Design and Implementation of Privacy-preserving Recommendation System Based on MASK

Yonghong Xie, Aziguli Wulamu and Xiaojing Hu
School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, China
Email: xieyh@ustb.edu.cn, ustbhxj@hotmail.com

Xiaojie Zhu
School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
Email: zxjju@163.com

***Abstract*—In view of the issue of internal data in the construction of recommendation system may be leaked by external person. A privacy-preserving recommendation system based on MASK is designed to address it. The system consists of two parts: privacy-preserving subsystem and recommendation subsystem. First, disrupt original data to form new privacy-preserving data according to a certain probability. Second, conduct data mining with the new privacy-preserving data in recommendation subsystem. Third, its mining results will be used to help to make recommendation for users. Experimental results show that it can protect the original data without decreasing the accuracy of recommendation system.**

***Index Terms*—privacy-preserving data mining, recommendation system, MASK**

I. INTRODUCTION

In recent years, data mining of privacy-preserving is a cutting-edge research direction in the field of data mining, which has achieved rich achievement in many research fields, such as analysis of outliers, association rules, decision tree, clustering and so on [1-5]. However, in practice application, only a small amount of typical mining problems of privacy-data have been used to summarize out some relative meaningful basic tools [6,7].

The main content of this study is privacy protection in recommendation system. Existing privacy-preserving recommendation systems concern and protect most of customer's personal information, for example, customer's personal browse and search information in local client (such as Cookies)[8, 9, 10]. Nevertheless, their contribution to the solution of the privacy-leaking problems associated with external personnel has proved to be limited when external personnel participating in the construction of recommendation system takes charge of data processing. Assuming that the data processing module in a company's recommendation system is performed by external personnel *A*, then *A* can get the original history information of customers directly or through some certain experiments [11]. If the implied knowledge of these raw data (e.g. behavior characteristics of high-quality customers or other rules) be illegally obtained by person with ulterior motives, it will seriously affect the enterprise's core competitiveness.

In view of the possible privacy-preserving problems, a privacy-preserving recommendation system based on MASK algorithm is designed and implemented to handle these issues in this paper, the characteristics of the system mainly as follows:

1) Adopting mining algorithm of privacy-preserving data to protect users' history information, so it is difficult to obtain the real data even the crimes know the technology of the recommendation system.

2) The parameters of privacy-preserving are not decided independently by human, within the scope of artificial, it is determined by program calculation that choose the optimum parameter values to suitable for the current data which increase system's intelligence.

3) Ask no hard requirements for the storage form and attribute characteristics of real data. All these can be directly dealt by the application of current system.

4) Through mining users' history data, a knowledge base based on users' history information is set up to provide a credible environment for the reliability of recommendation system. Meanwhile, it reduces the load and response delay time of recommendation system since the mining part is performed offline.

II. BACKGROUNDS

MASK (Mining Associations with Secrecy Konstraints) algorithm is a classic mining algorithm of association privacy-preserving data put forward by Indian scholars Rizvi [12, 13], which is able to maintain highly privacy and obtain accurate mining results at the same time.

MASK algorithm is based on random transformation technology that its mining dataset is formed by real dataset through probability transformation. And the

Manuscript received January 7, 2014; revised June 1, 2014; accepted July 1, 2014.

This work is supported by the National Key Technology R&D Program in 12th Five-year Plan of China (No. 2013BAI13B06) corresponding author, Aziguli Wulamu. (ustbhxj@hotmail.com)

doi:10.4304/jsw.9.10.2607-2613

implementation of MASK algorithm is based on classic Apriori algorithm which produces frequent 1-itemsets first, then generates frequent k-itemsets, finally, obtains strong association rules [14]. The difference between MASK algorithm and Apriori algorithm is the counting method of support of itemsets. Apriori algorithm is applied to real database mining, so it just need statistic the tuple number of items including candidate itemsets. For example, assuming the 2-itemsets represent for 11, it only need compute the tuple number of items that contains the 2-itemsets namely 11. While, improved MASK algorithm estimates support of itemsets based on deformation data resulting from real data. Likewise, for the same 2-itemsets, its original itemsets 11 become one of 00, 01, 10 and 11 after transformation, the real support of the 2-itemsets can be calculated only after considering above four situations. Similarly, calculate the support of n-itemsets in a real database based on deformation datasets, it needs to consider 2^n situations that generated from the deformation of original real n-itemsets [15].

