PRIVACY-PRESERVING COLLABORATIVE FILTERING SYSTEM FOR BOOK-CROSSING DATASET

Alper YARGIÇ
Bilecik Şeyh Edebali University, Software Engineering Department

Elif Tuğçe AÇIL
Bilecik Şeyh Edebali University, Computer Engineering Department

ABSTRACT

Web services that store and use their users' sensitive data can cause privacy violation issues. Using personal preferences to generate predictions may increase individuals' privacy risks in collaborative recommendation systems. Users who worry about privacy violations may be willing to provide false information and sometimes refuse to use these services. As a result, the recommender system's prediction generation quality will decrease because it is an undeniable fact that the accuracy of prediction is directly related to the quality of the collected user data.

It is crucial to discuss the privacy risks that may arise from the use of such systems and to protect user data privacy with accepted privacy protection mechanisms to alleviate user concerns. In this study, we evaluate the randomized perturbation-based privacy protection mechanism on a traditional memory-based collaborative filtering system that used the Book-Crossing dataset. We also compared recommendation accuracy over varying levels of privacy to find a balance between accuracy and privacy issues. Experimental results based on real-world user data show that a privacy-preserving scheme maintains the confidentiality of personal preferences without severely compromising prediction accuracy.

Keywords: Privacy-Preserving Collaborative Filtering, Recommender Systems, Randomized Perturbation.

INTRODUCTION

With the spread of the Internet, people tend to do their daily routines through online services. However, obtaining the desired information in a large amount of data is a significant challenge. Information filtering approaches such as Recommendation Systems (RSs) are valuable tools that make it easy for users to find relevant information from online services. Collaborative Filtering (CF) is a widely used data filtering technique within RSs.

CF systems have some significant challenges such as data sparsity, scalability, and shilling attacks (Su and Khoshgoftaar, 2009). Privacy protection is one of the crucial issues to overcome among these challenges. CF systems have the fundamental principle that people

who act similarly in the past tend to agree in the future. Therefore, it needs users' past preference values. Storing and using user preference values cause privacy problems for individuals. Collecting user preference values may reveal personal information that should be kept confidential, such as their lifestyle, shopping habits, and financial situation. Users who feel insecure about data privacy can sometimes give the system false or fabricated ratings to trick the CF system and hide their private data (Yargic and Bilge, 2017). However, the success of prediction generation is directly related to the quantity and quality of collected user preferences (Bilge et al. 2013). Unrealistic or fabricated data lead to inaccurate predictions for users and adversely affect the success of the CF system. Eliminating user concerns is a remarkable way of solving prediction accuracy issues in CF systems.

Privacy-Preserving Collaborative Filtering (PPCF) techniques are used to alleviate user data privacy concerns and ensure system security. There are various approaches such as cryptography, k-anonymity, differential privacy, obfuscation, and perturbation-based methods to provide data privacy in PPCF systems (Wei et al. 2018). Data manipulation approaches such as obfuscation and Randomized Perturbation (RP) techniques are commonly used privacy protection methods to alleviate privacy issues in CF systems (Kamal et al. 2017).

RP-based disguising techniques hide actual user ratings before being sent to the server, thus user data privacy is protected (Polat and Du, 2005). Guided by this method, Bilge and Polat (2013) proposed an RP-based PPCF approach using a bisecting k-means clustering-based algorithm. In another study, Gong (2011) used RP techniques to ensure user privacy on distributed user profiles. Polatidis et al. (2017) proposed improvements for RP-based PPCF system, which used multiple levels and different ranges on random values. In addition, Liu et al. (2017) proposed a privacy protection approach that could provide more privacy than existing RP-based protection mechanisms by combining RP techniques with a differential privacy approach to improve the levels of privacy. In another hybrid approach, Su et al. (2019) proposed a procedure that used differential privacy and RP to ensure privacy was preserved. The use of more than one criterion in recommendation systems allows producing more personalized predictions, while at the same time it causes more violations of user privacy (Yargıç and Bilge, 2017). Another RP-based approach was proposed to mitigate emerging privacy risks by generating masking data that protected privacy at variable levels depending on the amount of information (Yargıç and Bilge, 2019).

