DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4550

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A REVIEW OF ATTACKS AND ITS DETECTION ATTRIBUTES ON COLLABORATIVE RECOMMENDER SYSTEMS

Saakshi Kapoor, Dr.Vishal Gupta and Rohit Kumar
Department of Computer Science and Engineering,
UIET, Panjab University,
Chandigarh, India,

Abstract: Today, there is lots of information available over the Internet but it's very difficult to filter out the required information from this overload of information. Thus a solution to this problem, came as "Recommender Systems", they can predict outcomes according to user's interests. Although Recommender Systems are very effective and useful for users but the mostly used type of Recommender System i.e. Collaborative Filtering Recommender System suffers from shilling/profile injection attacks in which fake profiles are inserted into the database in order to bias its output. This paper is aimed at discussing various attacks that can affect Recommender Systems and the attributes that are used for the detection of these attacks.

Keywords: Recommender Systems, Shilling attacks, Generic attributes, Model specific attributes.

1. INTRODUCTION

Weare living in an era of information overload that means, we get more information than what we actually want and sometimes even the information we get is not actually relevant to what we actually wanted. Thus one tool developed to tackle such problems is Recommender System. Recommender Systems[4] can filter out information required by user from the vast amount of information available using certain characteristics and thus this concept is very helpful in overcoming the problem of information overload. Recommender Systems can broadly be categorized as Content-based[12], Collaborative[3, 13, 14] and Hybrid[2] Recommender Systems (Table 1 gives an overview of different techniques of recommender systems). Collaborative Recommender Systems are quite helpful in many ways but they are still prone to shilling or profile injection attacks due to their natural openness. In these attacks, malicious users are inserted into existing dataset in order to influence the result of Recommender Systems. Mostly these attacks are generated by product sellers or developers who aim to promote their own product or demote their competitor's product.

Based on different assumptions attack models can be divided in different categories such as push[16] or nuke[16] attacks and standard[3] or obfuscated[15] attacks which we will be discussing in detail further.

In this paper we present a review of different attacks on Collaborative Recommender Systems and different attributes used for their detection. This paper is organized as follows: section 2 describes different attack types, section 3 describes various detection attributes and finally in section 4 we conclude the paper along with possible future scope.

Table 1.Overview of Recommender System Techniques

Sr. No.	Technique	Characteristics	Techniques used	Pros	Cons	Examples
1.	Content-based Filtering [1,12]	Recommend items to user X similar to previous items rated highly by X.	Bayesian Classifiers, Cluster Analysis, Decision trees, Artificial Neural Networks	It can recommend users with unique taste	overspecialization	Pandora Radio, Rotten Tomatoes, Jinni
2.	Collaborative Filtering[3,13,14]	Recommend to users, items that were liked by other users who exhibited similar tastes.	Bayesian networks, Clustering, Artificial neural networks, Linear regression, Probabilistic models, Graph	Works for any kind of item	Sparsity, Popularity bias, cold start.	GroupLens, Amazon, LinkedIn

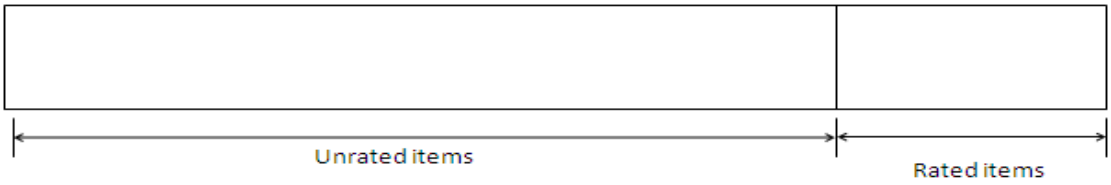
© 2015-19, IJARCS All Rights Reserved

			theory, Matrix Factorization			
3.	Hybrid Filtering[2,4]	Combines both content based and collaborative filtering techniques to overcome disadvantages of both and generate better results.	Works by combining algorithms used in content and collaborative filtering technique and incorporating one component as part of model for the other.	It can overcome the cons of both content-based and collaborative filtering approach.		Netflix, NewsDude