III. Design and Implementation of Privacy-preserving Recommendation System

A. Overall Framework of System

The framework of privacy-preserving recommendation system in this paper is modified and designed on basis of that of traditional recommendation system shown in figure 1, adding the parts of data perturbation and parameter selection of privacy-preserving. The overall framework of privacy-preserving recommendation system is shown in figure 2.

Privacy-preserving recommendation system in this paper is also applicable for centralized data and distributed data since it is based on distributed storage environment and reveals its superiority in operating efficiency when dealing with huge amounts of data [16]. The whole system is made up of two relatively independent subsystems: privacy-preserving subsystem and recommendation subsystem. In privacy-preserving subsystem, program chooses the optimum parameters to disrupt original data to get new data after a normalization processing for original data file. In recommendation subsystem, utilize mining algorithm of privacy-preserving data (MASK) to conduct a data mining for new data that have been protected, and select Top - N products from the mining results to comprise users' corresponding recommendation list according to the history information of users.

B. Privacy-preserving Subsystem

Privacy-preserving subsystem is responsible to disrupt original data to reach a certain protective effect which is the key step of privacy prevention. It consists of three modules: data preprocessing, selection for optimal parameter and data perturbation. Processing flow of the subsystem is normalization processing for original data first, then according to the characteristics of original data, select an optimal perturbation parameter for it, perturb original data with 0-1 probability on basis of the selected parameter, finally, generate new data with protective factors.

1) Data preprocessing

It is widely recognized that original data is hard for unification processing in data mining, therefore, a relevant normalization processing and conversion of data storage format for original data are needed that will be suitable for program processing. All the raw data in this module will be converted to the mode of "user–itemsets", and each line data represents all operations of each one of users. Eventually, the history information data of users after preprocessing would be converted into the following patterns: userid \t itemid1, itemid2, itemid3…

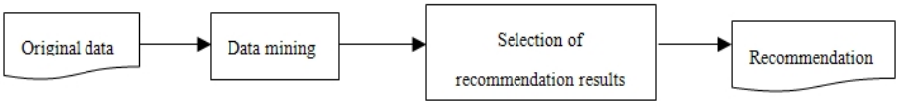


Figure 1. Framework of traditional recommendation system.

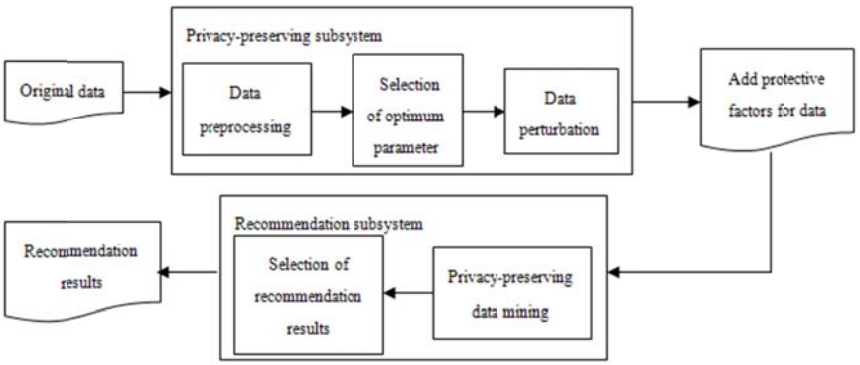


Figure 2. Framework of privacy-preserving recommendation system.

