RESEARCH ARTICLE OPEN ACCESS

Strategies of detecting Profile-injection attacks in E-Commerce Recommender System: A survey

Partha Sarathi Chakraborty, Dr. Sunil Karforma
Assistant Professor, University Institute of Technology, The University of Burdwan, Burdwan
Reader, Dept. of Computer Science, The University of Burdwan , Burdwan

Abstract

E-commerce recommender systems are vulnerable to different types of shilling attack where the attacker influences the recommendation procedure in favor of him by inserting fake user-profiles into the system. From one point of view, the attacks can be of type push or nuke-either to promote or to demote a product. On the other hand, attacks can be classified as high-knowledge or low-knowledge attack depending on the amount of system knowledge required for making the attack successful. Several research works have been done in the last two decades for defending attacks on recommender systems. In this paper, we have surveyed the major works done in this area by different researchers. After a brief explanation of different attack types and attack models, we discussed the attack detection strategies proposed by the researchers mainly under five categories- Generic and model specific attribute based, rating distribution based, outlier analysis based, statistical approach based and clustering based.

***Index Terms*— profile-injection attack, rating distribution, popular item, prediction shift, user profile, attack model.**

I. INTRODUCTION

Over the time, the number of customers who like to buy products online is increasing in a rapid way. Often these online customers face the problem of choosing the right product they want from a huge collection of products offered by the e-commerce websites. To help the customers in choosing their desired product, many e-commerce websites use collaborative filtering algorithms.

As the system is open, the customers put their rating on different items online; it is vulnerable to different types of attacks. In profile-injection attacks, attackers insert fake user profiles into the system in order to either promote or demote a product. When the intent of the attacker is to promote a product, the corresponding attack is called a push attack and when the intent is to prevent a product from appearing in the recommendation list then that type of attack is called nuke attack. In case of push attack, the maximum rating is given to the target item and in case of nuke attack, the minimum rating is given. Now the question is how the attacker creates fake user profile which will look like a genuine one so that it can influence the recommendation process in favor of the target item without letting the system know its true identity. In their works, researchers have discussed different attack models. A brief discussion on them has been given in the subsequent section. In this study, we present a survey on the different ways of detecting profile injection attacks against collaborative filtering based recommender systems.

The paper is organized, as follows: Section 2 discusses earlier works related to surveying on detection of profile-injection attacks in recommender systems. In section 3, a brief introduction of different attack models have been given. In section 4, attack detection schemes proposed by the researchers have been categorized and discuss in detail. Finally, a conclusion of the survey has been given in section 5.

II. RELATED WORKS

The Concept of profile injection attack is first introduced by O Mahony[1][2]. Then onwards researchers have proposed different attack models and a various ways of defending attacks generated following those attack models. In their paper, Mehta et al. [3] surveys some robust collaborative filtering algorithms. In another paper[4], authors examined the robustness of several recommendation algorithms that use different model-based techniques: particularly techniques based on k-means and probabilistic latent semantic analysis (pLSA) that compare the profile of an active user to aggregate user clusters, rather than the original profiles. Another survey has been made by Zhang [5] on shilling attack models, algorithm dependence, attack detection, and attack evaluation metrics. Gunes et al.[6] have given a overall picture of different attack types, detection methods, robustness analysis, cost-benefit analysis in this problem domain. In this paper, we have concentrated only on the detection strategies devised for profile-injection attacks proposed by different researchers over the time. We have discussed the detection strategies under different categories named Generic and model specific attributes based, rating distribution based, outlier analysis based, statistical approach based and clustering based.

III. TYPES OF ATTACK

The way in which the attacker will create fake user profiles mainly depends on the amount of knowledge he has about the database of user ratings for different products. According to these criteria, the profile-injection attacks can be categorized as low-knowledge attacks and high-knowledge attacks. Random, Bandwagon, Reverse bandwagon, Love/Hate, Hybrid are examples of low-knowledge attacks whereas Average, Favorite attacks falls in the category of high-knowledge attacks. An attack can be categorized as push attack or nuke attack depending on the intent of the attacker. When the attacker tries to promote a product (or a group of products), then the attack is called as push attack. When the target is to demote a product (or a group of products) then that attack is termed as nuke attack. There are some attack types which can be used both for push and nuke attack. Random, Average, Probe, Hybrid, Favorite attacks are of this type. Bandwagon and segment attack can be used only for push attack whereas Reverse bandwagon, Love/Hate can be used only for nuke attack.