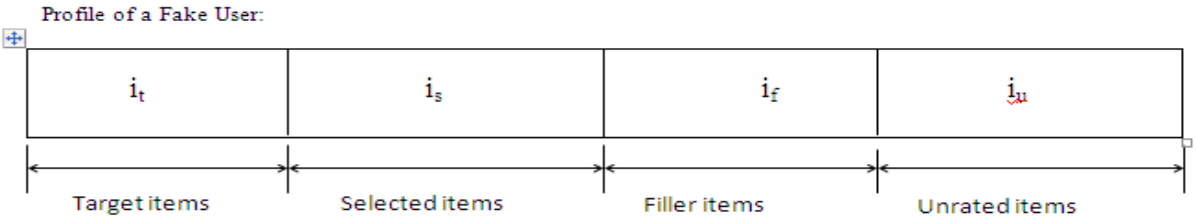
2. ATTACK PROFILES AND ATTACK MODELS

With the advancement of recommender systems, various techniques are employed to influence the output of recommender systems to promote or demote a particular product. These types of attacksare particularly observed in Collaborative Filtering based Recommender Systems which are known as profile injection or Shilling attacks [19], in which malicious users insert fake profiles into the rating database in order to bias the system's output. The general description of the profile of a true user and fake user are characterized below:

Profile of a True User:



Profile of a Fake User:



From above description of trusted and fake user profile it is clear that to attack a recommender system, attack profile need to be designed as statistically identical to genuine profile as possible. So the attacks are based on how an attacker selects ratings for target, selected and filler items. Figure 1gives an overview of different types of attacks.

© 2015-19, IJARCS All Rights Reserved

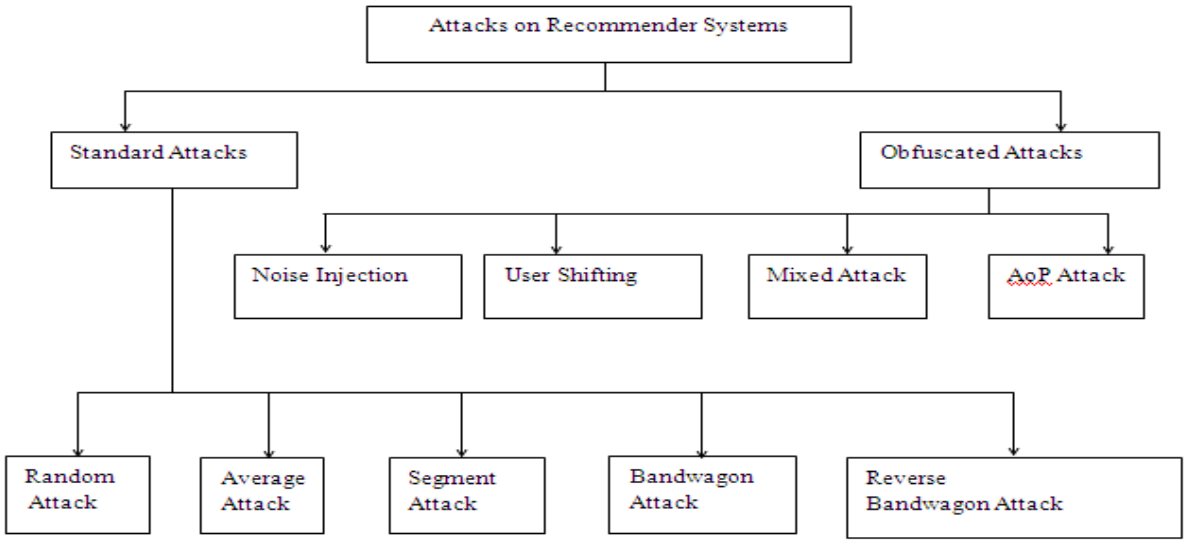


Figure 1 Various Attacks on Recommender Systems

Some of these attacks are described below:

