The Movie Recommendation System Based on Differential Privacy

Min Li[1], Yingming Zeng[2], Yue Guo[1], and Yun Guo[1](✉)

[1] College of Cyber Science, Nankai University, Tianjin, China
{limintj,2120180514,guoyun}@nankai.edu.cn
[2] Hangtian INC, Beijing, China
yingmingblue@163.com

Abstract. In the past decades, the ever-increasing popularity of the Internet has led to an explosive growth of information, which has consequently led to the emergence of recommendation systems. A series of encryption measures were adopted in the current recommendation systems in the cloud to protect users' privacy security. However, there are many other privacy attacks in this recommendation system of the cloud-based device. Therefore, this paper studies the encryption interference of setting differential privacy protection mechanism for user data in user's private devices based on untrusted servers. A dynamic privacy budget allocation method is proposed based on localized differential privacy protection technology and specific scenes recommended by movies.

Keywords: Differential privacy · Privacy budget · Movie recommendation · Collaborative

1 Introduction

With the development of information technology, users have a variety of ways to obtain information. What users spend the most time on the thing that no longer where to get information, but to find the content they are interested in among the numerous information. In this environment, recommendation system emerges as The Times require. Personalized movie recommendation service has been widely used nowadays. Personalized recommendation system usually needs a lot of user data to provide high quality recommendation services. And the data leakage means the user's privacy leakage. Most existing recommendation systems [1–6], such as Netflix movie recommendation system, are based on scenes that are trusted by servers. In the most common collaborative filtering algorithms, trusted servers need to collect all user data and analyze user behavior to perform such personalized recommendations. In the recommended method above, only privacy protection scenarios are considered when publishing to three parties, and no more scenarios of being attacked are taken into account. For example, when user data is transferred from the device to the cloud, an attacker can eavesdrop on the transmission channel and launch a "man-in-the-middle attack," so the data should be encrypted during transmission. In addition, attackers can directly hack into the cloud, servers and steal user

Y. Xiang et al. (Eds.): SocialSec 2020, CCIS 1298, pp. 318–328, 2020.
https://doi.org/10.1007/978-981-15-9031-3_28

data. This requires some encryption algorithms to protect the data stored on the cloud. In addition, people inside the server may also leak user data. Under this recommendation mode, users' privacy data cannot be effectively protected.

Therefore, this paper studies the encryption interference of setting differential privacy protection mechanism for user data on user's private device [7] based on untrusted server. In this paper, on the basis of existing research, the difference of privacy protection technology and classic movie recommendation algorithm, emphatically explores the application of localization difference privacy protection technology to solve the privacy issue in movie recommendation algorithm, the main contributions include: (1) based on localization difference privacy protection technology and film recommend specific scenarios, puts forward a method to dynamically allocate budget for privacy. The equal probability of the user's viewing frequency of the movie type is assigned to each node of the privacy prefix tree, and then the noise satisfying Laplace distribution is added according to the allocated privacy budget. (2) the user-based collaborative filtering algorithm is improved according to the actual movie scenes. In the process of user similarity calculation, a matrix similarity calculation method is used instead of the traditional vector-based similarity calculation method to find the similar group of target users. Let's define this process as DP-MRE (Differential Privacy-Movie Recommendation System).

2 Theoretical Basis

2.1 Differential Privacy Definition

Dwork defined differential privacy [8–10] as a method similar to data encryption in 2006. Differential privacy assumes that the attacker owns all the information except the target information.

Let the data set D_1 and D_2 have the same property structure, and the symmetry difference between them is denoted as $D_1 \Delta D_2$, and $|D_1 \Delta D_2|$ denotes the number of records in $D_1 \Delta D_2$. If $|D_1 \Delta D_2| = 1$, then D_1 and D_2 are said to be adjacent data sets.