In general, each attack profile can be decomposed into four subsets: a set of unrated items, a set of filler items, set of selected items with particular characteristics determined by the attacker and fourthly, one or more target items. For some models, the selected items are chosen because of their high similarity with the target item and for the other attack models, this subset remains empty. In case of Random Attack [7] a highest or lowest rating value is assigned to the target item depending on the push or nuke attack and random ratings are assigned to the filler items. The selected item set is kept empty for this attack model. In average attack [7], the rating of each filler item corresponds to the mean rating of that item across the database. The profiles that are created following the bandwagon attack model[8] try to associate the target item with frequently rated items by putting them in the selected item set. In the contrary, in Reverse Bandwagon attack model[8] poorly rated items are put in the selected item set to make the nuke attack more effective. The characteristics of the Segment attack[8] is it targets a specific group of users with similar tastes and tries to make the system's response in favor of those users. Love/Hate Attack is a very simple attack and requires no system knowledge. In this attack model, the attack profile consists of minimum(maximum) rating value for target items and maximum(minimum) rating value for filler items for nuke(push) attack.

IV. DETECTION STRATEGIES

A. Generic and model specific attributes based

In order to classify attack profiles, several researchers have used attributes derived from the user profiles in their classification task instead of the raw rating profiles. These attributes fall in two categories-attributes or metrics which are designed for all attack models are reported in the literature as generic attributes and attributes which are specific to a particular type of attack model are called as model-specific or type specific attributes. Both of these two types of attributes try to capture different statistical features of user profiles.

In order to distinguish between genuine and fake profiles Chirita et al.(2005) [9] proposed several metrics. They are Number of Prediction-Differences (NPD), Standard Deviation in User's Ratings, Degree of Agreement with Other Users, Degree of Similarity with Top Neighbors and Rating Deviation from Mean Agreement (RDMA).

NPD of an user X is defined as the number of net prediction changes after removing the profile of X. Standard Deviation in User's Ratings measures the degree of dispersion of a particular rating value for an item from the average ratings of that user. Degree of Agreement with Other Users is the average deviation in a user's ratings from the average rating of each item. The metric Degree of Similarity with Top Neighbors is defined as the average of the similarity values with the Top-K neighbors of a user. Rating Deviation from Mean Agreement (RDMA) measures the deviation of agreement with other users on a set of target items weighted by the inverse rating frequency for these items.

In her paper, Chirita[9] proposes an algorithm which computes all the above mentioned metrics for each user profile and profiles with very high values for NPD, Average Similarity, Degree of agreement with other users, and RDMA along with very low value for Standard Deviation in User Ratings are identified as fake profiles. After Chirita[9] several works have been made by the researchers following the same line of thought i.e by using generic attributes in identifying attack profiles. In their work, Burke et al.[8][10] extended the thought of Chirita et al and provided two variations of the attribute Rating Deviation from Mean Agreement (RDMA). They are Weighted Deviation from Mean Agreement (WDMA) and Weighted Degree of Agreement (WDA). WDMA differs from RDMA in the way that it puts more weight on rating deviations for sparse items which, in turn, provides higher information gain. Weighted Degree of Agreement (WDA) measures capturing the sum of the differences of the profile's ratings from the item's average rating divided by the item's rating frequency. Other than these two RDMA based measures or derived attributes, Burke et al.[8],[10] proposed another attribute named Length Variance(LengthVar) which measures the extent of differences in the number of ratings of a given profile from the average number of ratings in the system- k-NN classifier has been used for identifying attack profiles.

Williums et al.[11] used three classification algorithms namely simple nearest-neighbor classification using kNN, decision-tree learning using C4.5, and support vector machine(SVM) classifier. The attributes they have used are Rating Deviation from Mean Agreement (RDMA), Weighted Degree of Agreement (WDA), Weighted Deviation from Mean Agreement (WDMA), Degree of Similarity with Top Neighbors (DegSim) and Length Variance (LengthVar). They have also shown that support vector machine performed better than the other two in detecting attack profiles generated from known attack models.

The attributes so far discussed, are “generic” in nature. That means, all the above-mentioned attributes derived from the user profiles represented the statistical signature of the profiles irrespective of the nature of the attack models. In order to better understand the difference between attack profiles and the eccentric but authentic profiles, Burke[8],[10] introduced some model-specific attributes namely Filler Mean Variance, Filler Mean Difference and Profile Variance attribute for average attack model and Filler Mean Target Difference (FMTD) attribute for Segment Attack Model. In classifying attack profiles, Williums et al.[11] has also used the model-specific attributes mentioned above along with Filler Average Correlation attribute for Random Attack model. Bhaumik R, Mobasher B, Burke[12] have applied several generic attributes to k-means clustering algorithm and identified profiles belonging to small clusters as attack profiles.