1. Random Attack: In Random Attacks, attack profiles are generated such that their ratings are chosen randomly based on the overall distribution of user ratings in database, except target item. It is very simple to implement but has limited effectiveness (i_s=0, i_f=random, i_t=maximum).
2. Average Attack: In Average Attacks, attack profiles are generated such that the rating for filler items is the mean or average rating for that item across all the users in the database. Although it is a very effective attack but requires prior knowledge about the system (i_s=0, i_f=average, i_t=maximum).
3. Segment Attack: Segment Attack basically targets a specific group of users who may already be interested in the target item. Alternatively, we can say that it increases recommendations for a target product to a certain group of users (i_s=maximum, i_f=minimum, i_t=maximum).
4. Bandwagon or Popular Attack: In Bandwagon Attacks, profiles are generated such that besides giving high ratings to the target items, it also contains only high values for selected items and random values to some filler items (i_s=maximum, i_t=maximum, i_f=random/average).
5. Reverse-Bandwagon Attack: Reverse Bandwagon is a variant of Bandwagon Attack except for the fact that in Bandwagon Attack only high ratings were assigned to target items but here in Reverse Bandwagon Attack, low ratings are given to target and selected items(i_s=minimum, i_t=minimum, i_f=random/average).

All the type of attacks which are discussed above are Standard Attacks [3, 15, 16, 19, 20] and you might have noticed that during our discussion about attacks we are constantly using the term filler items [3], so what basically are filler items. It is simply the ratio between number of items rated by user and number of entire items in dataset. Next we will be discussing about Obfuscated type of Attacks [10, 15, 19].

6. User Shifting: In these types of attacks we basically increment or decrement all ratings for a subset of items per attack profile by a constant amount so as to reduce the similarity between attack profiles.
7. Mixed Attack: In Mixed Attack, attack is on the same target item but that attack is produced from different attack modules.
8. Noise Injection: This type of attack is carried out by adding some noise to ratings according to a standard normal distribution multiplied by a constant, β, which is used to govern the amount of noise to be added. This added noise can be used to affect the generated output.
9. Average over Popular Attack (AoP): AoP attack [15] was designed to obfuscate average attack by choosing filler items with equal probability from top x% of most popular items rather than from whole database.

In addition to above mentioned categories for classification of attacks, attacks can also be categorised as: push[16] and nuke[16] attacks where, in push attacks, higher ratings are given to target items, so as to promote a product while in nuke attacks, lower ratings are given to target items, so as to demote a product. Table 2 gives an overview of different attributes of certain attack models.

Table 2.Overview of attack models

Attack model	Target items(i_t)(push/nuke)	Selected items(i_s)	Filler items(i_f)	Unrated items(i_n)
1. Random attack	Maximum/minimum	0	Random	Ø
2. Average attack	Maximum/minimum	0	Average	Ø
3. Segment attack	Maximum/minimum	maximum	Minimum	Ø
4. Bandwagon attack	Maximum/minimum	maximum	Random/average	Ø

© 2015-19, IJARCS All Rights Reserved

5. Reverse-bandwagon attack	Minimum/maximum	minimum	Random/average	Ø
6. AoP attack	Maximum/minimum	0	x% popular items	Ø

3. DETECTION ATTRIBUTES

Detection attributes can be described as some descriptive statistics that can be used to capture some of the major characteristics that make an attacker's profile look different from genuine user's profile. These can be categorised into two categories as: generic attributes [18, 20] and type-specific attributes [18, 20].Table 3gives an overview of few of these attributes.

3.1 Generic attributes: These are the attributes that can be used for almost all attack types and these are not specific to any particular attack type.