2) **Selection for the optimum parameter**

After data normalization, select an optimum parameter *P* as the perturbation probability for next-step data. Firstly, conduct repeatedly random subsampling with the datasets that needed to be protected, then simulate experiments with data obtained from subsampling processing, find out an optimum perturbation probability *P* promising a highly protection for private data as well as making protected-data relatively similar with original data. To simplify the procedure, the range of *P* is (0.5, 1), because, by means of 0-1 probability distribution, p produces the same utility as that of 1-*P* in the process of perturbing data. (This conclusion has been proved in literature [11]).

There are two standards for evaluating parameter in this module:

(1) mean absolute error *MAE* (*P*) of similarity among items before and after data perturbation

MAE (*P*) is a standard for evaluating the difference between original data and new perturbed-data, and the value of *MAE* (*P*) is the smaller the better. Computation formula of *MAE* (*P*) is as (1), $sim_r(i,j)$ is the real similarity of item *i* and item *j*, $sim(i,j)$ is the similarity of item *i* and item *j* after being perturbed. *m* is data items.

$$MAE(P) = \frac{\sum_{i,j \in m}(sim_r(i,j) - sim(i,j))}{C_m^2} \quad (1)$$

(2) protection degree *P* (*P*) of data

P (*P*) represents the protection degree of privacy data information, the greater the value of *P* (*P*), the higher level protection for data, it means the lower probability of data to be remade. The calculation method of *P* (*P*) as (2), *R* (*P*) is original data probability that reconstructed by protected data. a is the percentage weight of privacy of 1 and 0, it is the average support of item, $R_1(P)$ and $R_0(P)$ respectively represent the probability of 1 and 0 that can be reconstructed by protected data.

$$P(P) = \big(1 - R(P)\big) * 100 \quad (2)$$

$$R(P) = aR_1(P) + (1-a)R_0(P) \quad (3)$$

$$R_1(P) = \frac{S_0 \times P^2}{S_0 \times P + (1-S_0) \times (1-P)} + \frac{S_0 \times (1-P)^2}{S_0 \times (1-P) + (1-S_0) \times P} \quad (4)$$

$$R_0(P) = \frac{(1-S_0) \times P^2}{S_0 \times (1-P) + (1-S_0) \times P} + \frac{(1-S_0) \times (1-P)^2}{S_0 \times P + (1-S_0) \times (1-P)} \quad (5)$$

In recommendation system, choose *P* making *P*(*P*) / *MAE* (*P*) maximum as the optimal parameter, because such chosen parameter in privacy-preserving subsystem can make private data been protected as highly as possible and the new data obtained after data perturbation as similar as original data.

3) **Data perturbation**

In data perturbation module, parameter *P* chosen from last module is used for probability perturbation of original data. The benchmark of data perturbation is: $X = \{X_i\}$ is a random variable, $X_i = 0 \; or \; 1$, $Y_i = X_i \; XOR \; \overline{r_i}$ is the variable after $X = \{X_i\}$ is perturbed. $\overline{r_i}$ is the supplement of disturbing factor r_i, the probability of r_i conforms to 0-1 probability distribution, its distribution probability can be expressed as (6). According to the characteristics of 0-1 distribution, the probability of each dataitem *X* keeping its original value is *P*.

$$P(r_i = 1) = P, P(r_i = 0) = 1 - P \quad (6)$$

C. Recommendation Subsystem

Recommendation subsystem is the core part of this system, made up by data-mining module and selection module of recommendation results. It is mainly responsible for mining new data and choosing a corresponding recommendation list for each one of users according to the characteristics and results after data mining.

1) **Data mining module**

Compared with traditional data mining module of recommendation system, recommendation system designed in this paper allows external personnel to participate in and ensures no leakage of private data. Because data handled in data mining part is not the real data, it is the new data generated after the real data was perturbed. Three main tasks in this module: refactoring support of real itemsets, calculate similarity among itemsets, statistics similar items of each itemset.