B. Rating Distribution Based

Observing the distribution of item ratings over time for a user can lead researchers to find out attack events in a recommender system. A number of works have been done in this direction. A time series based approach have been developed in paper Zhang et al.[13], which can reveal the presence of a wide range of profile injection attacks. They have grouped a certain number of consecutive ratings into windows and compute the sample average and sample entropy in each window. They have analyzed the time series of the computed sample average and sample entropy to detect attack events. For best detection of attack events, an optimal window size has been derived theoretically. A heuristic algorithm that adaptively changes the window size has also been proposed by them from practical scenario where the number of attacks is unknown. Bhowmik[14] have proposed a Time Interval Detection Scheme where mean and standard deviation of the ratings of an item are monitored for the first k-th time intervals, assuming there are no biased ratings. When new rating comes into the system the mean rating for the new time interval t after the k-th interval, $\overline{x}^{t}{}_{i}$ is calculated and compared with a threshold value by the following condition (1).

$$\overline{x}^{t}{}_{i} > \mu^{k}{}_{i} + Z_{\frac{\alpha}{2}} \frac{\sigma_{i}^{k}}{\sqrt{n}} \quad \text{and}$$

$$\overline{x}^{t}{}_{i} < \mu^{k}{}_{i} - Z_{\frac{\alpha}{2}} \frac{\sigma_{i}^{k}}{\sqrt{n}} \qquad (1)$$

where $\mu^{k}{}_{i}$ and $\sigma^{k}{}_{i}$ are the mean and standard deviation for the first k-th interval for an item i. If $\overline{x}^{t}{}_{i}$ is outside the range, the t-th interval is identified as an attack interval. In their paper, Chakraborty and karforma [15] constructed the rating windows and the time windows from the rating time series of an item. A number of consecutive ratings are taken for constructing the rating windows. Similarly, timestamp values of those ratings are also grouped into time windows of equal size. For better understanding of the distributional changes in the rating windows and time windows over time, if any, due to attack, they made both the rating windows and time windows overlapping in nature.

In order to get indication about any distributional change in the ratings of the item over time, the standard deviation of each rating window is measured. When the system is under attack, the standard deviation of the rating windows corresponding to attack events and the adjacent windows will have a very low value of standard deviation compared to the other rating windows. As this may happen, in case of some products, due to reasons other than attack also, the gap between the timestamp values of the first and last ratings of each time window has been considered in parallel with low standard deviation values in rating windows in declaring an attack event. Threshold values for rating and time windows have been calculated using the statistical characteristics of the rating time series.

C. Outlier Analysis Based

In this approach of attack detection, the fake user profiles, which are injected into the system during attack event, are considered as outliers. In literature, different categories of outlier detection techniques have been reported namely distance based approach, density based approach, clustering based approach and depth based approach. Chakraborty and karforma[16] applied a clustering based approach and used Partition Around Medoid(PAM) algorithm in detecting the fake attack profiles. As indicated in Mehta[17], the attack profiles are highly correlated in nature and percentage of attack with respect to the total database size must be small, members of the small clusters are considered as attack profiles. Following the definition of small cluster given in Loureiro,A., L. Torgo and C. Soares [18], Chakraborty and karforma [16]have identified those profiles as attack profiles that belong to a cluster having size lesser than half the average number of points in the k clusters. Authors have noticed that detection rate is not satisfactory for attack profiles with very low filler percentages-most of the attack profiles resides in large clusters. To identify those profiles in large clusters a PAM-based outlier detection algorithm[19] have been used. Clustering based approach is advantageous over the other approaches of outlier detection in the sense that it is totally unsupervised. But at the same time it should also be noted that clustering algorithms are not optimized for outlier detection. Chakraborty and karforma [20] used Cluster-Based Local Outlier Factor (CBLOF) proposed by He, Xu, Deng[21] and a distance-based approach named k-NN Based Outlier Detection proposed by Knorr[22] in detecting attack profiles.

