Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 188 (2021) 61–68

Procedia Computer Science

www.elsevier.com/locate/procedia

CQVIP Conference on Data Driven Intelligence and Innovation

Privacy Recommendation Based on Bhattacharyya Coefficient

Yong Wang[a,b], Li Wang[a], Ling Zhao[a], Xun Ran[a], Siyuan Deng[a*]

[a]*Key Laboratory of Electronic Commerce and Logistics, Chongqing University of Posts and Telecommunications, Chongqing and 400065, China*

[b]*Guangxi Key Laboratory of Cryptography and Information Security, Guilin University of Electronic Technology, Guilin and 541004, China*

Abstract

In the era of big data, how to achieve the trade-off between recommendation accuracy and privacy protection has become a hot topic in the field of recommendation. In this work, a recommendation scheme with personal privacy preservation is designed. Besides, it does not depend on co-ratings. In order to solve the problem of excessive sensitivity when the Laplace noise is introduced directly, the Bhattacharyya coefficient that is used as the similarity metric is normalized. Moreover, a personalized privacy protection scheme that considering the difference of users is used for the recommendation system. Experimental results show that, compared with the traditional differential privacy collaborative filtering scheme, the RMSE and MAE of the proposed scheme are significantly improved. The proposed scheme not only ensures the accuracy of the prediction results but also effectively guarantees the privacy protection of user data. It also provides an insight into the design of recommendation services considering privacy protection.

Peer-review under responsibility of the scientific committee of the CQVIP Conference on Data Driven Intelligence and Innovation (CCDDII2021)

Keywords: Differential privacy; Privacy protection; Collaborative filtering; Recommendation system; Bhattacharyya coefficient

*Corresponding author at: Key Laboratory of Electronic Commerce and Logistics, Chongqing University of Posts and Telecommunications
E-mail address: wangyong_cqupt@163.com;CQwangli996@163.com

1877-0509
Peer-review under responsibility of the scientific committee of the CQVIP Conference on Data Driven Intelligence and Innovation (CCDDII2021)
10.1016/j.procs.2021.05.053

1. Introduction

In the era of big data, the phenomenon of "information overload" is becoming more and more serious. It is necessary to provide personalized recommendation services for consumers from mass information [1]. Recommendation systems based on collaborative filtering can help users quickly discover the items they are interested in by analyzing the historical behavior of users. Today, recommendation system has been widely applied in e-commerce and social networking [2]. However, the computation in the recommendation system needs a large amount of user information, which increases the risk of user information leaked [3, 4]. Therefore, how to ensure user privacy security while maintaining high-quality recommendations has become a hot topic in current research.

As a privacy protection model, differential privacy protects data security and also takes into account the utility of data [5]. At present, differential privacy protection is widely used in recommendation systems due to its easy combination and simple implementation [6]. McSherry et al. [7] divided the recommendation system into the learning stage and the prediction stage. In the learning stage, noise implementation interference is added to the item similarity covariance matrix. Zhu et al. [8] disturbed the similarity after neighbor selection to reduce the probability that the attacker speculated the similarity between users and items, and applied the index mechanism to neighbor selection to prevent the attacker from speculating "who is the user". Hua et al. [9] prevented untrusted recommenders from using users' rating information, then they adopt the objective function perturbation to achieve differential privacy protection. Yang et al. [10] solved the problem of data sparsity through project clustering by K-means clustering algorithm and introduced the Laplace mechanism to protect the rating matrix. Yin [11] proposed a recommendation algorithm based on differential privacy protection and time factor.

There are two main problems in the existing studies. One is that the similarity measurement method is difficult to calculate the similarity without common ratings, which leads to the introduction of excessive noise in the privacy protection algorithm and affects the validity of the recommendation. Second, the difference in users' privacy requirements has not been fully considered. In fact, different users have different demands for privacy protection [12, 13]. Unified privacy standards are easy to introduce unnecessary noise, resulting in the decline of recommendation quality. In order to address these problems, a new scheme considering the Bhattacharyya Coefficient (BC) and personalized differential privacy is proposed. First, in order to address the problem of the lack of common ratings in traditional similarity measurement methods, the BC is introduced as the similarity measurement standard. However, the direct introduction of BC into the privacy protection algorithm has the problems of high sensitivity and excessive noise, which degrades the effectiveness of the algorithm. Therefore, the BC similarity is reformed for privacy protection. Based on this, different privacy budgets are designed for different users according to their various requirements for privacy protection, which effectively improves the accuracy of the recommendation system. By improving the accuracy of recommendations and reducing the amount of noise, the proposed scheme enables the recommendation system to provide accurate recommendation services for users while guaranteeing the security of users' privacy information.