(1) refactoring support of the real itemsets

In consideration of using perturbed-data in recommendation system, it needs to re-estimate the support of real frequent itemsets, and we use the method of MASK algorithm. Assuming *T* is the corresponding matrix of real datasets, matrix *T* is deformed to matrix *D*, its probability of deformation is p. The counts of 1 and 0 in column *i* of *T* is C_1^T and C_0^T, the counts of 1 and 0 in column *i* of *D* is C_1^D and C_0^D. According to 0-1 probability distribution [17, 18]. Equations as follows:

$$M = \begin{bmatrix} P & 1-P \\ 1-P & P \end{bmatrix}, C^D = \begin{pmatrix} C_1^D \\ C_0^D \end{pmatrix}, C^T = \begin{pmatrix} C_1^T \\ C_0^T \end{pmatrix}$$

$$C_1^T \times P + C_0^T \times (1-P) = C_1^D \quad (7)$$

$$C_0^T \times P + C_1^T \times (1-P) = C_0^D \quad (8)$$

$$C^T = M^{-1} C^D \quad (9)$$

Solve out the equations, support C_1^T of 1 - itemsets in real data can be estimated by matrix *D*. Similarly, the real support of n - itemsets can be estimated using the same calculation, the only difference is that *M* is an $2^n * 2^n$ matrix, C^T and C^D are both the $2^n * 1$ matrixes [15]. Due to the support of n - itemsets calculated in MASK algorithm is 2^n-order, we used an improved algorithm of MASK in this system [19], the calculation method of inverse matrix *M* as follows:

$$M_k^{-1} = \frac{1}{2P-1} \begin{pmatrix} M_{k/2}^{-1} & \\ & M_{k/2}^{-1} \end{pmatrix} \begin{pmatrix} PE_{k/2} & (P-1)E_{k/2} \\ (P-1)E_{k/2} & PE_{k/2} \end{pmatrix} \quad (10)$$

(2) calculate the similarity among itemsets

Data sparsity is serious in actual e-commerce that contributed greatly to the influence of zero transaction calculating similarity. The more the number of zero transaction, the greater likelihood of the weight of popular products be enlarged and the less true of

calculating similarity among itemsets that will be affecting the accuracy of recommendation. In order to prevent the influence of zero transaction in the historical information, we choose cosine measure that without been affected by zero transaction to calculate the similarity among itemsets. The specific calculation formula as (11), $sup\_count(A,B)$ is the times of A and B occurs at the same time.

$$cos(A,B) = \frac{sup\_count(A,B)}{\sqrt{sup\_count(A)} \times \sqrt{sup\_count(B)}} \quad (11)$$

(3) statistics similar items of each itemset

For the convenience of selecting recommendation results, similar-items of each one of items in this module will be put in order. Arrange similar-items of each item by descending order respectively according to similarity (sim) count and support (sup) count. The result form after ranking as follows: Item\t Item1, sim, sup; Item2, sim, sup; Item3, sim, sup…

2) **Selection module of recommendation result**

Users are the processing object in this module. For each user, get its I-itemsets of history data, then get all corresponding similar-items of each one of items in I-itemsets, and arrange them by descending order according to the count of similarity, support, choose the first N items which not in I-itemsets as recommendation results, that is the recommendation list of each user.

IV. Experimental Results And Analysis

We test the effectiveness of the framework of privacy-preserving recommendation system from two aspects with three different groups of datasets. First test the optimum perturbation parameters obtained by program whether can increase privacy degree in the case of ensuring a highly recommendation precision as far as possible or not, then test the recommendation accuracy of system.

A. Experimental Data

1) **EachMovie dataset**

EachMovie dataset is a dedicated database of the research system center of Digital Equipment Corporation, which acquired from the internet of Digital Equipment Corporation. EachMovie dataset collecting from 18 months data, evaluation information of 1628 movies from 72916 users, evaluation data using discrete type is graded to (0, 2, 4, 6, 8, 10). Experimental data used in this paper is the evaluation data of 300 films from 450 users selected randomly from above dataset, and reserves only the films with its evaluation level ≥ 4.

2) **MovieLens dataset**

MovieLens dataset is from GroupLens project team of Minnesota University who developed MovieLens to be an investigative recommendation system based on Web to receive marks of films from users and provide corresponding recommendation list of films for users. MovieLens dataset contains 100000 score data of 1682 films from 943 users, each of these users scores at least 20 films. In this dataset, there are 19 categories of all the films, 20 feature attributes sorted out of 1682 movies as the attribute tables.

