Defending recommender systems: detection of profile injection attacks

Chad A. Williams · Bamshad Mobasher · Robin Burke

Received: 4 May 2007 / Revised: 19 July 2007 / Accepted: 20 July 2007 / Published online: 21 August 2007
© Springer-Verlag London Limited 2007

Abstract  Collaborative recommender systems are known to be highly vulnerable to profile injection attacks, attacks that involve the insertion of biased profiles into the ratings database for the purpose of altering the system’s recommendation behavior. Prior work has shown when profiles are reverse engineered to maximize influence; even a small number of malicious profiles can significantly bias the system. This paper describes a classification approach to the problem of detecting and responding to profile injection attacks. A number of attributes are identified that distinguish characteristics present in attack profiles in general, as well as an attribute generation approach for detecting profiles based on reverse engineered attack models. Three well-known classification algorithms are then used to demonstrate the combined benefit of these attributes and the impact the selection of classifier has with respect to improving the robustness of the recommender system. Our study demonstrates this technique significantly reduces the impact of the most powerful attack models previously studied, particularly when combined with a support vector machine classifier.

Keywords  Attack detection · Bias profile injection · Collaborative filtering · Recommender systems · Attack models · Support vector machines

1 Introduction

Recommender systems have become a staple of many e-commerce web sites, yet significant vulnerabilities exist in these systems when faced with what have been termed “shilling” attacks [5,2,9,15]. We use the more descriptive phrase “profile injection attacks”, since promoting a particular product is only one way such an attack might be used. In a profile injection attack, an attacker interacts with a collaborative recommender system to build within it a number of profiles associated with fictitious identities with the aim of biasing the system’s output.

It is easy to see why collaborative filtering is vulnerable to these attacks. A user-based collaborative filtering algorithm collects user profiles, which are assumed to represent the preferences of many different individuals and makes recommendations by finding peers with like profiles. If the profile database contains biased data (many profiles all of which rate a certain item highly, for example), these biased profiles may be considered peers for genuine users and result in biased recommendations. This is precisely the effect found in [9] and [15].

Our prior work [2,3] identified a number of attack models, based on different assumptions about attacker knowledge and intent. The overall conclusion is that an attacker wishing to “push” a particular product (make it more likely to be recommended) or to “nuke” it (make it less likely to be recommended) can do so with a relatively modest number of injected profiles, with a minimum of system-specific
knowledge and with only the kind of general knowledge about likely user ratings distribution that one might find by reading the newspaper. We also know that profile injection attacks are not merely of theoretical interest, but have been uncovered at e-commerce sites.

As prior work has shown, if commercial recommendation systems are not protected, there is a very real risk the quality of the predictions and thus the consumer trust in the site can be compromised by attackers. The goal of this work is to address this vulnerability and provide tools and techniques web site owners may apply to protect their recommender services. Through techniques such as the one outlined in this paper, additional security and trust can be added to increase the robustness of recommendation systems used in the future for commercial sites.

The primary contribution of this paper is a description of an approach to detecting profile injection attacks with supervised classification. The technique is based on identifying characteristics of profiles that may be engineered to increase the influence of a malicious profile on the collaborative system. This is accomplished through a three pronged strategy to creating attributes to facilitate attack classification. This strategy combines attributes for detecting general ratings anomalies, similarity to reverse engineered attacks, and target concentrations; for use in a supervised approach to attack classification. A classifier is then built to distinguish attack profiles from genuine user profiles by constructing training data from authentic profiles and attacks generated by reverse engineered attack models. The combined effectiveness of this approach is then evaluated with the supervised classification algorithms k nearest neighbor (kNN), C4.5, and support vector machine (SVM). This study shows this defense technique when combined with the detection attributes described in this work and a robust classifier such as SVM, can nearly eliminate the impact of the most effective reverse engineered profile injection attacks for all but the largest attacks. We examine the impact the dimensions of attack type, attack intent, filler size, and attack size have on the effectiveness of such a detection scheme.

In Sect. 4, we provide a detailed description of our detection technique and the attributes used in this study. These attributes include both generic attributes that capture expected distribution of user data within profiles, as well as attributes based in the characteristics of well-known attack models. This is followed by our empirical analysis of the resulting detection classifier in Sect. 5.

2 Background and motivation

Researchers have shown that collaborative recommender systems, the most common type of web personalization system, are highly vulnerable to attack. Attackers can use automated means to inject a large number of biased profiles into such a system, resulting in recommendations that favor or disfavor given items. Since collaborative recommender systems must be open to user input, it is difficult to design a system that cannot be so attacked. Researchers studying robust recommendation have therefore begun to study mechanisms for defending against such attacks.

Defense against profile injection can take many forms. Some collaborative algorithms are more robust than others against such attacks. Recent research has focused on techniques that can be used to protect the predictive integrity of collaborative recommenders from this type of malicious biasing. This work falls into two categories: techniques that increase the robustness of the recommender; and techniques for detecting and discounting biased profiles, like this work.

Motivating example

In this paper we consider attacks where the attacker’s aim is to introduce a bias into a recommender system by injecting fake user ratings. In a profile injection attack, an attacker interacts with the recommender system to build within it a number of profiles with the aim of biasing the system’s output. Such profiles will be associated with fictitious identities to disguise their true source.

An attack against a collaborative filtering recommender system consists of a set of attack profiles, each containing biased rating data associated with a fictitious user identity, and including a target item, the item that the attacker wishes the system to recommend more highly (a push attack), or wishes to prevent the system from recommending (a nuke attack).

We provide a hypothetical example to help illustrate the vulnerability of collaborative filtering algorithms, and will serve as a motivation for defending against such attacks. Consider, as an example, a recommender system that identifies books that users might like to read using a user-based collaborative algorithm [7]. A user profile in this hypothetical system might consist of that user’s ratings (in the scale of 1–5 with 1 being the lowest) on various books. Alice, having built up a profile from previous visits, returns to the system for new recommendations. Figure 1 shows Alice’s profile along with that of seven genuine users. An attacker, Eve, has inserted attack profiles (Attack1–3) into the system, all of which give high ratings to her book labeled Item6. Eve’s attack profiles may closely match the profiles of one or more of the existing users (if Eve is able to obtain or predict such information), or they may be based on average or expected ratings of items across all users.

Suppose the system is using a simplified user-based collaborative filtering approach where the predicted ratings for Alice on Item6 will be obtained by finding the closest