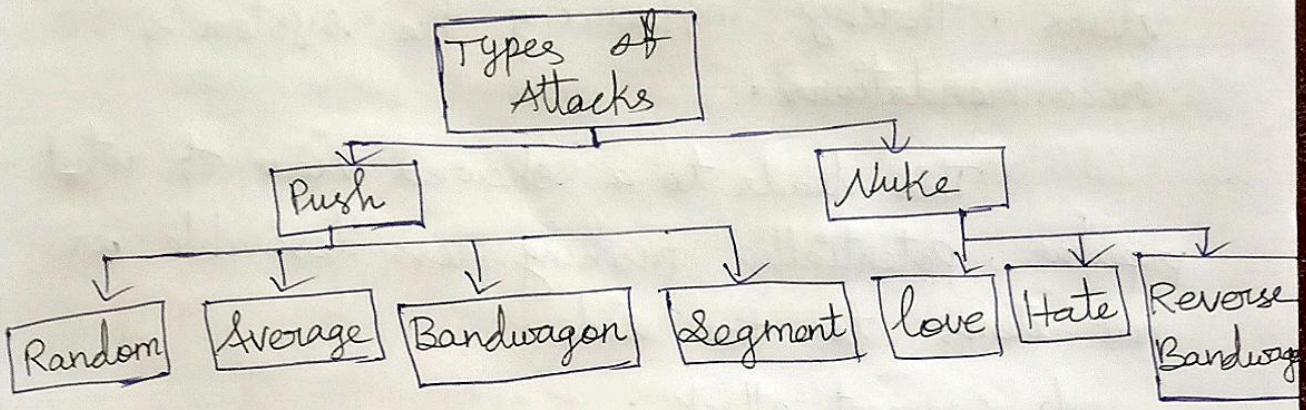
Unit 4 - Recommender systems unit 4 . 2021 regulation

Recommender System (Anna University)

Unit-4

Attack-Resistant Recommender Systems

1) Types of Attacks:-



1. Push Attacks:

a) Random Attack :-

A random attack in the context of recommender systems typically refers to a type of attack where the attacker randomly interacts with items in the system to manipulate or disrupt the recommendation process.

The goal of such attacks is often to introduce noise or bias into the data used by the recommender system, thereby affecting the accuracy and quality of recommendations.

b) Average Attack :

An average attack is a type of shilling attack where malicious users insert fake profiles that rate items around the average rating for each item.

This attack aims to manipulate the system into recommending specific items by injecting biased profiles that don't deviate significantly from the existing rating distribution.

c) Bandwagon attack :-

A bandwagon attack is a type of shilling attack where malicious actors create fake user profiles to inflate the popularity of specific items, thereby influencing the system's recommendations.

This leads to a skewed view of what is popular., potentially pushing less desirable or new items to the side.

d) Segment attack :-

Segment attack is a type of shilling attack that targets a specific group of users with similar interests, biasing the system's recommendations in favor of against certain items for that group.

Instead of trying to influence all users, attackers focus on manipulating recommendations for a particular segment of the user base.

2) Nuke Attacks :-

a) Love Attack :-

Love attack is a type of malicious manipulation where attackers artificially influence the system's recommendations by injecting fake user profiles or ratings that favor specific items.

b) Hate Attack :-

Hate attack is a type of malicious activity where an attacker manipulates the system by providing extremely low ratings to a target item while giving high ratings to other, often irrelevant items.

c) Reverse Bandwagon attack :-

A reverse bandwagon attack is a profile injection attack where malicious users inject fake profiles to make a target item less likely to be recommended.

This is achieved by associating the target item with items that are generally disliked by users, leading the system to predict low ratings for the target.

---

2) Detecting Attacks on Recommender Systems :-

Detecting attacks on recommender systems requires the implementation of robust techniques & algorithms that can identify suspicious patterns, anomalies or manipulations within the system.

They are

Anomaly Detection :-

Anomaly detection techniques are used to identify unusual or suspicious patterns in user behaviour, item ratings or system interactions.

By monitoring deviations from ~~malicious activity~~ expected norms, anomaly detection algorithms can flag potentially malicious activity such as unusual spikes in user feedback, abnormal rating distributions, signaling the presence of an attack.

Behavioral Analysis :-

Behavioral analysis can be used to detect malicious attacks by identifying deviations from normal user behavior patterns such as rating patterns or interaction frequency.

Data sanitation and Preprocessing:-

These are crucial for building accurate and reliable and recommendation systems, especially when protecting against attacks like shilling and data poisioning.

These techniques ensure data quality, consistency and relavance, leading to better model performance and attack resilience.

Model Robustness Checks:-

Model robustness checks in detecting attacks involve evaluating how well a model performs when subjected to malicious attacks, such as shilling attacks, where attackers attempt to manipulate recommendations.

