AIN SHAMS UNIVERSITY

Faculty of Computer & Information Sciences

Information Systems Department

Enhancing privacy approach of recommender systems

A Thesis submitted to Information Systems Department,
Faculty of Computer and Information Sciences, Ain Shams University,
in partial fulfillment of the requirements for
the degree of Master of Science in information Systems

By

Reham Mohamed Kamal

B.Sc. in Computer and Information Sciences,
Information system Department,
Faculty of Computer and Information Sciences, Ain Shams University.

Under Supervision of

Dr. Rasha Ismail

Associate Professor, Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University.

Dr. Wedad Hussein

Assistance Professor, Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University.

Acknowledgement

I deeply thank God for providing me strength, power and patience all the way and inspiring me from the beginning of my work.

I want to express my gratitude to Professor Dr. Rasha Ismail for guiding me through every step in our work. I also record my thanks to Dr. Wedad Hussein for being supportive and understanding.

I would like to thank Dr.AbdelAziz Elbatta, Modern university of information and technology, for his efforts and supporting me.

Finally, I thank my husband, my family and my work colleagues. I wouldn't accomplish anything without their support and love. They are my bless that pushes me throughout every stage in my life.

Abstract

In the last few decades, recommendation systems has received an iconic representation in the field of information technology. With the noticed rapid advancement of data mining, the issue of privacy has become an inevitable necessity. Hence, the mainstream challenge that accompanies data mining is developing a cutting-edge strategy to protect private information. In this work, we present two frameworks for enhancing and preserving the privacy of data in recommendation systems along with their experimental results and discussions.

In the first framework, we proposed a hybrid strategy data perturbation and query restriction (DPQR) with an improved version of MASK (Mining association rules with secrecy Konstraints) scheme to decrease the complexity of traditional MASK scheme. This hybridization resulted in 49.7% as recommendation precision and privacy degree of 97.4% while the traditional MASK scheme gives only 80% privacy degree. We enhanced our results by adopting non-linear programing and solved the privacy problem as a system of equations by setting the privacy equation to be our objective function, the privacy degree was raised to 99.6% and the recommendation precision reached 59%.

In the second framework, we implemented the DPQR strategy to hide sensitive association rules to overcome the limitation of (Modified Decrease Support of Right Hand Side item of Rule Clusters) MDSRRC algorithm that can hide multiple sensitive items in the right hand side by calculating the sensitivity of each item in the sensitive rule and delete the one with maximum sensitivity value then it repeats the calculations again till no sensitive items exist. The MDSRRC suffers long runtime as it requires multiple database scans, while DPQR can hide sensitive association rules in one database scan. We tested the performance of our framework by measuring hiding failure (HF), artificial rules (AR) and lost rules (LR). As for the HF it was 0% while AF and LR was 0.29% and 42% respectively. Also the runtime improvement is 96.22% compared with MDSRRC algorithm.

Table of contents

List of figures:

List of tables:

Chapter 1 Introduction

1.1 Overview

Data today represent a critical asset. An increasing number of organizations collect data, very often concerning individuals, and use them for various purposes. For example, scientific organizations need to collect data about individuals for medical purposes. Needless to say, that collecting data about individuals is essential for demographic and market analysis or study. Therefore, sharing an individual's or organization's data with third parties is an inevitable challenge to privacy. Hence, the issue of privacy has become pivotal in recent studies and researches [1].

Recommender systems are a subclass of information filtering that can provide recommendation or suggestions for users, or can predict user rating and preferences. It depends on developing a framework of detailed personal data, taking advantage of the users' preferences such as ratings, consumption history, as well as personal profiles. In one sense, one must admit the usefulness of data mining; in another sense, one should not undermine the risks that arise consequently [1]. The main issue is that recommender systems may abuses the original data for the interest of the system. Therefore, an accurate system that preserves or restricts the private data available for the recommendation system is urgently needed [2].

The techniques to generate recommendations for users strongly rely on information gathered from the user. The user can provide this information as in profiles, or the service provider can observe users' actions, such as click logs. On one hand, more user information helps the system to improve the accuracy of the recommendations. Alternatively, the information about users creates a grave privacy risk since there is no solid assurance for the service provider not to mismanage the users' data.