Define 1 ε-Differential Privacy. Assume $\varepsilon > 0$ is a real number and M is a random algorithm that takes the data set as input. M(x) is a query result obtained for the random algorithm M. R is a subset of M of x. For all adjacent data sets D_1 and D_2 as well as all subsets R of M(x) of non-single element, the algorithm M satisfies the ε-differential privacy if the following equation is satisfied:

$$\Pr[\mathrm{M}(D_1) \in \mathrm{R}] \leq e^{\varepsilon} \times \Pr[M(D_2) \in \mathrm{R}]$$

Nature 1 (Sequence). With algorithm $M_1, M_2, \cdots, M_n$, its privacy protection budget respectively $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$, so for the same data set D, composed of these algorithms combination algorithm of $M(M_1(D), M_2(D), \cdots, M_n(D))$ provide $\sum_{i=1}^{n} \varepsilon_i$-differential privacy protection, provide privacy protection level for the sum of total budget.

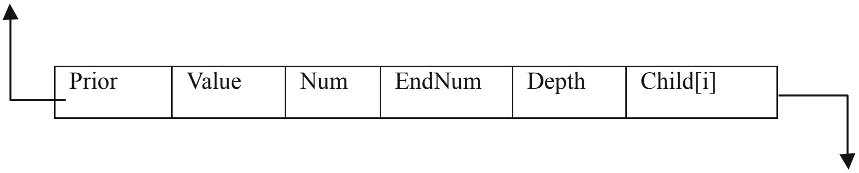
Laplace Noise Mechanism. When the initial query results are obtained, the Laplace mechanism implements $\boldsymbol{\varepsilon}$-differential privacy protection by adding noise following the

Laplace distribution to the original results. The mean value is 0, the Laplace distribution of the scale parameter is $Lap(\sigma)$, and its probability density function is:

$$p(x) = \frac{1}{2\sigma}\exp\left(-\frac{|x|}{\sigma}\right)$$

2.2 Differential Privacy Definition

The movie recommendation system based on differential privacy protection proposed in this paper combines with the characteristic structure of prefix Tree [11] to construct (DP-Tree) based on the historical record information of users' watching movies. The privacy prefix Tree is an improved prefix Tree. The data structure of the privacy prefix Tree is shown in Fig. 1:



Prior	Value	Num	EndNum	Depth	Child[i]

Fig. 1. Data structure of privacy prefix tree

In Fig. 1, *Prior* is the previous pointer pointing to the parent node; *Value* is the stored Value; *Num* is the number of times the value appears; *Depth* is the Depth of the value; *Child[i]* is an array of Pointers to Child nodes, and *EndNum* stores the current node in each path as the number of end nodes. The records of all movie types watched by users are abstracted into a privacy prefix Tree (DP-Tree) whose Root node is Root. Each node in the tree represents a movie type.

3 Design of the Differential Privacy Protection Method

3.1 The Steps of the Differential Privacy Protection Method

In the differential privacy protection method, the main steps are roughly divided into two steps: the first step is to select the appropriate privacy budget parameters and allocate the appropriate privacy budget for the protected data. The second is to add some noise interference to the protected data.

The amount of noise added is closely related to the allocation of the privacy budget ε. The value of the privacy budget ε is inversely proportional to the added noise. The privacy budget not only determines the level of differential privacy protection, but also determines the noise addition, which is the core parameter of differential privacy protection method. This article focuses on how to allocate your privacy budget. For the purpose of this article is based on the difference of privacy movie recommendation system, depending on the type of users to watch film of history data privacy prefix tree structure, privacy prefix tree high frequency sequence of film type and, in large probability on this type of movie for a user's interest degree is higher, the probability of being attacked the greater, in order to prevent privacy budget be exhausted and need budget to allocate more commonly used data privacy.

3.2 Privacy Budget Allocation Scheme Based on Prefix Tree

The film recommendation system based on differential privacy introduced in this paper is based on the data under the tree structure for protection and encryption. As shown in Fig. 2, is the information of users based on the prefix tree structure chart, of which the user to watch the movie, in accordance with the film type extract features, and in accordance with the prefix tree constructed the privacy prefix tree structure, the specific method is that the user watched a movie, extract the genre of the film, storage in the form of sequence to the tree of the substructure, each film type series combination is a path in a tree, and record the frequencies of each node and the frequencies of each node as a pseudo leaf node. In order to reasonably allocate the privacy budget to the privacy prefix tree, this paper allocates the privacy budget to each node in the privacy prefix tree in an equal proportion allocation method. In the privacy prefix tree, the abstract root node R is not a real movie type, so it will not consume the privacy budget. All other subtree nodes need to be assigned a privacy budget.