3) **Microblog users' data**

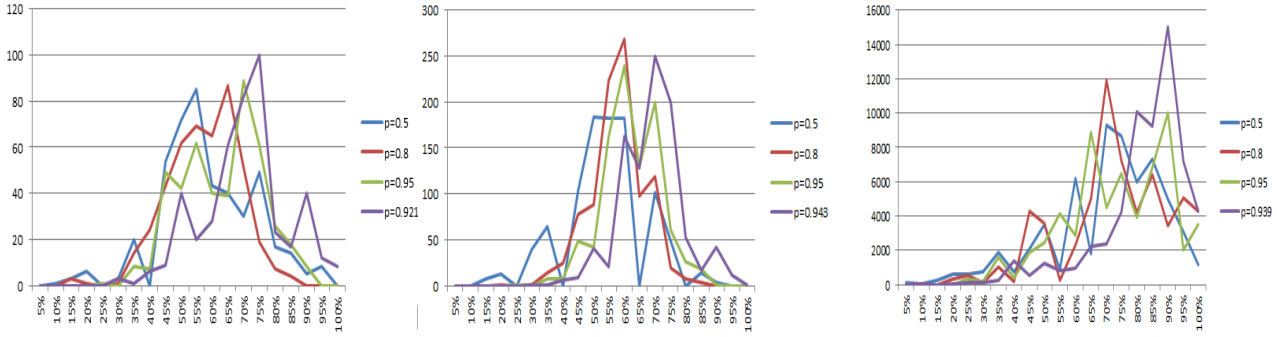
Microblog users' data is a kind of open source data, Experimental data in this study includes 60000 users' data of a microblogging site, the data about each user includes user id, name, province, city, registered time, data collection time, gender, whether the authenticated user or not, friends number, fans number, the number of published microblog, etc.

B. Result and Analysis

1) **Experiment of the effect of optimum perturbation parameter**

Data perturbation parameter is the key of recommendation system, if the parameter is too much small, the protection level of data privacy will be too low to reach the purpose of privacy protection, while, too larger parameter the cost of reducing result accuracy of recommendation to guarantee the protection level of privacy, so it is very important to select a perturbation parameter just right.

In order to examine the effect of automatic-selection part of data perturbation parameter, we use 0.5, 0.8, and 0.95 as the perturbation parameters for experiment and compare the results with that of automatic-selection parameters. Figure 3 is the statistical figures of experimental results of three groups of datasets with different perturbation parameters, figure (a) is the distribution of experimental results based on EachMovie in the case of different perturbation parameters, figure (b) is the distribution of experimental results based on MovieLens in the case of different perturbation parameters, figure (c) is the distribution of experimental results based on Microblog in the case of different perturbation parameters. P=0.921, P=0.943 and P=0.939 in figure (a), figure (b) and figure (c) are the perturbation parameters selected automatic by program for corresponding group of experiments.



(a) (b) (c)

Figure 3. The statistical figures of experimental results of three groups of datasets with different perturbation parameters. (a) experimental result of dataset based on EachMovie; (b) experimental result of dataset based on MovieLens; (c) experimental result of dataset based on Microblogging

As can be seen from figure 3, program can select different optimum perturbation parameter *P* according to the characteristics of data in different datasets. Vertical axis represents the number of users and horizontal axis represents its corresponding recommendation accuracy. For figure (a), when the number of users is about 100, recommendation precision of perturbation parameter selected by program reaches almost 75 percent, for figure (b), when the number of users is about 250, recommendation precision of perturbation parameter selected by program reaches almost 80 percent, for figure (c), when the number of users is about 15000, recommendation precision of perturbation parameter selected by program reaches almost 90 percent. Compared with other perturbation parameters provided in the experiment, it is obvious that, perturbation parameter selected by program is prioritized as the optimal parameter. As previously mentioned, such chosen parameter in privacy-preserving subsystem can make private data been protected as highly as possible and the new data obtained after data perturbation as similar as original data. Above experiments with different datasets have shown that optimum perturbation parameters selected by program made recommendation precision of experiments highest.