D. Statistical Approach Based

Bhaumik et al.[23] have shown that statistical process control (SPC) based approach can be effective in detecting items that are likely to be under attack. They investigated two SPC techniques-X-bar control limit and Confidence Interval control limit techniques. They have collected k items with similar rating distribution in the same category from the database. X-bar control limit plots how far away from the average value of a process the current measurement falls and defines upper and lower control limits following Shewart[24]. In case of Confidence Interval control limit technique, the upper and lower boundaries of the confidence interval are derived directly from the Central Limit Theorem and are considered as the threshold for push and nuke attacks respectively. Zhang et al.[25] detected random attacks by computing the log-likelihood of each rating profile given the low-dimensional linear model that best describes the original rating matrix. But attacks following average attack models could not be detected by their approach.

In their paper, Chakraborty and karforma[26] applied a change-point detection algorithm proposed by Moskvina[27]. The algorithm is based on Singular Spectrum analysis for detecting structural changes in the time series. Change-point detection is the process of discovering time-points in a time series where abrupt change in data occurs. The singular spectrum analysis breaks the original time series into a small number of independent and interpretable components such as a slowly varying trend, harmonic terms and a structure-less noise.

E. Clustering Based

We have already discussed works of Bhaumik R, Mobasher B, Burke[12] , Chakraborty and karforma [16] where clustering has been used for attack detection. Bhaumik R, Mobasher B, Burke[12] have done clustering of the user profiles using the generic attributes whereas in papers Chakraborty and karforma[16] and Al- Zoubi[19] clustering have been used as outlier detection techniques. O'Mahony et al.[28] utilized clustering technique to keep away probable attack profiles from being selected as neighbors and thus eliminating them from the recommendation generation process.

V. CONCLUSION

In this survey, we have discussed strategies of detecting profile injection attacks in e-commerce recommender systems. Some strategies have been applied directly on the rating data whereas other strategies have been applied on the attributes derived from the user profiles. We discussed the attack detection strategies proposed by the researchers under five categories. Relative advantage and disadvantages of these approaches have been discussed in some cases. We have mentioned only the broad outline of different strategies avoiding the detailed procedure intentionally.

REFERENCES

[1] O'MahonyMP, Hurley NJ, Silvestre GCM (2002a) Towards robust collaborative filtering. Lect Notes Computer Sci 2464:87–94.

[2] O'MahonyMP, Hurley NJ, SilvestreGCM(2002b) Promoting recommendations: an attack on collaborative filtering. In: Proceedings of the 13th international conference on database and expert systems applications, Aix-en-Provence, France, pp 494–503.

[3] Mehta B, Hofmann T (2008) A survey of attack-resistant collaborative filtering algorithms. IEEE Data Eng Bull 31(2):14–22.

[4] Sandvig JJ, Mobasher B, Burke RD (2008) A survey of collaborative recommendation and the robustness of model-based algorithms. IEEE Data Engineering Bulletin 31(2):3–13.

[5] Zhang FG(2009c)Asurvey of shilling attacks in collaborative filtering recommender systems. In: Proceedings of the international conference on computational intelligence and software engineering, Wuhan, China,
pp 1–4.

[6] Ihsan Gunes,Cihan Kaleli,Alper Bilge,Huseyin Polat, Shilling attacks against recommender systems: a comprehensive survey , Artificial Intelligence Review, December 2014, Volume 42, Issue 4,pp 767-799.

[7] Lam, S. And Riedl, J. Shilling recommender systems for fun and profit. In Proceedings of the 13th International WWW Conference (New York, NY)(2004).

[8] Burke, R.,Mobasher, B.,Williams, C., And Bhaumik, R. 2006b. Detecting profile injection attacks in collaborative recommender systems. In Proceedings of the IEEE Joint Conference on Ecommerce Technology and Enterprise Computing, E-Commerce and E-Services (CEC/EEE 2006, Palo Alto, CA)(2006).

[9] Chirita PA, NejdlW, Zamfir C (2005) Preventing shilling attacks in online recommender systems. In: Proceedings of the 7th annual ACM international workshop on web information and data management, Bremen, Germany, pp 67–74.

[10] Burke RD, Mobasher B,Williams CA, Bhaumik R (2006a) Classification features for attack detection in collaborative recommender systems. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, Philadelphia, PA, USA, pp 542–547.

[11] Williams CA, Mobasher B, Burke RD (2007a) Defending recommender systems: detection of profile injection attacks. Serv Oriented Comput Appl 1(3):157–170.

[12] Bhaumik R, Mobasher B, Burke RD (2011) 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 Breese JS, Heckerman D, Kadie K (1998).