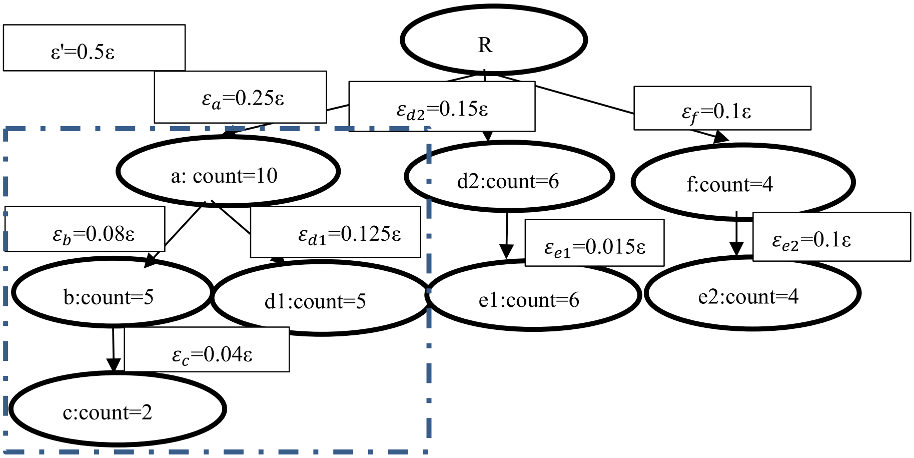


Fig. 2. Prefix tree privacy budget allocation scheme

Instead of storing the movie type directly in the prefix tree, the corresponding letter representation of the movie type is stored, as shown in Table 1.

Table 1. Mapping table of movie type letters

Film type	Love	Suspense	Action	Comedy	Plot	Ragedy
Letters mapping	a	b	c	d	e	f

The data of each node in the privacy prefix Tree (DP-Tree) structure shown in Fig. 2 is shown in Table 2.

Table 2. DP-Tree data structure

Prior	Value	Num	EndNum	Depth	Child[i]
–	R	–	–	0	[a, d2, f]
R	a	10	0	1	[b, d1]
a	b	5	3	2	[c]
b	c	2	2	3	–
a	d1	5	5	2	–
R	d2	6	0	1	[e1]
d2	e1	6	6	2	–
f	e2	4	4	2	–
R	f	4	0	1	[e2]

As shown in Fig. 2, assuming that the total privacy budget of the whole tree is, viewing frequencies of movie types a, d2 and f are 10, 6 and 4 respectively from the first layer structure of the tree. Then, the total privacy budget allocation proportion of the tree with node a as the root node is (10/20)ε, the privacy budget is allocated in the same way to $\varepsilon_b = (0.5 * 0.5 * 0.6)/2\varepsilon$. When a movie type is distributed in different sequences, the privacy budget of the movie type is equal to the sum of its allocated privacy budget, for example, the privacy budget of movie type $\varepsilon_d = \varepsilon_{d1} + \varepsilon_{d2} = 0.125\varepsilon + 0.15\varepsilon = 0.275\varepsilon$. According to the nature 1 of the differential privacy protection method, it can be concluded that:

$$\varepsilon = \varepsilon_a + \varepsilon_b + \cdots + \varepsilon_f$$

It can be seen that compared with other privacy budget allocation methods [12–16], this way of allocating privacy budget based on prefix tree based on the value of each node, not on the level structure alone. This allocation method can reasonably dynamically allocate the privacy budget in the case of big differences in tree structure, and it does not need to adjust the value of privacy budget allocation artificially.