2) Contrast experiments of recommendation precision

Above experimental results show, the program designed for optimum perturbation parameter is able to select reasonable perturbation parameter. However, a good perturbation parameter does not represent the recommendation effect of recommendation system. Evaluation measurements of recommendation precision mainly include methods of statistic accuracy and methods of decision support accuracy. We adopt *MAE* (Mean Absolute Error), one of the methods of statistical accuracy as the measure of recommendation precision. For each user, statistics the counts of same-itemsets of recommendation results from privacy-preserving system and traditional recommendation system, then the value of *MAE* is the probability of the counts of same-itemsets in recommendation results. Therefore, the value of *MAE* is the larger the better.

We take users of three groups of datasets as the experimental data. Choose thirty most likely films as a recommendation list for each user of first group and second group and choose thirty most likely friends having the same interests with each user as a friend recommendation list for the third group. We compare the probability of same-items in recommendation list of each user at the situation of adding protection strategy and without protection strategy.

The characters of two recommendation systems in contrast experiment as shown in table 1. From table 1, the difference between two recommendation systems is due solely to using different data mining algorithms, traditional recommendation system takes Apriori algorithm as the core algorithm of data mining, while privacy-preserving recommendation system uses MASK algorithm as the core algorithm of data mining. Actually, MASK algorithm is an algorithm adding privacy protection strategy on that basic of Apriori algorithm. All processing methods of datasets of both algorithms are almost the same. Now therefore, their results of contrast experiment have a certain reference value. Experimental hypothesis of recommendation results of traditional recommendation system based on Apriori are true, for each user, statistics the counts of same-itemsets of recommendation results from privacy-preserving system and traditional recommendation system, the value of *MAE* is the probability of the counts of same-itemsets in recommendation results. Measure formula for recommendation precision is as (12). A, B respectively represents the recommendation list provided by two recommendation systems for user in contrast experiment, $same\_num\,(A,B)$ statistic the counts of same recommendation items from two lists.

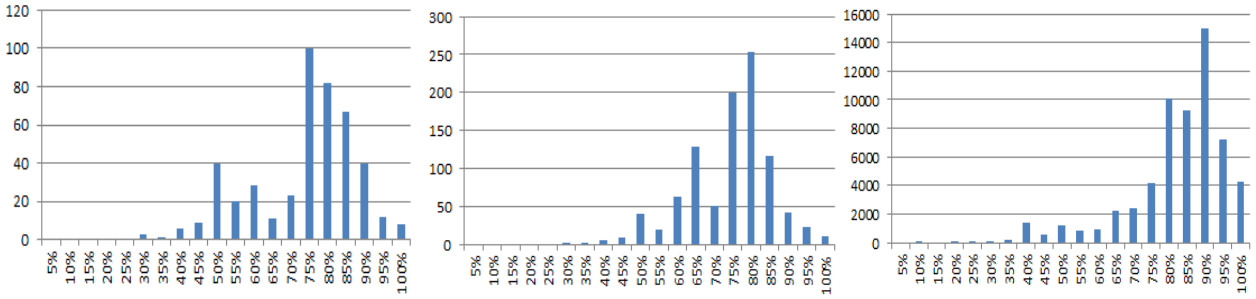
$$MAE = \frac{same\_num(A,B)}{30} \times 100\% \qquad (12)$$

Statistical distributions of experiment results after contrasting experiment with three groups of different datasets are shown as figure 4. Figure (a) is the distribution of recommendation accuracy from 450 users on EachMovie dataset, figure (b) is the distribution of recommendation accuracy from 943 users on MovieLens dataset, figure (c) is the distribution of recommendation accuracy from 6w users on Microblog dataset.

TABLE 1.

