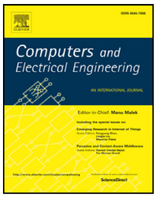
Computers and Electrical Engineering 000 (2018) 1–11

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng



Improving *k*-anonymity based privacy preservation for collaborative filtering[☆]

Ruoxuan Wei[a], Hui Tian[b], Hong Shen[b,c,*]

[a] Center of Information Technology, Ministry of Agriculture, China
[b] School of Computer Science, University of Adelaide, Australia
[c] School of Data and Computer Science, Sun Yat-Sen University, China

ARTICLE INFO

Article history:
Received 29 June 2017
Revised 9 February 2018
Accepted 9 February 2018
Available online xxx

Keywords:
Recommender system
Collaborative filtering
Privacy preserving
Anonymity

ABSTRACT

Collaborative Filtering (CF) is applied in recommender systems to predict users' preference through filtering the information or patterns. Privacy Preserving Collaborative Filtering (PPCF) aims to achieve privacy protection in the recommendation process, which has an increasing significance in recommender systems and thus attracted much interests in recent years. Existing PPCF methods are mainly based on cryptography, obfuscation, perturbation and differential privacy. They have high computational cost, low data quality and difficulties in calibrating the magnitude of noise. This paper proposes a (p, l, α)-diversity method that improves the existing k-anonymity method in PPCF, where p is attacker's prior knowledge about users' ratings and (l, α) is the diversity among users in each group to improve the level of privacy preserving. To achieve (l, α)-diversity, users in each equivalence class shall come from at least l $(l < k)$ clusters in α clustering results. Therefore, we firstly apply Latent Factor Model (LFM) to reduce matrix sparsity. Then we propose an improved Maximum Distance to Average Vector (MDAV) microaggregation algorithm based on importance partitioning to increase the homogeneity among the records in each group which can retain better data quality in (p, l, α)-diversity model. Finally, we apply t-closeness in PPCF. Theoretical analysis and experimental results demonstrate our approach assures a higher level of privacy preserving and less information loss than existing methods.

1. Introduction

Recommender systems help users discover items they might not have found by themselves [1]. Collaborative Filtering(CF) has become the most widely used technology in the field of recommender systems. The recommendations provided by CF are based on the assumption that a user will prefer items that similar users prefer. It provides users personal recommendation by analyzing users' historical transaction data and interests, and can be partitioned into the neighbourhood-based methods, model-based methods and hybrid methods. However, CF recommender system has the risks of leaking its users' privacy due to the way it works. How to balance recommendation quality and users' privacy has been a research hotspot.

The literature in CF recommender systems has shown some traditional privacy preserving approaches. which can be considered in two ways to protect users' privacy. The first category is mainly to protect users' preference by processing users'

[☆] Reviews processed and approved for publication by Editor-in-Chief Dr. M. Malek.

[*] Corresponding author.
E-mail address: tianhh@gmail.com (H. Tian).

https://doi.org/10.1016/j.compeleceng.2018.02.017
0045-7906/© 2018 Elsevier Ltd. All rights reserved.

profile information and rating data. For example, cryptographic methods [2], obfuscation methods and perturbation methods [3]. But these approaches have some shortcomings, such as unnecessary computational cost, the difficult of calibrating the magnitude of noise. The second category focuses on the protection of the recommendation algorithm. Such as randomized methods [4–6], which apply differential privacy to achieve privacy preservation. However, the differential privacy protection level is usually too high to ensure data quality. k-anonymity is gradually applied to the privacy preserving collaborative filtering (PPCF) for protecting users' preference [7,8], which is the main technique studied in this paper.