3.3 Prefix Tree Privacy Budget Allocation Algorithm

The privacy budget allocation algorithm based on the prefix tree is as follows. Where, *TMovie* stores the result of privacy budget allocation of movie-type nodes; DP-Tree movie type node <v, ε_v> and its privacy are calculated as ε_v in the queue set *TQueue*; P$_v$ is the viewing statistical frequency of the current node v; GetTop (LinkQueue Q, string r, float e) represents the queue function of queue header element.

Privacy budget allocation algorithm：

Input：Privacy budget **ε**，The prefix tree DP-Tree, The root node R

Output：Privacy budget allocation results set TMovie

```
Initialize the TMovie and TQueue collections to 0
IF（R==’ ’）
    $\boldsymbol{\varepsilon_R}$=0
    R→child(R)
Else
    Add the current node <R, $\boldsymbol{\varepsilon_R}$> to the TQueue
    While TQueue ≠ NULL  Do
        GetTop (TQueue,R,$\boldsymbol{\varepsilon_R}$)
        IF   R ∈ TMovie  Then
            $\boldsymbol{\varepsilon_R}$← TMovie
            TMovie ← <R,$\boldsymbol{\varepsilon_R} + \boldsymbol{\varepsilon}_{\mathbf{P}_R}$>
        Else
            TMovie ← <R,$\boldsymbol{\varepsilon}_{\mathbf{P}_R}$>
        End If
        If ( $\boldsymbol{P_R}$== $\boldsymbol{P_{R-parent}}$)
            ε←ε/2
        Else
            ε←(ε-$\boldsymbol{\varepsilon_{P_R}}$)/2
        For v
            $\mathbf{P}_v$← Probability of watching movie type v
            Add <v, $\boldsymbol{\varepsilon}_{\mathbf{P}_v}$> to TQueue
        EndFor
    End while
```

In the above algorithm, the *TMovie* and *TQueue* sets are initialized to empty after the input of the privacy budget ε, the prefixed prefix tree structure and the root node R of the prefix. Will the current node and the current node privacy are added to the *TQueue* budget (R for the root node), and then determine whether the current node and its parent weights are the same, if is the same accounts for half of the current budget privacy, if not the same with his brother calculate the current node weight ratio, such as half of the parent node privacy budget proportion distribution of the remaining half of privacy. Then loop the children of the current node.

4 Detailed Design of Film Recommendation System Based on Differential Privacy Protection

4.1 The Overall Framework of Film Recommendation System Based on Differential Privacy Protection

Figure 3 is the overall frame diagram of the movie recommendation system based on differential privacy protection. The system is composed of five components, in the first place in the user's local private equipment end users' private data collection, based on the user's personal data to construct privacy prefix tree and dynamically in accordance with this privacy budget allocations, and add meet the Laplace distribution noise, after using interference user data and public data together after a recommendation algorithm calculation item want to recommend to users. The meaning of each component in the figure is as follows:

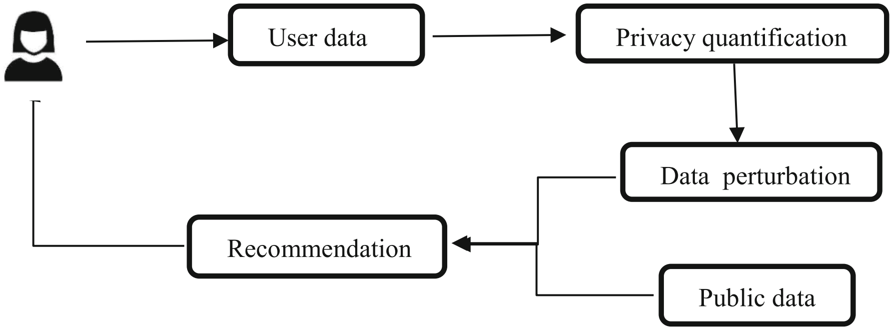


Fig. 3. Frame diagram of movie recommendation system based on differential privacy protection

Public data: to obtain Public information related to users' private data from internal or external resources. This article Users the MovieLens 1M data set, which contains 100 million ratings from 6,000 users on nearly 4,000 movies. This data set will be used as experimental data set and test data set for experimental verification in this paper.