CHARACTERISTICS OF RECOMMENDATION SYSTEMS IN CONTRAST EXPERIMENT

Contrasting System	privacy-preserving recommendation system	traditional recommendation system
System characters	add protection strategy	without protection strategy
experiment datasets	three groups of different datasets	three groups of different datasets
Core algorithm	MASK	Apriori
Recommendation results	each user produces thirty recommendations	each user produces thirty recommendations



(a) (b) (c)

Figure 4. Accuracy distribution of users' recommendation lists on three groups of datasets.

(a) recommendation accuracy of 450 users on EachMovie dataset; (b) recommendation accuracy of 943 users on MovieLens dataset;

(c) recommendation accuracy of 6w users on Microblog dataset

As can be seen from figure 4 that vertical axis shows the number of users and horizontal axis gives its corresponding recommendation accuracy, that is, the value of *MAE*. For figure (a), when the number of users is about 100, the value of *MAE* is almost 75 percent, for figure (b), when the number of users is about 250, the value of *MAE* is almost 80 percent, for figure (c), when the number of users is about 15000, the value of *MAE* is almost 90 percent. From the perspective of similarity of recommendation system, the recommendation accuracy provided by privacy-preserving recommendation system remains almost above 80%, namely at least average 24 items are the same in 30 recommended items. Furthermore, considerate comprehensively from the perspectives of similarity and diversity of recommendation system, privacy-preserving recommendation system is of a significant reference value for later research.

V. CONCLUSIONS

In this paper, privacy-preserving recommendation system utilizes distributed processing environment to meet the need of huge amounts of data in existing environment, and the data processing is an offline processing that reducing access time of online, that is, system meets the need of timely response of online, in other words, the recommendation system is feasible in principle. Experiments on three groups of different datasets show, recommendation system designed in this paper has a reference value in practical application. However, the limitations of MASK algorithm make system still a lot to be improved, for instance, the problem of increasing accuracy of recommendation results, cold-boot problem in recommendation results , the mark problem when users take historical data to forecast the score of itemsets, etc. All these problems are the next-step research focus.

ACKNOWLEDGMENT

This work is supported by the National Key Technology R&D Program in 12th Five-year Plan of China (No. 2013BAI13B06).

REFERENCES

[1] P. Zhang, Y. H. Tong, S. W. Tang, D. Q. Yang, X. L. Ma, "An Effective Method for Privacy Preserving Association Rule Mining," Journal of Software, Vol. 17, No. 8, pp. 1764-1774, August 2006. doi: 10.1360/jos171764

[2] W. W. Fang, B. R. Yang, J. Yang, C. S. Zhou, "Decision-Tree Model Research Based on Privacy-Preserving," Pattern Recognition and Artificial Intelligence, Vol. 23, No. 6, pp. 766-771, December 2010. doi:10.3969/j.issn.1003-6059.2010.06.004

[3] X. M. Chen, J. H. Li, J. Peng, H. L. Liu, J. Zhang, "A Survey of Privacy Preserving Data Mining Algorithms," Computer Science, Vol. 34, No. 6, pp. 183-186, June 2007. doi:10.3969/j.issn.1002-137X.2007.06.050

[4] J. Y. Xie, S. A. Jiang, W. X. Xie, X. B. Gao, "An Efficient Global K-means Clustering Algorithm," Journal of Computers, Vol. 6, No. 2, pp. 271-279, February 2011. doi:10.4304/jcp.6.2.271-279

[5] A. R. Xue, S. G. Ju, W. H. He, W. H. Chen, "Study on Algorithms for Local Outlier Detection," Chinese Journal of Computers, Vol. 30, No. 8, pp. 1455-1463, August 2007.

[6] F. Zhang, X. D. Sun, H. Y. Chang, G. S. Zhao, "Research on Privacy-Preserving Two-Party Collaborative Filtering Recommendation," Acta Electronica Sinica, Vol. 37, No. 1, pp. 84-89, January 2009.