User and Item Reputation Analysis:-

It is a crucial aspect of detecting attacks, particularly shilling attacks, where attackers manipulate ratings to boost certain users or items.

Adversarial Training and Testing:-

Adversarial training involves exposing the model to manipulated inputs during training enabling it to learn to identify and resist these manipulations.

Adversarial testing focuses on evaluating the model's behavior under adversarial conditions, identifying vulnerabilities and weaknesses.

3) Strategies for Robust Recommendation Systems :-

1. Data Quality Assurance :-

Ensure the quality of your data by regularly cleaning and preprocessing the dataset to remove outliers, noise or irrelevant information.

Implement data validation and data integrity checks to detect and handle inconsistent data.

2. Anomaly Detection :-

Employ anomaly detection algorithms to identify unusual patterns in user behavior, ratings or interactions.

3. Model Robustness :-

Use robust machine learning algorithms and models that can handle noisy or adversarial data effectively.

Regularly assess the model's performance and conduct stress tests to identify and rectify vulnerabilities.

4. Data Augmentation ~~and~~ Protection :-

Implement measures to detect and mitigate data augmentation attacks, such as detecting and removing fabricated interactions or fake profiles.

5. Privacy Preservation :-

Apply differential privacy techniques, to protect user privacy and ensure that individual preferences remain confidential.

6. Model defence and Adversarial Training :-

Incorporate adversarial training during the model's training phase to expose it to potential attack scenerios and enhance its resilience.

7. Content Verification :-

Verify the integrity of item content, reviews, and metadata to detect potential content poisioning attacks.

8. User and Item Reputation Analysis :-

Monitor user and item reputation to identify suspicious activities.

Develop trust-aware algorithms that consider reputation scores when making recommendations.

9. Secure data handling :-

Use encryption and secure data storage practices to protect sensitive user information and recommendation models.

Implement secure assess controls to prevent unauthorized access to user data or system resources.

10. User feedback Validations :-

Validate user feedback by techniques like sentiment analysis and opinion mining to filter out fake or manipulative feedback.

11. Feedback Loops & Continuous Monitoring :-

set up feedback loops that allow users to provide feedback on the recommendations they receive.

continuously monitor the system's performance and user feedback to detect and respond to emerging issues.

12. Regular Updates and Security Batching :-

Keep the recommender system software and libraries up to date to address vulnerabilities and security threats.

Stay informed about emerging security threats and adapt the system's defenees accordingly.

12. Education & Awareness :-

Educate users about the risks associated with manipulation and fraudlent behaviar within the system.

Promote awareness of security and Privacy best practices among users and system administrators.

—x—x—x—x—x—x—

4) Detecting Attacks using Algorithms in RS

1. Supervised attack detection algorithms
2. Unsupervised attack detection algorithms

Supervised attack detection algorithms are ability to learn from the underlying data. it is oftendifficult to obtain example of attack Profile detection methods are either individual Profile detection or Group profile detection methods.

Individual Attack Profile Detection :-

This method focus on identifying unique patterns of behavior associated with specific attacks rather than broad categories.

These methods often leverage machine learning & data analysis to distinguish between normal & malicious user activities and can be implemented online to detect attacks in real-time.

1. Number of prediction Differences (NPD) :-

For a given user, the NPD is defined as the number of prediction changes after removing that user from the system.

2. Degree of Disagreement with others (DA) :-

The degree of disagreement with others is a key factor in identifying potentially malicious users, especially in ~~malicious~~ recommender systems where users are expected to have similar tastes.

3. Rating Deviation from Mean Agreement :-

RDMA is a metric used to identify potential attack profiles, particularly those involved in shilling attack where malicious users manipulate ratings.

It quantifies how much a user ratings deviate from the average rating of items they have rated, provide a measure of their rating consistency & potential for manipulation.

$$RDMA = \sum (r_{u,i} - \mu_i) / R_i$$

4. Standard Deviation in user ratings :-

It can be used to identify anomalies & potentially malicious behavior. The formula of standard deviation is

$$\sigma = \sqrt{\Sigma (x_i - \mu)^2 / N}$$

5. Degree of similarity with top k neighbours :-

Using k-nearest neighbours, the degree of similarity with the top k neighbours is measured using a distance metric.

Commonly, euclidean distance or cosine similarity is used.

Group Attack Profile Detection :-

In these cases, the attack profiles are detected as groups rather than as individuals.

The basic principle is that the attacks are often based on group of related profiles, which are very similar.

Therefore, many of these methods perform the detection to use clustering strategies to detect attacks.

Preprocessing Methods :-

The most common approach is to use clustering to remove fake profiles.

Because of the way in which attack profiles are designed authentic profiles and fake profiles create seperate clusters.

This is because many of the ratings in fake profiles are identical, and are therefore more likely to create tight clusters.