User data: the User data is the User history data collected on the User's private device. This paper obtains the historical information record of the User's watching movies, such as the frequency of watching a certain type of movie and the User's rating information on the movie. This information is not subject to interference.

Privacy quantification: the recommendation system algorithm based on differential Privacy proposed in this paper builds the Privacy prefix tree according to the user's behavior record, and dynamically allocates the Privacy budget according to the frequency probability of each node in the Privacy prefix tree.

Data perturbation: according to the privacy prefix tree each node in the distribution of the privacy of our budget to each node adding meet the noise of the Laplace distribution, disturbance of the original Data, the Data is available to add a moderate amount of noise to protect the user's personal Data privacy. In addition, the disturbed data needs to meet two objectives: privacy protection security and recommended accuracy.

Recommendation: an untrusted third-party server obtains user information data after adding noise to build a user-movie type interest matrix, performs matrix similarity calculation based on multiple dimensions to find out similar user groups, and user-based collaborative filtering algorithm to recommend the desired information to users.

4.2 Privacy Security Analysis

The following is the security analysis of differential privacy protection based on the DP-MRE algorithm proposed in this paper. Let D_1, D_2 be the adjacent data set (that is, d$(D_1, D_2) = 1$), f(D_i) be the category set of user private data, C is the size of the public movie set, j is the user private movie data, and z(j) is the size of the Laplace noise added to the movie type j. For any r $= (r_1, \cdots, r_c) \in$ Range(DP − MRE), on the basis of the definition of differential privacy can know, if the algorithm DP - MRE content:

$$\frac{\Pr[\mathrm{DP} - \mathrm{MRE}(D_1) = r]}{\Pr[\mathrm{DP} - \mathrm{MRE}(D_2) = r]} \leq e^{\varepsilon}$$

The algorithm DP-MRE satisfies the constant ε-differential privacy protection.

According to the differential privacy protection proposed in this paper, the differential privacy protection is carried out on the user's local private device, so the privacy protection analysis only focuses on the steps of privacy budget allocation and noise addition, while there is no privacy leakage problem in the user similarity calculation and recommended steps. Therefore, privacy security analysis can be performed in the privacy budget allocation and noise addition steps. The analysis is as follows:

$$\frac{\Pr[\mathrm{DP} - \mathrm{MRE}(D_1) = r]}{\Pr[\mathrm{DP} - \mathrm{MRE}(D_2) = r]} = \prod_{j \in C} \frac{\Pr\big[\mathrm{DP} - \mathrm{MRE}(D_1)(\mathrm{j}) = r(j)\big]}{\Pr\big[\mathrm{DP} - \mathrm{MRE}(D_2)(\mathrm{j}) = r(\mathrm{j})\big]}$$

$$\geq \exp\left(- \sum_{j \in C} \frac{1}{z(j)} \left| f_j(D_1) - f_j(D_2) \right| \right)$$

$$\geq \exp(- \max_{d(D_1, D_2) = 1} \sum_{j \in C} \frac{1}{z(j)} \left| f_j(D_1) - f_j(D_2) \right| \geq e^{-\varepsilon}$$

So the DP-MRE satisfies:

$$\frac{\Pr[\mathrm{DP} - \mathrm{MRE}(D_1) = r]}{\Pr[\mathrm{DP} - \mathrm{MRE}(D_2) = r]} \leq e^{\varepsilon}$$

In the first step above, the independent injected noise on each category set is obtained from the combinational property of difference privacy and remains unchanged. In the second step, the injected Laplace noise and triangle inequality can be derived, and the above proof is completed.

5 Experimental Results and Analysis

In order to reflect the impact of differential privacy on the recommendation quality of the recommendation system (DP-MRE) in this paper, the precision rate and recall rate are used to evaluate the recommendation system model in this paper.