[7] C. Clifton, M. Kantarcioglu, J. Vaidya, X. D. Lin, M. Y. Zhu, "Tools for Privacy Preserving Distributed Data Mining," ACM SIGKDD Explorations Newsletter, Vol. 4, No. 2 , pp. 28-34, December 2002. doi:10.1145/772862.772867
[8] T. Chen, W. L. Han, M. Yang, "Personalized Recommendation System Based on Privacy Protection," Computer Engineering, Vol. 35, No. 8, pp. 283-285, April 2009. doi:10.3969/j.issn.1000-3428.2009.08.096
[9] H. Sun, H. Z. Hu, W. H. Dai, H. J. Mao, Y. Zhang, "Intelligent System for Customer Oriented Design and Supply Chain Management," Journal of Computers, Vol. 7, No. 11, pp. 2842-2849, November 2012. doi:10.4304/jcp.7.11.2842-2849
[10] W. X. Hong, S. T. Zheng, H. Wang, J. C. Shi, "A Job Recommender System Based on User Clustering," Journal of Computers, Vol. 8, No. 8, pp. 1960-1967, August 2013. doi:10.4304/jcp.8.8.1960-1967
[11] N. Ramakrishnan, B. J. Keller, B. J. Mirza, A. Y. Grama, G. Karypis, "Privacy Risks in Recommender Systems," IEEE Internet Computing, Vol. 5, No. 6, pp. 54-63, December 2001.
[12] S. J. Rizvi, J. R. Haritsa, "Maintaining Data Privacy in Association Rule Mining," Proceedings of the 28th international conference on Very Large Data Bases. VLDB Endowment, 2002.
[13] P. Andruszkiewicz, "Optimization for Mask Scheme in Privacy Preserving Data Mining for Association Rules," Rough Sets and Intelligent Systems Paradigms. Springer Berlin Heidelberg, 2007, pp.465-474. doi:10.1007/978-3-540-73451- 2_49
[14] R. Agrawal, T. Imieliński, A. Swami, "Mining Association Rules between Sets of Items in Large Databases," ACM SIGMOD Record. ACM, 1993. doi: 10.1145/170036.170072
[15] Z. L. Shen, J. G. Cui, "Improved Algorithm of Association Rule Mining in Privacy Preserving," Computer Engineering and Applications, Vol. 46, No. 8, pp. 133-136, April 2010. doi:10.3778/j.issn.1002-8331.2010.08.038
[16] Y. Q. Huang, Z. D. Lu, H. P. Hu, R. X. Li, "Privacy Preserving Distributed Data Mining Association Rules of Frequent Itemsets," Computer Engineering, Vol. 32, No. 13, pp. 12-14, July 2006. doi:10.3969/j.issn.1000-3428.2006.13.005
[17] C. X. Zhang, X. Z. Qian, "Optimization for MASK Algorithm in Privacy Preserving Data Mining," Computer Engineering and Design, Vol. 30, No. 14, pp. 3316-3318, July 2009.
[18] S. Y. Liu, Q. X. Yang, Q. M. Ma, "Research on Privacy Preserving Algorithm Based on Random Perturbation," Journal of North University of China (Natural Science Edition), Vol. 32, No. 5, pp. 596-599, October 2011. doi:10.3969/j.issn.1673- 3193.2011.05.013
[19] Z. H. Wu, S. Liu, J. G. Cui, "A Improved MASK Algorithm by Using Divide and Conquer Strategy," Microcomputer Information, Vol. 25, No. 36, pp. 78-80, December 2009.

Yonghong Xie She is an associate professor, who now works at School of Computer and Communication Engineering, University of Science and Technology Beijing. She was born in 1970. Her main research fields include knowledge discovery, intelligent system and database technology.

Aziguli Wulamu She is a professor, who now works at School of Computer and Communication Engineering, University of Science and Technology Beijing. She was born in 1969. Her main research fields include knowledge engineering (knowledge base), natural language processing and innovation theory.

Xiaojing Hu She is a master, who now studies at School of Computer and Communication Engineering, University of Science and Technology Beijing. She was born in 1990. Her main research interest is database technology.

Xiaojie Zhu She had obtained a master degree of School of Computer and Communication Engineering, University of Science and Technology Beijing. She was born in 1986. Her main research interest is recommendation system.