1.2 Motivation

Data Mining is the process of pulling out useful knowledge from large amounts of data. Data should be manipulated in such a sensitive way that information cannot be found through Data Mining techniques, which has increased the disclosure risks when the data is released to outside parties. This scenario leads to the research of privacy-preserving data mining where the challenge is to provide accurate results while preserving the privacy of data. [3]

The power of data mining tools to extract hidden information from large collections of data leads to increased data collection efforts by companies and government agencies. Naturally, this raised privacy concerns about collected data. Therefore, after the data miners collect large amounts of private data from data providers, the data needs to be perturbed[1] in different ways to avoid the privacy disclosure, as well as to keep some useful patterns for further data mining. [4]

1.3 Objective

Existing privacy-preserving recommendation systems attempt to protect most of the Customer's personal information like, the customer's browse and search information in a local client. Nevertheless, their role has proved to be limited when external staff participating in the construction of the recommendation system take charge of data processing.

Since the need to protect private information continue to strengthen, privacy protection issues in data mining became a pressing issue. Privacy Preserving Data Mining (PPDM) is a method which can obtain accurate data mining results even with imprecise access to the original data.

So, the objective of this thesis is to augment the privacy of data in recommender systems by applying privacy-preserving data mining techniques with minimum effect on recommendation accuracy.

In this work we suggest two privacy preserving frameworks for recommender systems, the first framework, is based on improved (Mining Association rule with Secrecy Konstraints) MASK scheme and DPQR (data perturbation and query restriction strategy). Where a modified version of MASK algorithm is implemented to improve the time-efficiency, the calculations to generate an inverse matrix is achieved by dividing the matrix into blocks. Also, the number of scanning database is optimized using the set theory. In DPQR strategy three perturbation parameters are chosen according to data characteristics to distort data with 0-1 probability distribution resulting in new sanitized data with protective factors, with the help of these two strategies we were able to generate accurate recommendations without disrupting data quality and privacy. Also in this framework we adopted non-linear programming strategy to select optimum privacy parameters .The results showed that the proposed framework is capable of providing the maximum security for the information available without decreasing the accuracy of recommendation. The second framework is based on hiding sensitive association rules by using DPQR strategy to

[1] Data perturbation is considered a relatively easy and effective technique for protecting sensitive data from unauthorized use.

hide sensitive association rules by applying the DPQR strategy only to those items contained in the sensitive rules in one iteration which obviously reduced the run time and generated accurate recommendation results, but with some limitations in the artificial rule (AR) and lost rule (LR) measures that caused the accuracy of recommendations to decrease than our first framework.

1.4 Thesis organization

The thesis chapters are organized as follows:

Chapter 2 gives a detailed background on the types of recommender systems and their techniques. Also it discusses the concepts and techniques of hiding sensitive association rules. Chapter 3 overviews the recent research efforts in the field of privacy preserving data mining techniques as well as the state of the art for hiding sensitive association rules. Chapter 4 presents and explains our MASK and DPQR framework along with its experimental results. Chapter 5 presents Hiding sensitive rules using DPQR framework as well as discussing the experimental results. Finally in chapter 6 we conclude our work and suggest the future work.

Chapter 2 Background

2.1 Introduction

The growth of internet resulted in information abundance. In other words there are a lot of information but with lack of knowledge, and this is mainly due to the amount of data being published on one side and the inadequacy of techniques to process the data to knowledge on the other side. This situation leads to the demand for new techniques that can help in discovering resources of interest among the massive options presented to us. All of this lead to the introduction of recommender systems which aim to recommend items of interest to specific users by predicting a user's interest in an item based on related information about the items, the users and the interactions between items and users.

Recommender systems are software tools that can give suggestions about items to users. It provides recommendation for user to help in decision making such as where to plan a tour, what items can be bought, what exciting news are there to read etc. It has the ability to cope with the information overload problem [5].

Both users and providers can benefit from recommender systems. As for the user, it recommend items that are of interest, narrows the number of choices and help in finding new items. As for the service providers, it ensures the uniqueness of personalized service for users, strengthens users' trust and loyalty, increases items sale and provide chances for promotions as well as gaining knowledge about users for better market analysis [5].