5.1 Influence of Accuracy

In order to objectively analyze the feasibility and effect of DP-MRE algorithm based on differential privacy protection proposed in the film recommendation system, this method is compared with S-DPDP algorithm based on differential privacy protection proposed by Shen et al. We set the difference privacy parameter as an independent variable, took different values for the privacy budget parameter in the experiment, and controlled a single variable to compare multiple recommendation algorithms. In addition, in order to more intuitively reflect the impact of privacy protection on the overall recommendation algorithm, this paper added the data recommendation algorithm Baseline without privacy protection to the experiment and compared it with it. Now, S-DPDP and DP-MRE data with privacy protection are compared with the data algorithm without privacy protection.

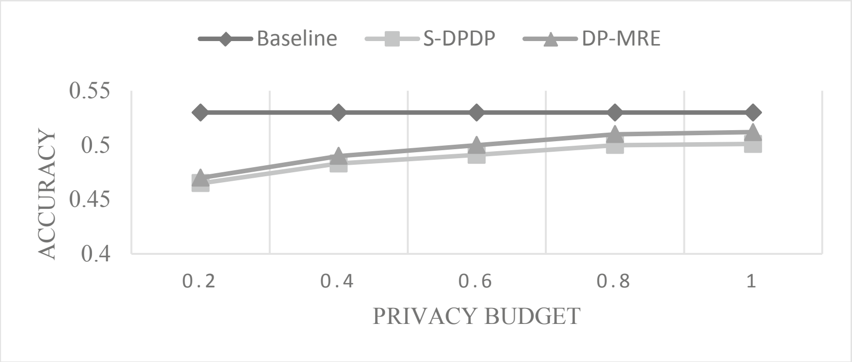


Fig. 4. Impact of difference privacy on the accuracy of the recommendation system

Figure 4 shows the impact of differential privacy protection on the accuracy of the recommendation system. By the experimental results show that the Fig. 4 in not in privacy protection recommendation system, the accuracy of the collaborative filtering recommendation system based on user recommendation about 0.53 or so, and privacy protection based on the difference of DP-MRE and S-DPDP will cause a certain degree of recommendation quality loss, when the differential privacy parameter epsilon close to 1, DP-MRE and S-DPDP algorithm recommended accuracy of about 0.51. With the increase of the parameter of the privacy algorithm, the accuracy of DP-MRE and S-DPDP algorithms gradually increases to the recommended level of the original data. Compared with S-DPDP, DP-MRE has a smaller loss of precision, because DP-MRE is a privacy budget allocated based on DP-Tree structure, which maintains the type combination sequence and frequency characteristics of the movies watched by users and distributes the Laplace noise reasonably, thus reducing the loss of recommended quality caused by noise addition. However, S-DPDP user an iterative algorithm to add noise, which blurs the similarity between users. Therefore, in the aspect of recommendation quality loss, DP-MRE is better than S-DPDP algorithm, but DP-MRE has a high time complexity in the privacy budget allocation process, which affects the overall system efficiency.

5.2 Influence of Recall Rate

Figure 5 for the influence of difference of privacy to recall rate, by the experimental results show that the Fig. 5 in not in privacy protection recommendation system, collaborative filtering recommendation system based on users recommend the recall rate of around 0.51 or so, DP-MRE based on differential privacy and S-DPDP will cause a certain degree of recommendation quality loss, but with the increase of difference algorithm parameter epsilon privacy, the recall rate gradual in the recommended level with the original data. In the recommendation results, the higher the precision rate and recall rate, the higher the quality of the recommendation system. According to the experimental results, the recall rate of DP-MRE is very similar to that of S-DPDP, which means that when the data set base is very large, the recall rate of the two recommendation algorithms is basically similar, but the recall rate of DP-MRE is still slightly higher than that of S-DPDP algorithm.