2. Our Algorithm

Collaborative filtering recommendations based on neighbor selection can provide users with high-quality recommendation services. However, the literature [8] shows that the attacker can infer the user's preferred items by checking the list of related items, resulting in privacy leakage. In addition, when the data is sparse, it is easy to fail to calculate the similarity due to the lack of common ratings [14]. In this paper, to prevent the related item list inference attack and improve the performance of the algorithm, similarity based on BC is used as the similarity measure. By introducing the Laplace noise into the similarity based on BC, the risk of privacy leakage can be prevented. At the same time, the method of Yang et al. [12] is used for reference, and different privacy budgets are allocated according to the different privacy requirements of users, so as to provide personalized privacy protection and ensure the accuracy of the recommendation algorithm more effectively.

2.1. Similarity calculation and normalization

Sensitivity is a key parameter in determining the levels of noise added to the recommendation system [15]. It refers to the maximum change to the calculation result caused by changing any record in the dataset. In the function of BC similarity $S(u,v)$, the values of the formula $\mathrm{loc}(r_{ui}, r_{vj})$ have no clear upper and lower bound, which means when the product of σ_u and σ_v is infinitesimal. The value of $\mathrm{loc}(r_{ui}, r_{vj})$ may be close to infinity or infinitesimal, which lead to the calculation of $\sum_{i\in I_U}\sum_{j\in I_V} BC(i,j)loc\left(r_{ui}, r_{vj}\right)$ may be quite different. Eventually, result in the sensitivity of $S(u,v)$ is too high to introduce differential privacy to protect its privacy.

Therefore, in order to successfully introduce the differential privacy into the similarity measurement based on BC and reduce the sensitivity of similarity calculation. The similarity formula based on BC is processed. The processing formula is present as follows:

$$S'(u,v)=\frac{S(u,v)-\min\limits_{v} S(u,v)}{\max\limits_{v} S(u,v)-\min\limits_{v} S(u,v)} \tag{1}$$

Where $\max\limits_{v} S(u,v)$ represent the greatest BC similarity between user u and other users, $\min\limits_{v} S(u,v)$ represent the minimum BC similarity between user u and other users.

Since the normalized similarity value range based on BC is [0, 1], the global sensitivity of the utility function is:

$$GS(S)=\max_{D,D'}\left\|S'_{D}(u,v)-S'_{D'}(u,v)\right\|=1 \tag{2}$$

2.2. Algorithm description

Referred to the method in literature [12], users are randomly divided into three groups: low privacy concern users, medium privacy concern users and high privacy concern users. And the corresponding allocation ratio is P_l、P_m and P_h. For the low privacy concern users, the privacy budget is ε_L, for users with medium and high privacy concerns, preferences are uniformly sampled in the range of $[\varepsilon_l, \varepsilon_m]$ and $[\varepsilon_m, \varepsilon_h]$ respectively. $\varepsilon_l, \varepsilon_m, \varepsilon_h \in [1,10]$. Smaller privacy budget values indicate stronger privacy requirements. Specific parameter settings are shown in the following table:

Table1. Parameter of privacy distribution

parameter	describe	Values
P_l	Ratio of low privacy concern users	0.09
P_m	Ratio of medium privacy concern users	0.37
P_h	Ratio of high privacy concern users	0.54
ε_l	Lower bound of the privacy parameter	1
ε_m	Median value of privacy parameters	3
ε_h	Upper bound on the privacy parameter	10

By introducing differential privacy to protect similarity, privacy leakage can be prevented [16]. The different privacy protection is provided for different users according to their privacy requirements. This method can not only

protect the user's privacy reasonably but also improve the recommendation performance. The specific steps of the algorithm shown as follows:

Table 2. Algorithm description

Algorithm: Privacy protection recommendation algorithm based on Bhattacharyya coefficient similarity(PDP-BC)

Input: dataset $D \in R^{m \times n}$, user's privacy budget parameter $\varepsilon_{\varphi}(u)$, number of neighbors K

Output: predicted rating $\tilde{r}_{u,i}$

1 for each user u do:

2 for each user v $(u \neq v)$ do:

3 calculate $S(u,v) = J\mathrm{acc}(u,v) + \sum_{i \in I_U} \sum_{j \in I_V} BC(i,j) loc(r_{ui}, r_{vj})$ 4 end for

5 calculate $\max_v S(u,v), \min_v S(u,v)$

6 for each user v $(u \neq v)$ do:

7 $$S'(u,v) = \frac{S(u,v) - \min_v S(u,v)}{\max_v S(u,v) - \min_v S(u,v)}$$

8 $$S''(u,v) = S'(u,v) + Lap(\frac{GS(S)}{\varepsilon_{\varphi}(u)})$$

9 end for

10 sort $S''(u,v)$ from largest to smallest

11 select K -nearest neighbors

12 end for

13 for each user u do:

14 for user v' in top K neighbors do:

15 $$\tilde{r}_{u,i} = \bar{r}_u + \frac{\sum_{v'=1}^{k} [S''(u,v')(r_{v',i} - \bar{r}_{v'})]}{\sum_{v'=1}^{k} S''(u,v')}$$

16 end for

17 end for

3. Algorithm Analysis

3.1. The utility analysis

In the collaborative filtering recommendation algorithm based on neighbors, similarity not only affects the selection of neighbors but also plays a key role in the final rating prediction. Using the BC similarity to find the distribution law of user ratings can effectively improve the traditional similarity measurement method over-reliance on common ratings. In the Movie-Lens dataset, for example, co-ratings accounted for only 4% of all ratings. This indicates that the effectiveness of the similarity measurement method based on common ratings is very low. Because the BC similarity is not restricted to common ratings, the utilization rate of ratings can reach 100%. In addition, to reasonably

protect the privacy of users in the recommendation system, two measures approaches are proposed. Firstly, to solve the problem that the similarity sensitivity of BC similarity is too high, the similarity formula is normalized. Then, a scheme of personalized differential privacy is introduced. The privacy budget is randomly allocated to users according to Section 2.2. By appropriately reducing the privacy budget and focusing on the privacy protection levels of users, the introduction of noise can be reasonably reduced and the accuracy of recommendation performance can be improved.

3.2. Safety analysis

Theorem 1: Given a dataset D, if user $u_i \in D$ has privacy budget $\varepsilon_\varphi(i)$, the proposed algorithm can provide user $u_i \in D$ with personalized differential privacy.

Proof: Assume the user's privacy budget parameter is $\varepsilon_\varphi(i)$, $GS(S)$ is global sensitivity of the similarity of the Bhattacharyya coefficient, $S_D'(u,v)$ represent the algorithm query function, and D' is the neighbor dataset. The noise added to the similarity is $N(x)$ and it obeys the Laplace distribution $Laplapce(0,b)$. According to the definition of differential privacy, we can get formulation as follows:

$$\frac{\Pr[S_D'(u,v)=s]}{\Pr[S_{D'}'(u,v)=s]} = \exp\left(\frac{\left|s-S_{D'}'(u,v)\right|-\left|s-S_D'(u,v)\right|}{b}\right) \le \exp\left(\frac{\left|S_{D'}'(u,v)-S_D'(u,v)\right|}{b}\right) \le \exp\left(\frac{GS(S)}{b}\right) = e^{\varepsilon_\varphi(i)}$$

Therefore, the proposed method could satisfy the privacy requirements of different users. Simultaneously, it could provide corresponding privacy protection for every user. Since no new noise is introduced in the subsequent data processing of the algorithm, when the similarity introduces noise satisfies the differential privacy, the whole algorithm also satisfies the differential privacy.

4. Experimental Results and Analysis

4.1. Selection of Datasets

The algorithm is tested on Movie-Lens-100K and Yahoo Music datasets. Movie-Lens-100K contains 100,000 ratings for 1,682 movies from 943 users, each user rated at list 20 movies. Yahoo Music consists of 270,000 pieces of music rated by 8,089 users on a scale of 1-5. In order to ensure the stability of the results, the ratio of the train set to the test set is 4∶1. For the other comparison algorithms involved, the privacy budget parameter defaults to $\varepsilon = 1$. To reduce the impact of the randomness of differential privacy on the experimental results, all algorithms take the average value of five experiments as the final experimental results.

The Root Mean Square Error (RMSE) and Mean Square Error (MAE) are commonly used as the indicator of prediction accuracy. The calculation formula as follows:

$$RMSE = \sqrt{\frac{\sum_{u,i}(r_{u,i}-\tilde{r}_{u,i})^2}{N}} \qquad MAE = \frac{1}{N}\sum_{i=1}^{n}\left|r_{u,i}-\tilde{r}_{u,i}\right|$$

Where N represent the total number of valid ratings, $\tilde{r}_{u,i}$ is the predicted rating.