Recommender systems are of great benefit, however privacy risks related to users' personal data collection and processing are usually undervalued or ignored. Many users are not adequately aware if and how much of their data is collected and if such data is sold to third parties, or how securely it is stored and for how long.

In this chapter, we will discuss and explain different types of recommender systems and the methodologies and techniques used for preserving the user privacy while generating accurate recommendations.

2.2 Recommender systems

All recommender systems act in a similar manner that is: in order to generate personalized accuarte recommendations, they gather information on the attributes, demands, or interrests of the user. Generally, the more detailed the information related to the user is, the more accurate the recommendations are.

The information aquired by recommender systems can be supplied automatically due to users interactions with the recommender systems and making choices. For example, page views on "Ebay" are used to automatically show a selection of recommended similar items (recommendations for you). Similarly, recommended videos on Youtube are altered to the recently viewed videos. Based on purchases by other users, items on Amazon are accompanied by package deals (frequently bought together) or related items (customers who bought this item also bought). Based on sites visited, Google serves and suggests personalized advertisements. Based on your friends and social interactions, Facebook suggests and recommends new friends. In LinkedIn, based on a user's cv and connections, it suggests interesting companies, job offers, and news. Vice-versa, LinkedIn also recommends people to recruiters posting new job openings. In this way, users build their own profile specifying their likes and dislikes, or containing general information (such as age and gender) about themselves., or the information can specifically added by the user [1].

A recommender system suggests a set of items (e.g. content, solutions, or other users) that is most relevant to a particular user of the system. Typically, recommender systems achieve this by predicting relevance scores for all items that the user has not seen yet. Items that receive the highest score get recommended (typically the top-N items, or all items above a threshold). Typically, systems look at similarities and correlations between items, similarities between users, or relations between particular types of items and particular types of users.

Recommender systems take into account a combination of multiple factors to provide good recommendations. They include: [1]

1. The type of data available for the system like personal data, demographic data , browsing data or contextual data.
2. The algorithm like Bayesian networks, genetic algorithms, probabilistic approaches, nearest neighbor strategy etc.
3. The technique used for filterng like collaborative filtering, content-based or hybrid-techniques.

The results of the recommender system are also influenced by:

1. The system performance.
2. Sparsity of the database.
3. Objective of the system.

In the next subsections , we will explain the types of recommender systems and algorithms used. Figure 2.1 shows these types. There are basic types of recommender systems and improved ones. The Improved recommender types are built upon basic recommender types by making hybridization between them or adding new information [1].

