Computer Communications 68 (2015) 17–24

Contents lists available at ScienceDirect

Computer Communications

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

Can user privacy and recommendation performance be preserved simultaneously?

Tingting Feng*, Yuchun Guo, Yishuai Chen

School of Electrical and Information Engineering, Beijing Jiaotong University, Beijing, China

ARTICLE INFO

Article history:
Available online 27 June 2015

Keywords:
Privacy protection
Balance precision
Similarity metric
Recommendation

ABSTRACT

In online systems of videos, music or books, users' behaviors are disclosed to the recommender systems to learn their interests. Such a disclosure raises a serious concern in the public for the leak of users' privacy. Meanwhile, some algorithms are proposed to obfuscate users' historical behavior records to protect users' privacy, at the cost of degradation of recommendation accuracy. It is a common belief that such tradeoff is inevitable. In this paper, however, we break this pessimistic belief based on the fact that people's interests are not necessarily limited to items which are geared to a certain gender, age, or profession. Based on this idea, we propose a recommendation-friendly privacy-preserving framework by introducing a privacy-preserving module between a recommender system and user side. For instance, to obfuscate a female user's gender information, the privacy-preserving module adds a set of extra factitious ratings of movies not watched by the given user. These added movies are selected to be those mostly watched by male viewers but interesting the given female user. Extensive experiments show that our algorithm obfuscates users' privacy information, e.g., gender, efficiently, but also maintains or even improves recommendation accuracy.

1. Introduction

Recommendation service plays an important role in online video systems, but the possible disclosure of users' privacy via recommendation raises serious concern in the public. The accurate recommendation is of cardinal significance for both the service providers and users. It helps users get satisfying products with ease and in turn increase the revenue for the providers [1]. Hence Netflix ever set $1 M Prize for 10% improvement of recommendation accuracy [2]. However, since a recommender system (RS) makes recommendations via mining user historical behavior to find their interests, it has been found that user personal information such as gender and age can be inferred as well. For instance, one's gender can be inferred from his or her records of watching videos [3,4], and one's age can be inferred from his or her web browsing history [5]. It worries some users that leak of personal information may lead to malicious attacks, e.g., prank calls, annoying advertisement. As a result, some users are even disgusted with personalized recommendation [6]. It is a burning issue of finding a solution to the dilemma of efficient recommendation and the protection of user identity information [7].

Many existing schemes have protected user privacy successfully during the recommendation processes. And they commonly solve this problem via sacrificing the recommendation accuracy [3,6,8]. In contrast, in this paper, we set up a simple and effective framework to protect user privacy information from being inferred, and retain the recommendation accuracy without degradation meanwhile. The key to the success of this framework is a privacy-preserving (PP) module that makes the user privacy information blurred but the interest pattern undistorted and even more distinct towards RS. Since both the advertisers and malicious attackers can divide users into two groups, i.e., target group and non-target group, we exam PP module via binary classification of user privacy in detail.

We illustrate its functionality with a simple example. In an online book system, to avoid old people being troubled with too many elderly-oriented advertisements, the PP module will add a set of extra factitious ratings of unread books which satisfy the following conditions: (1) their readers are mostly youth; (2) they are similar to the books that the user is interested in. Then, we set the ratings of selected books equal to the average ratings given by the youth. With the help of PP module, we can not only protect old people's age from being inferred but also provide them with accurate recommendations simultaneously. In Section 5, we introduce the work steps of PP module in detail.

We take gender preservation as an instance to elaborate our algorithm as does the existing work on privacy preservation in RS [3]. Gender is one of the most basic personal information cared by not only users themselves but also advertisers or other third parties who want to make their commercials more targeted. We evaluate our

* Corresponding author. Tel.: +86 13811152897.
E-mail addresses: 11111022@bjtu.edu.cn, tt06274031@gmail.com (T. Feng), ychguo@bjtu.edu.cn (Y. Guo), yschen@bjtu.edu.cn (Y. Chen).

http://dx.doi.org/10.1016/j.comcom.2015.06.013
0140-3664/© 2015 Elsevier B.V. All rights reserved.

framework on the MovieLens and Flixster datasets, which are commonly used in studies on RSs and privacy information obfuscation algorithms [3,9–12]. In particular, we use balanced precision [4] as the metric to evaluate the accuracy of gender inference to handle the serious imbalance between different user groups in this dataset, e.g., the number of male users is much greater than that of female ones. Such an imbalance is common in real systems [5]. And we also make a complexity analysis of our method in Section 6. The main contributions of this paper are as follows.

- We treat the recommendation-friendly privacy protection (RFPP) problem as a multi-objective optimization problem and propose an efficient heuristic algorithm to solve it. As a result, we can preserve privacy without loss of recommendation accuracy simultaneously.
- We introduce a PP module, which serves between users and RSs to disguise user ratings before they are disclosed to the RS and the privacy inferrer (PI). With this module, the privacy information is blurred but the interest preference of user keeps undistorted.
- We propose a new similarity metric of movies based on the average difference of ratings and the number of users. Evaluation results show that our metric outperforms the existing ones.
- We add factitious ratings based on her/his interests and the average ratings from the opposite privacy group. Our scheme performs better both in the recommendation accuracy and the user privacy protection than the existing works.