In this study, we used the Book-Crossing (BX)[1] dataset with a RP-based PPCF system in a conventional memory-based collaborative filtering system. We also compared recommendation accuracy over varying levels of privacy to find a balance between conflicting accuracy and privacy goals. Real-world data-based experimental results are presented in the following sections.

[1] http://www2.informatik.uni-freiburg.de/~cziegler/BX/

EXPERIMENTAL METHODOLOGY

Within the scope of this study, we investigate the prediction accuracy effects of the RP approach for the BX dataset. For this purpose, the well-known memory-based similarity calculation procedure is used in the CF system to generate predictions with disguised data (Polat and Du, 2005).

In this work, the data disguising process consists of four main steps as follows:

(i) *Z*-score normalization to provide standardization of ratings,

(ii) Determination of privacy level over σ parameter according to user privacy needs and dataset rating range,

(iii) Generating random numbers according to *normal* distribution and σ parameter,

(iv) Adding generated random numbers to z-score normalized ratings and obtaining disguised ratings.

In summary, if we name the z-score normalized rating vector as U and the random numbers vector as R, the disguised ratings will be equal to $U + R$. For more information on the data disguising and prediction estimation procedures, we direct the reader to Polat and Du (2005).

Dataset

The BX dataset consists of three separate sections: user information, book information, and the preference values given to the books by the users. For experimental purposes, we used the BX table with numerical preference values. The preference values contain the book ratings, and the assigned ratings are 0 to 10 range. If a rating value is equal to zero, it means that a user has read a book but did not vote for it. Therefore, ratings with a value of zero are deleted from the table.

The BX dataset contains 1,149,780 ratings from 278,858 users. The fact that 99.9 percent of the available items were not rated made the dataset extremely sparse. This percentage of sparsity adversely affects prediction accuracy (Yargic and Bilge 2019; Adomavicious and Kwon 2007). Due to the sparsity of the dataset, we utilized a subset of the dataset collection in which each user has at least 15 ratings and each book has at least 10 ratings, called BX_15_10 datasets. The variation of sparseness and its details were given in Table 1.

Table 1. Comparison of sparsity degrees between BX and BX_15_10 dataset.

	BX dataset	**BX_nonzero dataset**	**BX_15_10 dataset**
Number of Unique Users	105,283	77,805	756
Number of Unique Books	340,553	185,972	1173
Number of Ratings	1,149,780	433,671	21,690
Sparsity Rate	99.9 %	99.9 %	97.5 %

Disguising BX_15_10 Dataset and Generating Predictions

According to Gross and Acquisti (2005), individuals' opinions about personal data privacy may differ. Therefore, experiments were performed on variable privacy control parameters to

investigate the effect of the utilized privacy protection mechanisms. The masking vectors were calculated with *normal* distribution using the data disguising procedure. Furthermore, a masking vector is generated based on the user's desired level of privacy using the σ_{max} parameter, which is determined by the service provider and represents the highest level. The σ parameter determined by the user in the range of (1, σ_{max}] values represents the privacy level of individuals. The σ_{max} parameter was used in the range of 2 to 10 since the BX dataset has rating values between 1 and 10. Therefore, all experiments were conducted within nine different σ_{max} levels to observe the prediction accuracy. The importance that users attach to data privacy varied, so variable values were used in the σ parameter. The σ_{max} parameter was multiplied by a number in the (0-1] range to emulate users' privacy selection behavior. Thus, all user ratings were disguised over a random σ parameter within the specified σ_{max} range. Finally, a differential entropy-based privacy measure metric introduced by Agrawal and Aggarwal (2001) was used to analyze the effect of RP-based disguising on user privacy.

In the prediction generation process, a neighborhood-based traditional CF approach was used. The obtained neighbors were ordered based on their similarity values, and the recommendation process was carried out over the *N* most similar users, where *N* represents the number of neighbors and remained constant across all experimental works. In order to determine the constant N value, predictions were created based on undisguised raw user ratings and the maximum accuracy was observed when the N value was 10.

