Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 33 – 42

Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Randomization-based Privacy-preserving Frameworks for Collaborative Filtering

Zeynep Batmaz[a], Huseyin Polat[a,*]

[a]*Computer Engineering Department, Anadolu University, 26470 Eskisehir, Turkey*
{*zozdemir, polath@anadolu.edu.tr*}

Abstract

Randomization-based privacy protection methods are widely used in collaborative filtering systems to achieve individual privacy. The basic idea behind randomization utilized in collaborative filtering schemes is to add randomness to original data in such a way so that required levels of accuracy and privacy can be achieved. Although there are various studies on privacy-preserving collaborative filtering using randomization, there are no well-defined privacy-preserving frameworks for collaborative filtering algorithms based on randomization. In this paper, we present eight randomization-based privacy-preserving frameworks for privacy protection in collaborative filtering schemes. We first group privacy-preserving methods into two broad categories. We then classify them based on private data. Final grouping is done while considering varying privacy concerns of individual users. The frameworks can be chosen according to individual users' expectations and be utilized for privacy protection. The well-defined privacy-preserving frameworks form a basis for privacy protection based on randomized perturbation and randomized response techniques in collaborative filtering studies.

Peer-review under responsibility of KES International

Keywords: Framework, privacy, randomization, collaborative filtering, data masking

1. Introduction

E-commerce helps online vendors collect user preferences about various products traded over the Internet. Such preferences are considered valuable and private asset because they can be used for recommendation purposes and help online vendors increase their sales/profits. Moreover, revealing such data might cause various privacy risks. To transform such data into knowledge, e-commerce sites employ recommender systems. Collaborative filtering (CF) is one of the most common recommender systems. CF schemes employ other people's data for generating predictions for single items and top-N recommendation lists. A traditional CF algorithm consists of three steps such as similarity computation, neighbor selection, and recommendation estimation [1].

Users provide their preferences about products they bought or showed interest to online vendors. An $n \times m$ user-item matrix is created to store such data, where n and m represent number of users and items, respectively.

* Corresponding author. Tel.: +90-222-321-3550 ; fax: +90-222-323-9501.
E-mail address: polath@anadolu.edu.tr

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.091

Without privacy protection, user might not feel comfortable to share their preferences with e-commerce sites. They either refuse to give data at all or tend to provide false data. High quality results can be derived from high quality data. To collect high quality enough data for recommendation purposes, privacy-preserving collaborative filtering (PPCF) schemes are used[2,3]. Users' preferences represented by numeric or binary ratings can be considered private. Furthermore, it might be more damaging for revealing whether a user bought an item or not. Hence, rated/unrated items can also be considered private data. Privacy protection methods aim to mask such private data (ratings and rated/unrated items). To achieve privacy, different privacy protection methods are used in PPCF systems[2]. The most common method is known as randomization in general. Randomization adds some randomness to original data. Since CF algorithms usually depend on aggregate data, it is still possible to estimate accurate recommendations from perturbed data. Level of randomness should be chosen in such a way so that accurate predictions can be estimated while preserving privacy. However, accuracy and privacy are conflicting goals. Moreover, users' privacy concerns might be different. Likewise, user preferences can be represented using numeric or binary ratings. These constraints require different privacy protection measures.

As presented in[2,3], there are numerous PPCF studies based on randomization. Different researchers propose to use randomization-based privacy-preserving techniques for privacy in CF systems. The methods can be grouped as randomized perturbation techniques (RPTs) and randomized response techniques (RRTs) for numeric and binary ratings-based PPCF schemes, respectively. There are also different privacy control parameters. Their values should be carefully chosen. Varying users' privacy concerns should be considered as well. The studies on randomization presented in[2,3] do not use steady privacy-preserving measures. It is more appropriate to employ common privacy-preserving frameworks for fair comparisons in CF schemes. Well-defined and structured privacy-preserving frameworks should be presented for CF systems.

Following the above-mentioned motivation, we design eight privacy-preserving frameworks for CF. The frameworks are structured for numeric and binary ratings-based schemes because they use RPTs and RRTs, respectively. Some users might consider true ratings only as confidential while some might consider true ratings and rated/unrated items as private data. Hence, different frameworks should be designed for two different confidential data considerations. Finally, due to varying privacy concerns, users might perturb their confidential data variable; or they might decide to use invariable data masking. Thus, two different approaches are proposed for variable and invariable data perturbation. Considering these cases, we design eight frameworks. These frameworks will form a common base for researchers in the CF research field.

2. Related Work

Agrawal and Srikant[4] propose to use value distortion on randomization as a privacy-preserving method. The true value x_i is masked with a random value r_i; and $x_i + r_i$ is shared. Note that r_i is drawn from a distribution with zero mean. The authors also discuss privacy levels provided by randomization based on uniform and Gaussian random number distributions. Polat and Du[5,6,7] propose RPTs in order to achieve confidentiality in CF systems. The authors consider invariable data disguising. In addition to invariable data perturbation, variable data disguising-based RPT are used in PPCF schemes[8,9]. Due to varying privacy concerns, users tend to disguise their private data in such a way so that required levels of privacy is achieved. This results inconsistent data masking. Compared to invariable data perturbation, variable data disguising provides higher privacy level. Users randomly decide random number distribution and the related standard deviation. Uniform or Gaussian random number distribution can be used to generate random noise. To disguise rated/unrated items, the authors propose to fill in some randomly chosen unrated item cells. Number of the filled cells can be selected uniformly over a range. Similarly, standard deviations of the random number distributions can be uniformly randomly selected over a specified range. Gong[10] utilizes RPTs to achieve privacy in both centralized and decentralized PPCF schemes. Basu et al.[11] also use RPTs to protect individual user ratings in Slope One predictors for CF. The authors focus on masking real ratings only. They also propose to disguise deviations computed by subtracting the related ratings off-line. Zhu et al.[12] utilize data perturbation to hide the true ratings of the neighbors.

Numeric and binary ratings hold different properties. RPTs are suitable to perturb numeric ratings while RRTs are appropriate for masking binary ratings. Warner[13] proposes a surveying technique that forms the basis of RRTs. The method allows respondents to answer sensitive questions while preserving privacy. Instead of asking sensitive