DETECTION OF SHILLING ATTACKS IN CF BASED RECOMMENDER SYSTEMS USING HYBRID APPROACH

Authors
Archi Garg
Ayushi Agarwal
Riya Kumari
Shruti Chauhan
B.Tech (IT & MI), University of Delhi, Delhi-110007, India

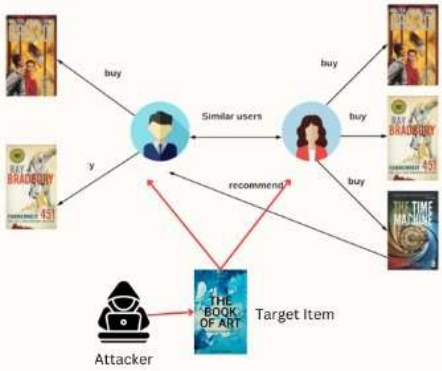
Cluster Innovation Centre — संकुल नवप्रवर्तन केंद्र
...Evolving Senses Dissolving Boundaries...

Mentor
Dr. Anjani Kumar Verma,
Assistant Professor,
University of Delhi, Delhi-110007, India

INTRODUCTION

The recommendation system is broadly used as a crucial tool in different areas from e-commerce sites like Amazon, to entertainment sites like Netflix to provide users the most appropriate recommendations.

Since such systems are open and accessible, they are prone to malicious attacks, primarily shilling attacks in which malicious user profiles are injected into the system to push or nuke the outcomes of targeted items. Therefore, it is very important to detect shilling attacks in order to protect the recommendation systems.



SHILLING ATTACKS

MOVIE ID	USER ID	RATING	TIMESTAMP
M1	U1	4	09:54:22
M2	U2	3	07:23:04

DATASET

OBJECTIVE

This project aims to be an extensive study of the shilling attack models, detection attributes, and detection algorithms. Moreover, it proposes an ensemble model approach to classify the attributes of the inserted profiles that are exploited by the detection algorithms

ATTACK MODELS

- Random Attack : Except for the target item, the items rated by attack profiels are chosen randomly.
- Average Attack : Items are chosen randomly based on the individual item's rating distribution.
- Bandwagon Attack : Generated attacker profiles are filled with higher ratings of popular items.
- Love/Hate Atack : Target item is given least & highest ratings.

:

ATTRIBUTES

$$\mathrm{RDMA}_u = \sum_{i=0}^{Nu} \frac{\frac{|r_{u,i} - \bar{r}_i|}{NR}}{N_u}$$

$$\mathrm{LengthVariance} = \frac{|\#score_j - \overline{\#score}|}{\sum_{i=0}^{N} (\#score_i - \overline{\#score})^2}$$

$$\mathrm{WDA}_x = \sum_{x=0}^{T_u} \frac{|r_{x,i} - \bar{r}_i|}{R_{x,i}}$$

$$\mathrm{WDMA} = \sum_{i=0}^{N_u} \frac{\frac{|r_{i,j} - \bar{r}_j|}{NR_j^2}}{N_u}$$

METHODOLOGY

Identify attack profiles (fake accounts, biased ratings/reviews)

↓

Create genuine profiles as benchmark for comparison

↓

Extract features (RDMA, WDA, length variance, WDMA)

↓

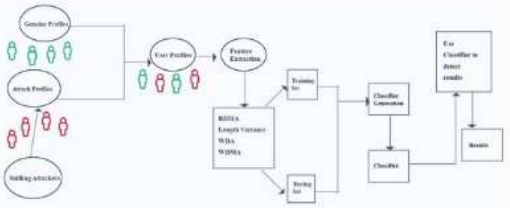
Split data into train and test sets

↓

Generate classifier with ML algorithm and evaluate on test set

↓

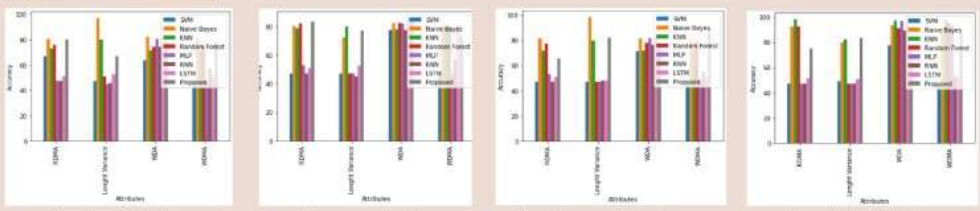
Generate results by predicting genuine/fake profiles.



CONCLUSION

- In this project, first we discussed the various shilling attacks and describe them.
- Second, we used the various detection attributes which are widely used in multiple detection techniques to analyze different algorithms for the detection of shilling attacks.
- Lastly, we proposed an ensemble model for the classification and identification of shilling attacks.
- In the future, we will try to build a working web application with built-in models used in the project.

RESULTS



Random Attack | Average Attack | Bandwagon Attack | Love/Hate Attack

- Conducted several experiments to compare the performance of the proposed algorithm with different classifiers and existing techniques
- Compared different classifiers for each attack model, using attributes RDMA, Length Variance, WDA, and WDMA
- Figures show the comparative performance of classifiers to identify the best classifier.
- Our proposed ensemble of SVM+KNN+Logistic Reg+Decision Tree outperformed the individual algorithms and we plan to further increase the accuracy.