4.2. Analysis of experimental results

In this study, four groups of experiments are conducted to analyze the performance of our scheme. The following comparison algorithms are selected: 1) BCCF algorithm [14]: Recommendation algorithm based on Bhattacharyya coefficient similarity without any differential privacy protection. 2) DP-BC algorithm: In the recommendation algorithm based on Bhattacharyya coefficient similarity, a uniform privacy budget is added to the similarity without considering the personalized differential privacy. 3) DP-PCC algorithm: the traditional user-based differential privacy collaborative filtering algorithm, and the similarity calculation method is Pearson. 4) DP-COS algorithm: the traditional user-based differential privacy collaborative filtering algorithm, the similarity calculation method is COS. 5) PPCF [12] algorithm: a personalized differential privacy scheme based on Laplace mechanism, which carries out differential privacy protection for similarity and carries out Johnson-Lindenstrauss transformation on the original data. 6) PSGD [17] algorithm: a typical algorithm combining differential privacy with matrix factorization recommendation.

The experimental results are shown in Table 3.

Table 3 the experimental results of different algorithms

Algorithm	dataset	index	K-nearest neighbor				
			20	40	60	80	100
BCCF	Movie-Lens-100K	RMSE	1.0430	1.0162	1.0091	1.0019	0.9970
		MAE	0.8103	0.7896	0.7857	0.7807	0.7780
	Yahoo Music	RMSE	1.3809	1.3306	1.3018	1.2898	1.2817
		MAE	1.0883	1.0474	1.0242	1.0138	1.0071
DP-BC	Movie-Lens-100K	RMSE	1.1400	1.0851	1.0590	1.0387	1.0300
		MAE	0.8936	0.8502	0.8308	0.8150	0.8082
	Yahoo Music	RMSE	1.5493	1.4848	1.4448	1.4190	1.3931
		MAE	1.2233	1.1718	1.1393	1.1184	1.0978
DP-PCC	Movie-Lens-100K	RMSE	1.2098	1.1532	1.1251	1.1075	1.0944
		MAE	0.9424	0.9060	0.8877	0.8763	0.8671
	Yahoo Music	RMSE	1.9432	1.8832	1.8445	1.8118	1.7850
		MAE	1.5061	1.4641	1.4393	1.4202	1.4046
DP-COS	Movie-Lens-100K	RMSE	1.2088	1.1514	1.1201	1.1034	1.0921
		MAE	0.9407	0.9046	0.8828	0.8719	0.8648
	Yahoo Music	RMSE	1.8749	1.8063	1.7593	1.7227	1.6932
		MAE	1.4877	1.4430	1.4149	1.3937	1.3771
PDP-BC	Movie-Lens-100K	RMSE	**1.0993**	**1.0528**	**1.0307**	**1.0191**	**1.0116**
		MAE	**0.8584**	**0.8233**	**0.8069**	**0.7979**	**0.7925**
	Yahoo Music	RMSE	**1.5193**	**1.4511**	**1.4120**	**1.3856**	**1.3647**
		MAE	**1.1971**	**1.1428**	**1.1120**	**1.0910**	**1.0745**
PPCF	Movie-Lens-100K	RMSE	1.1938	1.1355	1.1081	1.0902	1.0783
		MAE	0.9344	0.8967	0.8780	0.8659	0.8576
	Yahoo Music	RMSE	1.8769	1.8095	1.7654	1.7295	1.6995
		MAE	1.4889	1.4445	1.4184	1.3979	1.3814
PSGD	Movie-Lens-100K	RMSE	1.1442	1.1442	1.1442	1.1442	1.1442
		MAE	0.8965	0.8965	0.8965	0.8965	0.8965
	Yahoo Music	RMSE	1.5060	1.5060	1.5060	1.5060	1.5060
		MAE	1.2363	1.2363	1.2363	1.2363	1.2363

(1) Influence of personalized differential privacy on recommendation performance

By compared with the DP-BC algorithm, we can know that the proposed algorithm is better than the DP-BC algorithm. This is because DP-BC is adding the same privacy budget for all users, which increases noise and thus reduces the accuracy of recommendations. Furthermore, the algorithm in this paper sets personalized differential privacy according to the different privacy requirements of different users. The perturbation on data is reduced and the performance of the recommendation algorithm is improved. In addition, compared with the BCCF algorithm, although the introduction of differential privacy affects the recommended performance to some extent, it can be seen from Table 3 that the accuracy of the proposed algorithm is not greatly reduced.