1. Rating Deviation from Mean Agreement (RDMA): RDMA [11] can identify attackers by analysing the profile's average deviation per item or user. It is defined as:

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

where *Tu*is the number of items user xrated, $r_{x,i}$ is the rating given by user *x*to item *i* , $r_i$is the average rating of item *i* , $R_{x,i}$be the number of ratings provided for item *i* by all users and N_x is the number of users.

2. Weighted Degree of Agreement (WDA): WDA[5] can be calculated as the numerator of RDMA.

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

where *Tu*is the number of items user *x*rated, $r_{x,i}$ is the rating given by user *x*to item *i* , $r_i$is the average rating of item *i* , and $R_{x,i}$be the number of ratings provided for item *i* by all users.

3. Weighted Deviation from Mean Agreement (WDMA): WDMA [8] can help identify anomalies by placing a higher weight on rating deviations for sparse items.

$$\text{WDMA}_x = \frac{\sum_{x=0}^{T_u} \frac{|r_{x,i} - \bar{r}_i|}{R_{x,i}^2}}{T_u}$$

where *Tu*is the number of items user *x*rated, $r_{x,i}$ is the rating given by user *u* to item *i* , $r_i$is the average rating of item *i* , and $R_{x,i}$be the number of ratings provided for item *i* by all users.

4. Length Variance (LengthVar): LengthVar[5] is used to capture how much the length of a given profile varies from average length in the dataset. It is particularly effective in detecting attacks with large filler sizes.

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

Where $\#score_j$ is the total number of ratings in the system for user j, and N is the total number of users in the system.

5. Degree of Similarity with Top Neighbours (DegSim): DegSim [9] is used to capture the average similarity of a profile's *k* nearest neighbours.

$$\text{DegSim} = \frac{\sum_{i=1}^{x} Z_{i,j}}{x}$$

Where $Z_{i,j}$ is the Pearson correlation between users i and j, and x is the number of neighbours.

There are certain other generic attributes as well. Some of them are H_v-score, TWDMA (calculated by incorporating trust into RDMA)[17], UnRAP (unsupervised retrieval of attack profiles)[6,7].

3.2 Type-Specific Attributes: Attributes which will be used for certain specific attack types, like some attributes will be for average attack, some for random attack, etc.

1. Filler Mean Variance (FMV)[7]: It is generally used for average attack and is defined as follows:

$$\text{FMV}_u = \sum_{i \in L_f} \frac{(r_{x,i} - \bar{r}_i)^2}{|L_f|}$$

Where L_f is the filler item set, $r_{x,i}$ is the rating given by user u to item i and r_i is the average of ratings assigned to item i.

2. Filler Mean Target Difference (FMTD)[7]: It is generally used for segment and bandwagon attack and is defined as follows:

$$\text{FMTD}_u = \left| \frac{\sum_{i \in L_s} r_{x,i}}{|L_s|} - \frac{\sum_{i \in L_f} r_{x,i}}{|L_f|} \right|$$

Where L_s is the selected item set, L_f is the filler item set and $r_{x,i}$ is the rating given by user u to item i.

3. Mean Variance (MeanVar): MeanVar[5] is generally being used for identification of average attack and is defined as follows:

$$\text{MeanVar}\ (r_{target}, j) = \frac{\sum_{i \in (P_j - r_{target})} (r_{i,j} - \bar{r}_i)^2}{|k|}$$

Where P_j is the profile of user j, r_{target} is hypothesized target item, $r_{i,j}$ is the rating user j has given item i, and r_i is the mean rating of item i across all users.

Other type specific attributes include FMD [5], FAC [5], Profile variance, etc.