The rest of this paper is organized as follows. The related work is presented in Section 2. In Section 3, we describe the RFPP problem, and in Section 4, we give out the prerequisite knowledge. Section 5 introduces the algorithm framework, and Section 6 shows our experiment analysis. Finally, we make a conclusion in Section 7.

2. Related work

There are mainly three aspects of related work about RS and user privacy.

Recommendation enhancement with extra information. The social relationship or other related information of users increasingly received research interests to be used to improve recommendation accuracy [12–14]. For example, Yang et al. studied Top-k recommendation problem using social network data from Epinions and Flixster [12]. Observing that users may have different opinions on their shared interests, Gurini et al. presented a sentiment-based approach to Twitter user recommendation [15]. For comparison, they all utilize extra data resources to improve RSs' performance. However our method does not need extra data like social relationship and sentiment information.

Protection of data privacy. How to protect user's privacy, especially in personalized services becomes a research focus in recent years. To decrease users' psychological burden, the work of privacy-preserving collaborative filtering (PPCF) begins with Canny's paper [16], and many related works based on PPCF have been proposed, such as [17]. Moreover, some works utilized randomized perturbation techniques (RPT) to realize PPCF and make data private or avoid greatly exposing users' privacy [18–21]. Others make use of the clustering-based PPCF [22] or the semi-trusted third party [23,24]. Furthermore, since anonymity is generally conceived to be an integral part of user's right to privacy, Kambourakis contributes to acquiring a comprehensive view of the anonymity research via examining how anonymity is put to work in practice [25]. These works mainly focus on protection of individual user's interests pattern, and in this paper, we consider the protection of demographic information of individual users.

Protection of demographic information. Although different users may have different privacy concerns, personal demographic information is most private. Hence it is taken into consideration in research of privacy protection. Berkovsky et al. described a privacy-enhanced collaborative filtering by distributing the stored profiles to several repositories [6]. They also proposed an algorithm, which blocks the properties of a single user, but not the aggregated preferences of a large group of users, to be inferred. This algorithm achieved this at the cost of some loss of recommendation accuracy [8]. It brings about a concern that when most of the users in a group have the same demographics like gender, once some users have their gender disclosed, the whole group may suffer the risk of disclosure. Besides, Weinsberg et al. blur user gender via adding extra factitious ratings to users' existing records but also with some loss of recommendation accuracy [3]. To our knowledge, none of these schemes can improve the recommendation accuracy when protect user demographic information. And our early work only protects user gender information for video systems [26]. The recommendation-friendly obfuscation method proposed in this paper is the first general one to solve this dilemma by only obscuring users' privacy information but not their interest information via adding factitious data for a number of scenarios.

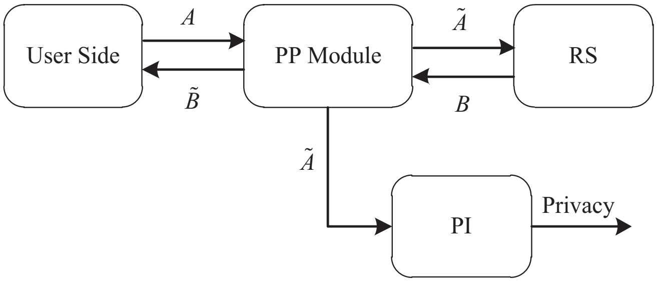

Fig. 1. Illustration of problem framework.

3. Problem description

Is it possible to preserve privacy without degrading recommendation quality? We regard this problem as RFPP problem, and first give the definition of it. In this section, we also make an analysis of the RFPP problem.

3.1. Problem definition

The problem studied in this paper is how to design a PP module to minimize the loss of recommendation accuracy and privacy disclosure risk for users, given user behavior records. Specifically, the PP module functions as a filter between user side and RS and the possible PI, as shown in Fig. 1. A transformation is applied to the user behavior data feeding into the PP module, which obfuscates the demographic information without distorting the user's interest information. RS and PI both work on the data outgoing from PP module. In the reverse direction, recommendation to the users from the RS is rectified in PP module.

The original PI is chosen according to the type of the protecting privacy. It can be trained as a multi-classifier or a linear classifier for the demographic information like race or age. Since almost all advertisers and malicious have a defined target user group, we take the binary classifier as PI to divide users into target group and non-target group. Since the gender prediction is typical binary classification problem, we take gender as the protection privacy to introduce the function of PP module in the following subsections.

Moreover, the framework can be applied in a number of scenarios. Though we take movie systems to exam the function of our method, and the related systems with items like music, CDs, and books also fit the framework in Fig. 1. Once they have recommendation requirement, they need to improve recommendation and protecting user privacy simultaneously. Moreover, browsing, listening, watching, rating, etc. such behavior records can be utilized as same as watching records. Thus, the generality of our method is obvious, and we take