Cross-validation methodology was used to comprehensively analyze the negative impact of masking data on prediction accuracy on all user preferences. 10-fold cross-validation was utilized in the prediction generation process. For this purpose, each user's rating values were divided into ten groups, which were divided into nine for the train set and one for the test set. For the prediction accuracy assessment, the predictions generated for the test set were compared to the raw user ratings, and the Mean Absolute Error (MAE) was calculated. This process was repeated for each fold, and predictions for each rated item were generated. Averaging the entire MAE values produces the resulting error value. Each set of experiments was repeated five times to mitigate deviations in prediction accuracy due to the randomness of the generated masking number vectors. The averages of the results obtained at each iteration of the experimental sets were presented as final accuracy values. The details of the experimental studies, and the effects of the tests on privacy and accuracy were discussed in the following section.

RESULTS AND DISCUSSION

In this section, we present the empirical results of various privacy parameters to evaluate the prediction generation success of the PPCF on the BX_15_10 dataset. In the utilized privacy protection protocol, each user determined the privacy level by choosing a random σ value

over the range (1, σ_{max}] and disguising the user rating values by adding random noise to the actual rating values. The random noise vector disguised actual ratings with random numbers from a zero-mean distribution compatible with the σ value. Therefore, the chosen σ value defined how the disguising of the actual ratings was strongly related to the observed accuracy losses, so σ_{max} was changed from 2 to 10 to analyze such losses.

To simplify the comparison of privacy levels, a zero-one normalization was applied and the results were presented. The privacy levels of random noise vectors obtained with *normal* distribution by using variable σ values were shown in Figure 1. With an increasing σ value, the randomness of the masking data was increased, and hence the level of privacy was also increased. In summary, there was a linear relationship between user privacy and the σ parameter, and an increase in σ positively affected the level of privacy.

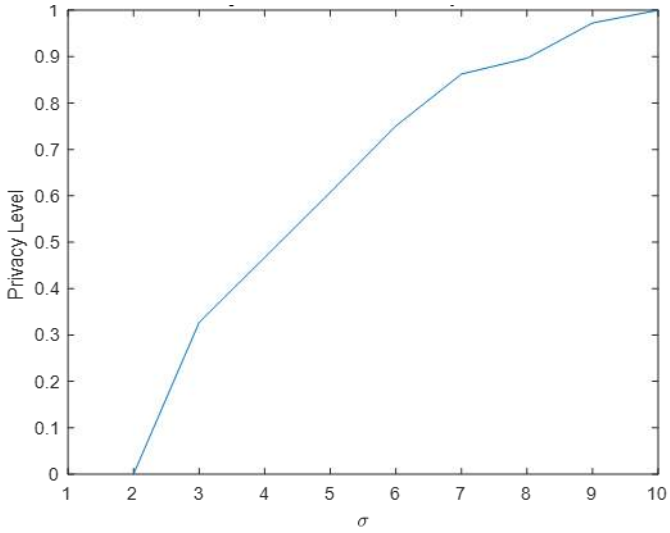

Figure 1. Privacy levels with variable σ parameters

With the increasing σ value, the value range of the random number vector and the associated confidentiality level will increase. However, the optimal σ value to be used in the disguising process for the user is not a parameter that can only be decided on the obtained privacy level. Therefore, when determining the optimal σ for a user, there must be a balance between privacy and prediction accuracy. The comparison of the RP process in terms of recommendation accuracy on the BX data set was shown in Figure 2. The fundamental principle used in PPCF systems was to provide data privacy while minimizing loss of prediction generation accuracy. According to the privacy levels shown in Figure 1, it was observed that the increase in privacy level obtained with σ in the range of (1, σ_{max}] s1owed down for values greater than 7. For this reason, the ideal σ_{max} parameter for the BX_15_10 data set was determined as 7 to balance between privacy and prediction accuracy. The result was a reasonable loss of prediction accuracy while keeping the level of privacy as high as possible.