Table 3.Overview of detection attributes

Sr. No.	Attribute	Attribute Type	Equation	Description
1.	RDMA	Generic	$\text{RDMA}_x = \frac{\sum_{x=0}^{T_u} \frac{\lvert r_{x,i} - \bar{r}_i \rvert}{R_{x,i}}}{N_x}$	Rating Deviation from Mean Agreement [19]
2.	WDA	Generic	$\text{WDA}_x = \sum_{x=0}^{T_u} \frac{\lvert r_{x,i} - \bar{r}_i \rvert}{R_{x,i}}$	Weighted Degree of Agreement [6]

© 2015-19, IJARCS All Rights Reserved

3.	WDMA	Generic	$WDMA_x = \frac{\sum_{x=0}^{T_u} \frac{\lvert r_{x,i} - \bar{r}_i \rvert}{R_{x,i}^2}}{T_u}$	Weighted Deviation from Mean Agreement [11]
4.	LengthVar	Generic	$LengthVar = \frac{\lvert \#score_j - \overline{\#score} \rvert}{\sum_{i=0}^{N} (\#score_i - \overline{\#score})^2}$	Length Variance [6]
5.	DegSim	Generic	$DegSim = \frac{\sum_{i=1}^{x} Z_{i,j}}{x}$	Degree of Similarity with Top Neighbours [12]
6.	TWDMA	Generic	$TWDMA_u^P = \frac{RDMA_u}{Trust(u)}$	Calculated by incorporating trust into RDMA. [28]
7.	UnRAP	Generic	$H_v(u) = \frac{\sum_{i \in I} (r_{ui} - r_{Ui} - r_{uI} + r_{UI})^2}{\sum_{j \in J} (r_{ui} - r_{uI})^2}$	Unsupervised retrieval of attack profiles [7,9]
8.	FMV	Type specific: average attack	$FMV_u = \sum_{i \in L_f} \frac{(r_{x,i} - \bar{r}_i)^2}{\lvert L_f \rvert}$	Filler Mean Variance [9]
9.	FMTD	Type specific: segment and bandwagon attack	$FMTD_u = \left\lvert \frac{\sum_{i \in L_S} r_{x,i}}{\lvert L_S \rvert} - \frac{\sum_{i \in L_f} r_{x,i}}{\lvert L_f \rvert} \right\rvert$	Filler Mean Target Difference [9]
10.	MeanVar	Type specific: average attack	$MeanVar(r_{target}, j) = \frac{\sum_{i \in (P_j - r_{target})} (r_{i,j} - \bar{r}_i)^2}{\lvert k \rvert}$	Mean Variance [6]
11.	FMD	Type specific: average attack	$FMD_u = \frac{1}{\lvert U_u \rvert} \sum_{i=1}^{\lvert U_u \rvert} \lvert r_{u,i} - \bar{r}_i \rvert$	Filler Mean Difference [6]
12.	FAC	Type specific : random attack	$FAC_u = \frac{\sum_i (r_{u,i} - \bar{r}_i)}{\sqrt{\sum_i (r_{u,i} - \bar{r}_i)^2}}$	Filler Average Correlation [6]

4. CONCLUSION AND FUTURE WORK

The issue of Shilling attacks is a major concern in the field of Recommender Systems, to maintain its trustworthiness we need to either design Recommender Systems in such a way that they are resistant to such attacks or design algorithms which can detect attacks easily and effectively. Furthermore, we should also aim at developing detection attributes for obfuscated attacks.

5. REFERENCES

1. Prem Melville and VikasSindhwani. "Recommender systems." In Encyclopedia of machine learning, pp. 829-838. Springer US, 2011.
2. Robin Burke,"Hybrid recommender systems: Survey and experiments." User modeling and user-adapted interaction 12, no. 4 (2002): 331-370.
3. BamshadMobasher, Robin Burke, RunaBhaumik, and Jeff J. Sandvig. "Attacks and remedies in collaborative recommendation." IEEE Intelligent Systems 22, no. 3 (2007).