[13] Zhang S, Chakrabarti A, Ford J, Makedon F. Attack detection in time series for recommender systems. In: KDD ’06:Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 809–814(2006).

[14] R. Bhaumik, C. Williams, B. Mobasher, and R. Burke. Securing collaborative filtering against malicious attacks through anomaly detection. In Proceedings of the 4th Workshop on Intelligent Techniques for Web Personalization (ITWP’06), at AAAI’06, Boston, (2006).

[15] P. S. Chakraborty, S. Karforma, A Time Series based Technique for Detecting shilling attacks in Recommender Systems, Proceedings of International Conference On Computing and Systems-2013 (ICCS-2013), Dept of Comp. Sc. ,The University of Burdwan.

[16] P. S. Chakraborty, S. Karforma, Detection of Profile-injection attacks in Recommender Systems using Outlier Analysis,International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013, Dept of Comp. Sc.,The University of Kalyani. Published in Procedia Technology, Elsevier.

[17] Mehta, B.: Unsupervised shilling detection for collaborative filtering. Association for the Advancement of Artificial Intelligence (2007). www.aai.org

[18] Loureiro,A., L. Torgo and C. Soares, 2004. Outlier Detection using Clustering Methods: a Data Cleaning Application, in Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany.

[19] Al- Zoubi, M. B., An Effective Clustering-Based Approach for Outlier Detection, European Journal of Scientific Research, Vol. 28, No. 2, 2009, pp. 310-316.

[20] P. S. Chakraborty, S. Karforma, Effectiveness of Proximity-based Outlier Analysis in Detecting Profile-injection attacks in E-Commerce Recommender Systems, second International Conference in Information Systems Design and Intelligent Applications (INDIA-2015), Dept of Comp. Sc.,The University of Kalyani.

[21] He, Z., Xu, X., Deng, S.: Discovering cluster-based local outliers. Pattern Recogn. Lett. 24, 1641–1650 (2003)

[22] Knorr, E., Ng, R.: Algorithms for mining distance-based outliers in large datasets. In: VLDB Conference (1998).

[23] R. Bhaumik, C. Williams, B. Mobasher, and R. Burke. Securing collaborative filtering against malicious attacks through anomaly detection. In Proceedings of the 4th Workshop on Intelligent Techniques for Web Personalization (ITWP’06), at AAAI’06, Boston, (2006).

[24] Shewart, W. A. 1931. Economic Control of Quality of manufactured Product. Van Nostrand.

[25] S. Zhang, J. Ford, and F. Makedon. Analysis of a low-dimensional linear model under recommendation attacks. The 29th Annual International ACM Conference on Research & Development on Information Retrieval (SIGIR 2006), Seattle, WA, August 6-11,(2006).

[26] P. S. Chakraborty, S. Karforma, Detection of profile-injection attacks in Recommender Systems using SSA-based Change-Point Analysis, Mathematical Sciences International Research Journal, Vol 1, No. 3, ISSN No:2278-8697, ISBN NO: 978- 93-81583-57-9.

[27] Moskvina, V. and Zhigljavsky A.A. (2003) An Algorithm Based on Singular-Spectrum Analysis for Change-Point Detection, Communica tion in Statistics. Statistics and Simulations, 32, 319-352.

[28] O’MahonyMP, Hurley NJ, Silvestre GCM (2003) Collaborative filtering-safe and sound. Lect Notes ComputSci 2871:506–510.

[29] Mehta B (2007) Unsupervised shilling detection for collaborative filtering. In: Proceedings of the 22nd international conference on artificial intelligence, Vancouver, BC, Canada, pp 1402–1407.

[30] Mehta B, Nejdl W (2009) Unsupervised strategies for shilling detection and robust collaborative filtering. User Model User Adapt Interact 19(1-2):65–97.

Author’s Profiles

Partha Sarathi Chakraborty has done his M..E.(in I.T.) degree from West Bengal University of Technology and M.C.A from I.G.N.O.U.. Before that he obtained Master degree in Economics from Kalyani University. He is currently serving as Assistant professor of I.T. and C.S.E. department of University Institute of Technology, The University of Burdwan.

Dr. Sunil Karforma has completed B.E. (Computer Science and Engineering) and M. E. (Computer Science and Engineering) from Jadavpur University. He has completed his Ph. D. in the field of Cryptography. He is presently holding the post of Associate Professor and the Head of the Department in the Department of Computer Science, The University of Burdwan.