(2) Comparison with traditional privacy protection algorithms

In this section, the DP-PCC algorithm and DP-COS algorithm are selected to illustrate the superiority of BC-based similarity computation. From Table 3 we know that when $K=100$. In the Movie-Lens-100K dataset, the values of RMSE and MAE for the DP-PCC algorithm are 1.094 and 0.867. Those values of the proposed algorithm are 1.011 and 0.792, which are about 8.3% and 7.5% higher than that of the DP-PCC algorithm respectively. Compared with the DP-COS, the RMSE and MAE of the proposed algorithm are 8.1% and 7.2% higher respectively. In addition, even without considering the personalized differential privacy scheme, the DP-BC algorithm still get a better recommendation performance than DP-PCC and DP-COS algorithm. The results of the experiment in Yahoo Music dataset are similar overall to those of the Movie-Lens-100K dataset. Due to the sparseness of Yahoo Music, the performance improvement of this algorithm is more obvious. It shows that the measurement method based on BC similarity is more accurate in the calculation of similarity and can improve the differential privacy recommendation quality.

(3) Comparison with different differential privacy protection algorithms

This section analyzes the performance of the proposed algorithm by comparing it with some recent similar algorithms. The comparison algorithms selected as follows: 1) PPCF algorithm. 2) PSGD algorithm. As can be seen from Table 3, the value of RMSE and MAE gradually decreases with the increase of neighbors. In addition, since the prediction results of the PSGD algorithm have nothing to do with the number of neighbors K, it is a straight line. In the Movie-Lens-100K dataset, compared with the PSGD algorithm, the performance of the proposed algorithm is better. PPCF algorithm is the same type as the proposed algorithm in this paper. Compared with the PPCF algorithm, the algorithm is better, because the algorithm selects neighbors based on the similarity of BC and avoids over-reliance on common ratings. In the Yahoo Music dataset, when $K=20$, the PSGD algorithm is slightly better than the proposed algorithm. However, with the number of neighbor's K increases, the performance of the proposed algorithm is better than that of PSGD.

(4) The impact of privacy budget on recommendation performance

To provide the suggestion for the privacy parameters of personalized differential privacy, the influence of different privacy protection levels on the performance of the DP-BC algorithm is analyzed. The number of neighbors is $K=100$. Due to limited space, only RMSE results are given. The result are shown in Figure 1.

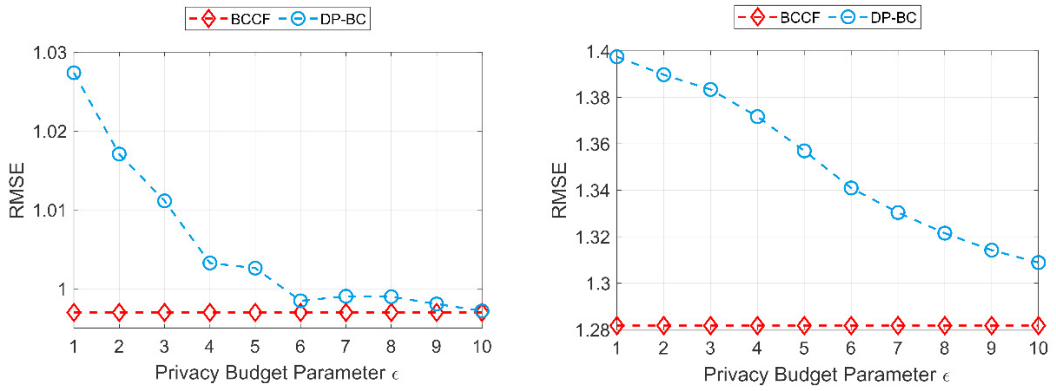

Figure 1. Different privacy budget on Move-Lens-100K (left side)/Yahoo Music (right side)

From the results in Figure 1 left side, we can see that with the privacy budget increases, RMSE decreases faster as the levels of privacy protection decreases. It indicates that higher utility costs are required when maintaining a higher level of privacy protection ($\varepsilon=1$). When the privacy budget is greater than 6, RMSE starts to decline slowly, and the Laplace noise has little influence on the algorithm. The result in the right side are similar to the left side. It is shown that DP-BC is sensitive to privacy budget parameters, and the greater the privacy budget is, the smaller the impact of differential privacy on the performance of the recommendation algorithm will be.