To overcome the problem of unassured security guarantee with high prediction accuracy in the neighbourhood-based CF recommender systems, we propose a novel method, (p, l, α)-diversity, which has a better performance than other methods. First, we apply Latent Factor Model (LFM) to reduce matrix sparsity. Then we propose an improved MDAV microaggregation algorithm based on importance partitioning [9] to increase homogeneity among records in each group and (p, l, α)-diversity model where p is attacker's prior knowledge about users' ratings and (l, α) is the diversity among users in each group to improve the level of privacy preservation. Finally, we give the realization of t-closeness in PPCF. Experimental results show that this method has a higher level of privacy preservation with less information loss.

The rest of this paper is organized as follows: We introduce the existing PPCF methods on CF recommendation systems in Section 2 and preliminaries in Section 3. In Section 4, we present our algorithm, an improved (p, l, α)-diversity approach to Privacy Preserving Collaborative Filtering. Section 5 shows the results of our experiments and performance evaluation. Finally, we conclude the paper in Section 6.

2. Related work

A considerable amount of literature has been published on PPCF, including cryptographic methods [2,10], obfuscation methods [11–13], perturbation methods [3,14], randomised methods [5,6,15]. There is a trade-off between the accuracy of recommendations provided to users and the level of privacy preserving. We divide these methods into two categories: protection on users' original data and protection on recommendation algorithm. The first category is mainly for protecting users' preference by processing users' profile information and rating data before using these information, such as cryptographic methods, obfuscation methods, perturbation methods and k-anonymity methods [1,16]. Cryptographic methods provide the most reliable security because they don't add noise to the original ratings. But extra and unnecessary computational cost impacts limits its application by normal users. Obfuscation methods use a number of random values to replace part of the users' original rating. Weinsberg et al. [11] designed an approach for effectively adding ratings to a user's profile for obfuscating the user's gender against inferring user's sensitive information. In these methods, the key is the percentage of replaced ratings. Perturbation methods mainly add noise that meet a certain distribution to the users' original rating. Polat et al. [3] proposed a randomized perturbation technique on CF recommendation systems to protect users' original ratings while still guarantee accurate recommendations. Gong [14] presented a CF algorithm based on randomized perturbation techniques and secure multiparty computation. The randomized perturbation methods are used in the process of user data collection to protect privacy. But it is difficult to control the magnitude of noise.

The second category aims at protection on recommendation algorithm, and differential privacy mechanism [17] has been widely used in this category. McSherry et al. [5] first introduced differential privacy into CF recommendation algorithm. They add Laplace noise into the covariance matrix to protect the original nearest neighbours. But the introduction of Laplace noise is not conducive to recommend accurately. Zhu et al. [15] proposed a Private Neighbour Collaborative Filtering (PNCF) algorithm which introduced Recommendation-Aware Sensitivity and a re-designed differential privacy mechanism in selecting the nearest neighbors. They used a truncated parameter λ to improve the prediction accuracy.

Patsakis et al. [1] introduced Statistical Disclosure Control (SDC) and microaggregation concepts to guarantee k-anonymity to protect users' original data. Unfortunately, the records in groups built by microaggregation are too close to resist on attacks. This motivates us to investigate how to build l-diversity model on sensitive attributes based on the construction of k-anonymity. It was pointed out in [8] that all information in the recommendation dataset could be quasi-identifier attributes, and can be regarded as confidential information. It is assumed that an attacker must have obtained some prior knowledge about the ratings. The more knowledge the attacker has, the more dangerous they are. In this paper, we assume that p items in the dataset are attacker's piror knowledge about users' ratings and remaining items are sensitive attributes. Then we propose a (p, l, α)-diversity model to improve the level of privacy protection.

3. Preliminaries

In this section, we introduce some fundamental concepts including KNN collaborative filtering (we use K in capital here to distinguish the k in k-anonymity), attack model and k-anonymity.

3.1. K Nearest Neighbour Collaborative Filtering

Because of easy implementation and high prediction accuracy, K Nearest Neighbor Collaborative Filtering is widely used in recommender systems. It provides recommendations to users based on the user's K nearest neighbours. There are two main stages: Neighbour Selection and Rating Prediction. The Neighbour Selection stage locates K most similar neighbours of