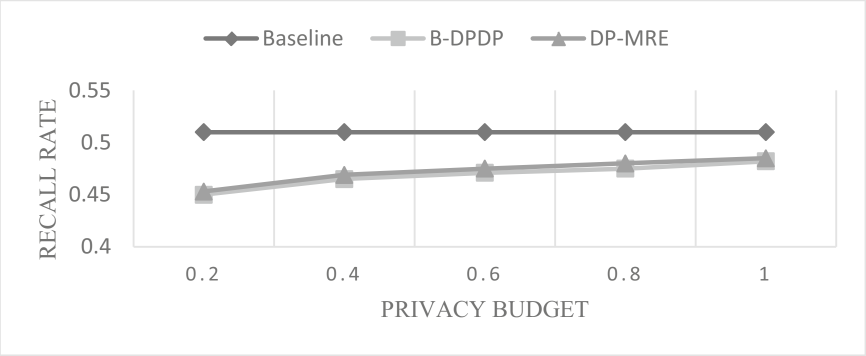


Fig. 5. Impact of difference privacy on recall rate

6 Conclusion

This paper mainly introduces how to apply the differential privacy protection technology to the movie recommendation system to solve the user privacy protection problem in the recommendation process and ensure the recommendation performance will not suffer too much loss. In conclusion, this paper first dynamically adds noise to the sensitive information of users locally to ensure the privacy and security of users, then sends the user data with noise to the server for similarity calculation, and recommends the movie to users according to the user-based collaborative filtering algorithm. In this way, users' privacy data will not be violated during the whole recommendation process. The performance and effect of the recommendation system are verified by experiments. Although we have done some research work in the field of differential privacy and recommendation system, there are still many fields to be further studied in the application of differential privacy, and many security problems in the recommendation system have not been solved. Therefore, the relevant research still has a long way to go.

References

1. Chaudhuri, K., Sarwate, A., Sinha, K.: Near-optimal differentially private principal components. In: NIPS, pp. 989–997 (2012)
2. Chaudhuri, K., Vinterbo, S.A.: A stability-based validation procedure for differentially private machine learning. In: NIPS, pp. 2652–2660 (2013)
3. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: a survey of recent developments. ACM Comput. Surv. **42**(4), 14:1–14:53 (2010)
4. Thakurta, A.G., Smith, A.: (Nearly) optimal algorithms for private online learning in full-information and bandit settings. In: NIPS, pp. 2733–2741 (2013)
5. Hardt, M., Ligett, K., Mcsherry, F.: A simple and practical algorithm for differentially private data release. In: NIPS, pp. 2339–2347 (2012)
6. McSherry, F., Mironov, I.: Differentially private recommender systems: building privacy into the net. In: KDD, pp. 627–636 (2009)
7. Ye, Q., Meng, X., Zhu, M., et al.: A review of localized differential privacy. J. Softw. **29**(07), 159–183 (2018)
8. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: privacy via distributed noise generation. In: Vaudenay, S. (ed.) EUROCRYPT 2006. LNCS, vol. 4004, pp. 486–503. Springer, Heidelberg (2006). https://doi.org/10.1007/11761679_29
9. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006). https://doi.org/10.1007/11681878_14
10. Dwork, C., Mcsherry, F., Talwar, K.: The price of privacy and the limits of LP decoding. ACM Symposium on Theory of Computing. ACM (2007)
11. Vágvölgyi, S.: Descendants of a recognizable tree language for prefix constrained linear monadic term rewriting with position cutting strategy. Theor. Comput. Sci. **732**, 60–72 (2018)
12. Hay, M., Rastogi, V., Miklau, G., et al.: Boosting the accuracy of differentially private histograms through consistency. Proc. VLDB Endow. **3**(1/2), 1021–1032 (2009)
13. Chen, R., Fung, B.C.M., Desai, B.C.: Differentially private transit data publication: a case study on the Montreal transportation system. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 213–221. ACM, New York (2012)
14. Shang, T., Zhao, Z., Shu, W., et al.: Algorithm of big data decision tree based on isometric privacy budget allocation. Eng. Sci. Technol. **51**(02), 134–140 (2019)
15. Wang, X., Han, H., Zhang, Z., Yu, Q., Zheng, X.: Budget allocation method for tree index data differential privacy. Comput. Appl. **38**(07), 1960–1966 (2008)
16. Demin, H., Liao, Z.: Differential privacy location privacy protection method for m-fork average tree. J. Small Micro Comput. Syst. **40**(03), 76–82 (2019)