© 2015-19, IJARCS All Rights Reserved

4. F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh. "Recommendation systems: Principles, methods and evaluation." Egyptian Informatics Journal 16, no. 3 (2015): 261-273.
5. Zhihai Yang, and ZhongminCai. "Detecting abnormal profiles in collaborative filtering recommender systems." Journal of Intelligent Information Systems (2016): 1-20.
6. Quanqiang Zhou and Fuzhi Zhang. "A Hybrid Unsupervised Approach for Detecting Profile Injection Attacks in Collaborative Recommender Systems." Journal of Information &Computational Science 9, no. 3 (2012): 687-694.
7. ZhuoZhang, and Sanjeev R. Kulkarni. "Detection of shilling attacks in recommender systems via spectral clustering." In Information Fusion (FUSION), 2014 17th International Conference on Information Fusion, pp. 1-8. IEEE, 2014.
8. Chad Williams,RunaBhaumik, J. J. Sandvig, BamshadMobasher, and Robin Burke. “Evaluation of profile injection attacks in collaborative recommender systems.” Technical report. Available from http://facweb. cti. depaul. edu/research/techreports/TR06-006. pdf.(accessed 17.03. 12), 2008.
9. Wei Zhou, Junhao Wen, Yun Sing Koh, ShafiqAlam, and Gillian Dobbie. "Attack detection in recommender systems based on target item analysis." In Neural Networks (IJCNN), 2014 International Joint Conference on Neural Networks, pp. 332-339. IEEE, 2014.
10. QuanqiangZhou. "Supervised approach for detecting average over popular items attack in collaborative recommender systems." IET Information Security 10, no. 3 (2016): 134-141.
11. WilanderBhebe, and Okuthe P. Kogeda. "Shilling attack detection in Collaborative Recommender Systems using a Meta Learning strategy." In Emerging Trends in Networks and Computer Communications (ETNCC), 2015 International Conference on Emerging Trends in Networks and Computer Communications, pp. 56-61. IEEE, 2015.
12. GediminasAdomaviciusand Alexander Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." IEEE transactions on knowledge and data engineering 17, no. 6 (2005): 734-749.
13. DhohaAlmazro,GhadeerShahatah, LamiaAlbdulkarim, Mona Kherees, Romy Martinez, and William Nzoukou. "A survey paper on recommender systems." arXiv preprint arXiv: 1006.5278 (2010).
14. XiaoyuanSu, and Taghi M. Khoshgoftaar. "A survey of collaborative filtering techniques." Advances in artificial intelligence 2009 (2009): 4.
15. RunaBhaumik, BamshadMobasher, and Robin Burke. "A clustering approach to unsupervised attack detection in collaborative recommender systems." In Proceedings of the 7th IEEE international conference on data mining, Las Vegas, NV, USA, pp. 181-187. 2011.
16. FuguoZhang. "A survey of shilling attacks in collaborative filtering recommender systems." In Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference onComputational Intelligence and Software Engineering, pp. 1-4. IEEE, 2009.
17. QiangZhang, Yuan Luo, ChuliangWeng, and Minglu Li. "A trust-based detecting mechanism against profile injection attacks in recommender systems." In Secure Software Integration and Reliability Improvement, 2009. SSIRI 2009. Third IEEE International Conference onSecure Software Integration and Reliability Improvement, pp. 59-64. IEEE, 2009.
18. Mohammad Amin Morid, Mehdi Shajari, and Ali Reza Hashemi. "Defending recommender systems by influence analysis." Information retrieval 17, no. 2 (2014): 137-152.
19. IhsanGunes, CihanKaleli, Alper Bilge, and HuseyinPolat. "Shilling attacks against recommender systems: a comprehensive survey." Artificial Intelligence Review (2014): 1-33.
20. Chad A.Williams, BamshadMobasher, and Robin Burke. "Defending recommender systems: detection of profile injection attacks." Service Oriented Computing and Applications 1, no. 3 (2007): 157-170.

© 2015-19, IJARCS All Rights Reserved