacy. For the first issue, we propose the SDRM model. The model is not a traditional recommendation system based on similarity. This model uses HMM and graph to find the

dynamic preference model belonging to each user. Genre-based forecasting reduces the amount of computation required. At the same time, with changes to certain data, the prediction result is more stable. In general, the traditional recommendation model is too superficial, and every data has an impact on future predictions. The SDRM explores the selection of users' internal preferences. Therefore, a fraction of data changed has little impact on the predicted results. For the second issue, we query the probability of each user's privacy item based on random replacement items. The process is based on LDP, and the experiment results show that the noise adding does not change the performance in significant value, so the privacy of users is protected without affecting the utility of the data. After the random replacement, the data and the privacy query results are all the data we released to the public. SDRM-LDP guarantees the information security of users and reduces the risk of attacks. Both the final results of the recommendation model and the privacy protection model are satisfactory.

Data Availability

The data used to support the findings is generally unavailable due to public releasability constraints. Please contact the corresponding author for special release consideration.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Our research fund is funded by the Fundamental Research Funds for the Central Universities (3072020CFQ0602, 3072020CF0604, and 3072020CFP0601) and 2019 Industrial Internet Innovation and Development Engineering (KY10600200021 and KY10600200008).

References

[1] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 766–775, 2020.

[2] Z. Cai, X. Zheng, and J. Yu, "A differential-private framework for urban traffic flows estimation via taxi companies," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 12, pp. 6492–6499, 2019.

[3] M. B. Holbrook and R. M. Schindler, "Some exploratory findings on the development of musical tastes," *Journal of Consumer Research*, vol. 16, no. 1, pp. 119–124, 1989.

[4] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Transactions on Dependable and Secure Computing*, p. 1, 2016.

[5] Z. Cai and Z. He, "Trading private range counting over big IoT data," *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 2019, Dallas, TX, USA, 2019.

[6] Y. Wang, G. Yin, Z. Cai, Y. Dong, and H. Dong, "A trust-based probabilistic recommendation model for social networks," *Journal of Network and Computer Applications*, vol. 55, pp. 59–67, 2015.

[7] N. H. M. H. Aghdam and B. Mobasher, "Adapting recommendations to contextual changes using hierarchical hidden Markov models," *RecSys '15: Proceedings of the 9th ACM Conference on Recommender Systems*, pp. , 2015241–244, 2015.

[8] S. Cleger Tamayo, L. A. Hernandez Leyva, L. W. Osorio Gamez, and A. L. Scull Pupo, "Temporal dynamic study in personalization digital newspaper ahora!," *IEEE Latin America Transactions*, vol. 13, no. 8, pp. 2792–2797, 2015.

[9] N. Sahoo, P. V. Singh, and T. Mukhopadhyay, "A hidden Markov model for collaborative filtering," *MIS Quarterly*, vol. 36, no. 4, pp. 1329–1356, 2012.

[10] V. E. Epure, B. Kille, E. J. Ingvaldsen, R. Deneckere, C. Salinesi, and S. Albayrak, "Recommending personalized news in short user sessions," *RecSys*, pp. , 2017121–129, 2017.

[11] W. Huang, B. Liu, and H. Tang, "Privacy protection for recommendation system: a survey," *Journal of Physics: Conference Series*, vol. 1325, p. 012087, 2019.

[12] X. Chi, C. Yan, H. Wang, W. Rafique, and L. Qi, "Amplified locality-sensitive hashing-based recommender systems with privacy protection," *Concurrency and Computation Practice and Experience*, no. 1, 2020.

[13] F. J. Garcia Clemente, "A privacy-preserving recommender system for mobile commerce," *2015 IEEE Conference on Communications and Network Security (CNS)*, 2015, pp. 725-726, Florence, Italy, 2015.

[14] Z. He, Z. Cai, J. Yu, X. Wang, Y. Sun, and Y. Li, "Cost-efficient strategies for restraining rumor spreading in mobile social networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2789–2800, 2017.

[15] T. S. Al-Nazzawi, R. M. Alotaibi, and N. Hamza, "Toward privacy protection for location based recommender systems: a survey of the state-of-the-art," in *2018 1st International Conference on Computer Applications & Information Security (ICCAIS)*, Riyadh, April 2018.

[16] Z. He, Z. Cai, and J. Yu, "Latent-data privacy preserving with customized data utility for social network data," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 665–673, 2018.

[17] L. R. Welch, "Hidden Markov models and the Baum-Welch algorithm," *IEEE Information Theory Society Newsletter*, vol. 53, no. 2, pp. 194–211, 2003.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the *EM* algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–22, 1977.

[19] https://grouplens.org/datasets/movielens/100k/.

[20] https://grouplens.org/datasets/hetrec-2011/.