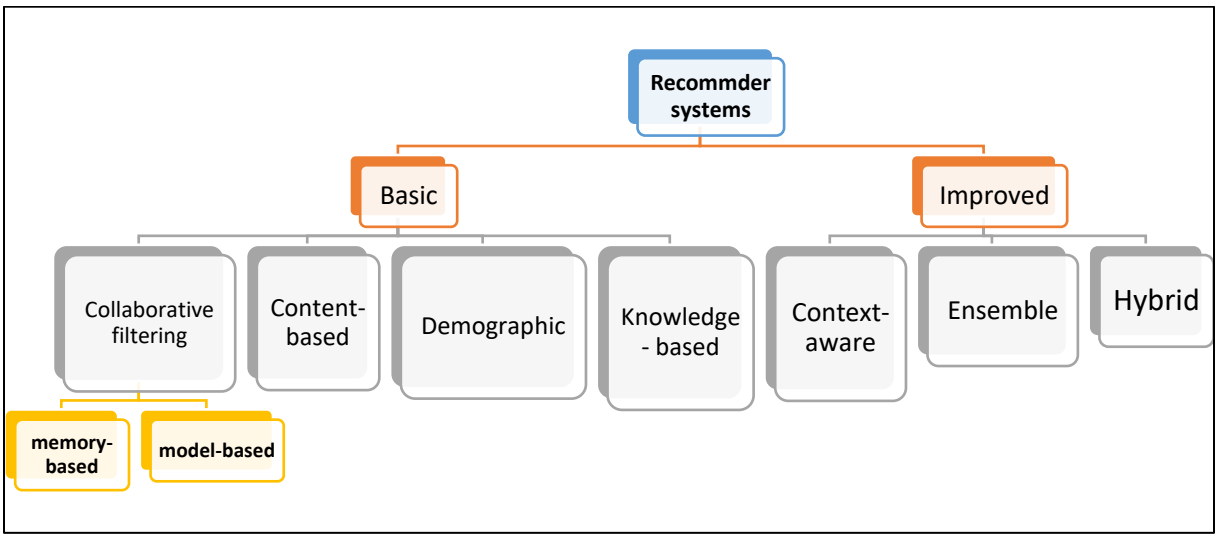

Figure 2.1 Recommender systems types

2.2.1 Collaborative Filtering

It is based on the human psychology of a person asking friends and family for suggestions, so that it helps the person to make a decision. In collaborative filtering, each user rates items. These ratings determine similarity between either users (similar users like similar items) or items (users like items similar to highly rated items). Different metrics exist to compute similarity. The historical data available helps to build the user profile and the data available about the item is used to make the item profile. Both the user profile and the item profile are used to make a recommendation system. [6]

Collaborative filtering technique can be categorized into two types: memory-based and model-based. Predictions for memory based approach make use of the user database completely. Statistical methods are used by the system to find like-minded set of users or neighbors who share similar interests with the active user [7].The

implementation of a memory based system can either be item-based or user-based as shown in Figure 2.2 .

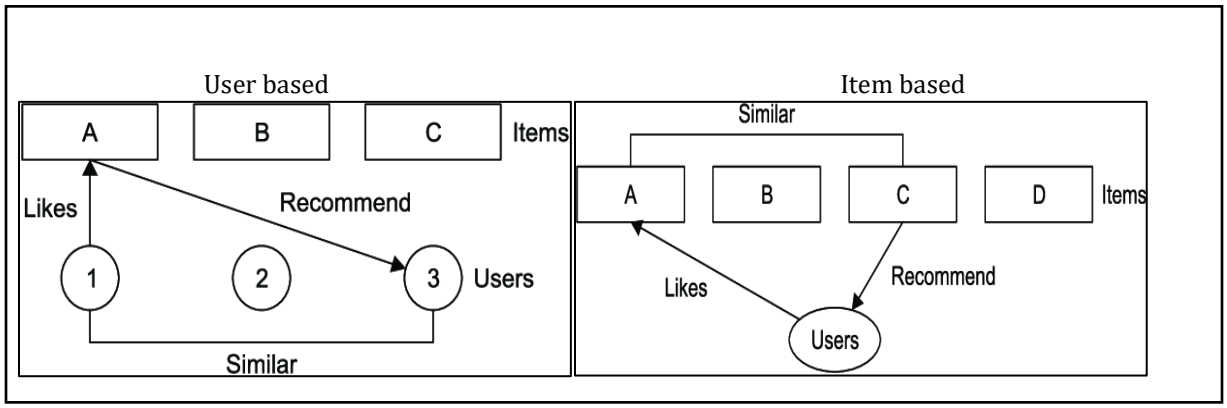

Figure 2.2 memory based collaborative filtering

In item-based collaborative filtering, we are interested in the relationship between different items that are purchased or viewed together. If two or more items are bought or viewed frequently by different users then those items most probably share a close relationship(Eg: Bread and Jam or Bread and Butter or Peanut Butter and Jam etc). So if once the relationship is established then the next time a user adds bread to his cart he'll be given jam or peanut butter as recommendations. These recommendations make more sense than recommending something totally unrelated. [6]

In user-based collaborative filtering the target is the users rather than items. We find the similarity between the users based on their behavior and ratings. This is achevived by buildng a detaild user profile for every user which grows with the interaction of the user with the system. Similarity shared between the users is one of the important factors of recommender systems. If a group of users share similar interest then some items liked by one user might not be rated or used by the other user, so recommending that item to the user has a very high probability of acceptance by the new user. This is also a very successful way of recommending items to users. [6]

Unlike the memory-based approach, the model based collaborative filtering method uses the user database to learn a model which is in turn used for making predictions. Models are created based on classification and clustering techniques to recommend items using user-item set. When designing a model that is capable of making predictions to a user the strength of both data mining and machine learning algorithms are collectively used [8].