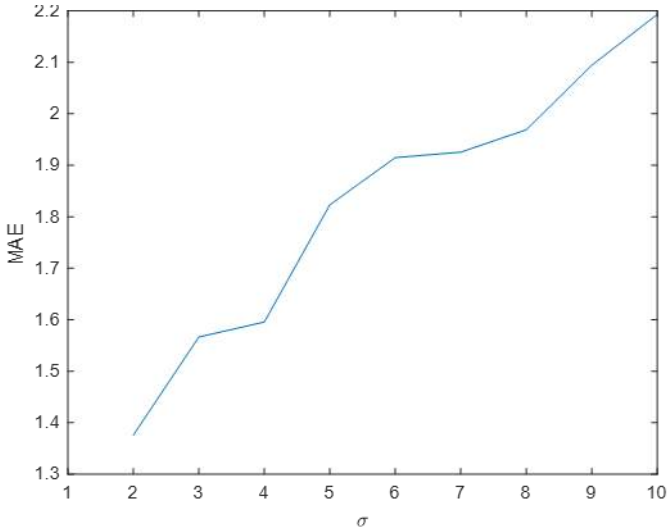

Figure 2. MAE with variable σ parameters

CONCLUSION

CF systems have privacy risks regarding collecting and using individuals' ratings. Many users have privacy concerns and are not willing to share their personal preferences. Failure the collecting user ratings hinders CF systems generate an accurate prediction. To alleviate the user data privacy problem there are many studies in the literature. RP techniques are one of the methods for creating the PPCF system and, protecting user privacy by adding noise to their genuine ratings. In this work, we evaluate the RP-based privacy protection mechanism on a traditional memory-based collaborative filtering system that used the BX dataset. We perform experiments on BX_15_10 the less sparse version of real-world BX preference data to scrutinize prediction accuracy. In order to assess the RP efficiency on the BX_15_10 dataset, we analyze the procedure based on provided privacy levels and obtained recommendation accuracy. Then, we measure privacy levels in terms of the amount of applied random noise. The BX_15_10 dataset was disguised with RP techniques for different perturbation levels, which was chosen according to the rating range of the genuine data set. The results showed a reasonable loss in recommendation accuracy while keeping the level of privacy as high as possible.

REFERENCES

Adomavicius, G., & Kwon, Y. (2007). New recommendation techniques for multicriteria rating systems. IEEE Intelligent Systems, 22(3), 48-55.

Agrawal, D., & Aggarwal, C. C. (2001, May). On the design and quantification of privacy preserving data mining algorithms. In Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (p. 247-255).

Bilge, A., Kaleli, C., Yakut, I., Gunes, I., & Polat, H. (2013). A survey of privacy-preserving collaborative filtering schemes. International Journal of Software Engineering and Knowledge Engineering, 23(08), 1085-1108.

Bilge, A., & Polat, H. (2013). A scalable privacy-preserving recommendation scheme via bisecting k-means clustering. Information Processing & Management, 49(4), 912-927.

Gong, S. (2011). Privacy-preserving collaborative filtering based on randomized perturbation techniques and secure multiparty computation. International Journal of Advancements in Computing Technology, 3(4), 89-99.

Gross, R., & Acquisti, A. (2005, November). Information revelation and privacy in online social networks. In Proceedings of the 2005 ACM workshop on Privacy in the electronic society (pp. 71-80).

Kamal, R., Hussein, W., & Ismail, R. (2017). Privacy preserving recommender system based on improved MASK and query restriction. In 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS) (p. 310-314). IEEE.

Liu, X., Liu, A., Zhang, X., Li, Z., Liu, G., Zhao, L., & Zhou, X. (2017). When differential privacy meets randomized perturbation: a hybrid approach for privacy-preserving recommender system. In *International Conference on database systems for advanced applications* (p. 576-591). Springer, Cham.

Polat, H., & Du, W. (2005). Privacy-preserving collaborative filtering. International journal of electronic commerce, 9(4), 9-35.

Polatidis, N., Georgiadis, C. K., Pimenidis, E., & Mouratidis, H. (2017). Privacy-preserving collaborative recommendations based on random perturbations. Expert Systems with Applications, 71, 18-25.

Su, C., Chen, Y., & Xie, X. (2019). Location recommendation with privacy protection. In Proceedings of the 2019 3rd International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (p. 83-91).

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. Advances in artificial intelligence, 2009.

Wei, R., Tian, H., & Shen, H. (2018). Improving k-anonymity based privacy preservation for collaborative filtering. Computers & Electrical Engineering, 67, 509-519.

Yargic, A., & Bilge, A. (2017). Privacy risks for multi-criteria collaborative filtering systems. In 2017 26th International Conference on Computer Communication and Networks (ICCCN) (p. 1-6). IEEE.

Yargic, A., & Bilge, A. (2019). Privacy-preserving multi-criteria collaborative filtering. Information Processing & Management, 56(3), 994-1009.