5. Conclusions

In order to balance the recommendation accuracy and privacy protection, BC is used as the similarity metric to solve the problem of the lack of common ratings in the recommendation system. To solve the problem of high sensitivity when the proposed scheme directly introduces differential privacy protection, the BC similarity calculation formula is modified. In addition, considering the different privacy requirements of users, personalized differential privacy is adopted to reduce the introduction of noise and improve the accuracy of recommendation. Experimental results show that the proposed privacy protection recommendation scheme keep the accuracy of recommendation while ensuring the privacy of users, and overcome the problem that the existing differential privacy collaborative filtering algorithm relies too much on common ratings. It provides a new idea for recommendation service design considering privacy protection in the context of big data.

Acknowledgements

This work was supported by the MOE Layout Foundation of Humanities and Social Sciences [No. 20YJAZH102] and the Foundation of Guangxi Key Laboratory of Cryptography and Information Security [No. GCIS201908].

References

[1] Pan Yinghui,Huo Yongfeng,Tang Jing,Zeng Yifeng,Chen Bilian. Exploiting relational tag expansion for dynamic user profile in a tag-aware ranking recommender system [J]. *Information Sciences*, 2021,545.

[2] Sushma Malik,Anamika Rana,Mamta Bansal. A Survey of Recommendation Systems [J]. *Information Resources Management Journal (IRMJ)*, 2020, 33(4).

[3] Wu Fan,Chen Zhen. A recommendation method of Japanese vocabulary learning based on embedded system and data intelligent analysis [J]. *Microprocessors and Microsystems*, 2021, 80.

[4] Jorgensen, Zach, et al. Conservative or liberal? Personalized differential privacy[C]. *2015 International Conference on Data Engineering*, 2015:1023-1034.

[5] Dwork C. Differential privacy: a survey of results[C]. *Theory and applications of models of computation (TCMC)*, 2008: 1-19.

[6] QIAO Yu. Review of Privacy Protection Strategies in Recommendation Systems [J]. *Network Security Technology and Application*, 2020(04):68-70.

[7] Frank McSherry,Ilya Mironov. Differentially private recommender systems [P]. *Knowledge discovery and data mining*, 2009.

[8] Tianqing Zhu,Gang Li,Yongli Ren,Wanlei Zhou,Ping Xiong. Differential privacy for neighborhood-based collaborative filtering [P]. *Advances in Social Networks Analysis and Mining*, 2013.

[9] Hua, J., Xia, C., & Zhong, S. Differentially Private Matrix Factorization[C]. *Proceedings of the 24th International Conference on Artificial Intelligence*.2015:1763-1770.

[10] Yang Shuxin, Zhu Kaili and Liang Wen.Differential Privacy for Context-Aware Recommender Systems[C]. *2019 IEEE 18th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)* 2019:356-360.

[11] Chunyong Yin,Lingfeng Shi,Ruxia Sun,Jin Wang. Improved collaborative filtering recommendation algorithm based on differential privacy protection [J]. *The Journal of Supercomputing*, 2019, 76(7), 1-14

[12] Yang M, Zhu T, Xiang Y, et al. Personalized privacy preserving collaborative filtering [J]. *GPC* 2017: 371-385.

[13] Xuying Meng,Suhang Wang, et al. Towards privacy preserving social recommendation under personalized privacy settings[J]. *World Wide Web*, 2019, 22(6), 1-29.

[14] Bidyut Kr. Patra,Raimo Launonen,Ville Ollikainen,Sukumar Nandi. A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data [J]. *Knowledge-Based Systems*, 2015, 82(C), 163-177.

[15] Wang, Ning, et al.Collecting and Analyzing Multidimensional Data with Local Differential Privacy[C]. *IEEE 35th International Conference on Data Engineering (ICDE)* 2019: 638-649.

[16] Zhengzheng Xian,Qiliang Li,Xiaoyu Huang,Lei Li. New SVD-based collaborative filtering algorithms with differential privacy [J]. *Journal of Intelligent & Fuzzy Systems*, 2017, 33(4):2133-2144.

[17] Arnaud Berlioz,Arik Friedman,Mohamed Ali Kaafar,Roksana Boreli,Shlomo Berkovsky. Applying Differential Privacy to Matrix Factorization [P]. *Recommender Systems*, 2015.