UDC 004.9 **doi: 10.20998/2522-9052.2020.2.05**

Ye. Meleshko, O. Drieiev, H. Drieieva

Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine

METHOD OF IDENTIFICATION BOT PROFILES BASED ON NEURAL NETWORKS IN RECOMMENDATION SYSTEMS

Abstract. The subject matter of the article is the process of increased the information security of recommendation systems. The goal of this work is to develop a method of identification bot profiles in recommendation systems. In this work, the basic models of information attacks by the profile-injection method on recommendation systems were researched, the method of identification bot profiles in recommendation systems using the multilayer feedforward neural network was developed and the experiments to test the quality of its work were conducted. The developed method is to identify bot profiles that attempt to change item ratings in a recommendation system in order to increase the occurrence frequency of target items in recommendation lists to all authentic users, or to certain segments of authentic users. When removing bot profiles' data from the database of the recommendation system before generating recommendation lists, the accuracy of the system and the correctness of recommendations are significantly increased, and authentic users get protection from information attacks. Random, Average and Popular attacks were used to model the attacks on a recommendation system. To identify bots, their ratings for system items were analyzed. The experiments have shown that the neural network that analyzes only the numbers of different ratings in a profile, detects bot profiles with high accuracy, that use Random attack regardless of the number of target items for each bot. At the same time, the developed neural network can detect bots that use Average or Popular attacks only when they have several target items. Also, the results of the experiments show that type I errors, when the system identifies authentic users as bots, is very rarely appear in the developed method. To improve the accuracy of the neural network, there can add to analysis also other data of user profiles, such as the timestamp of each rating and as segments of items, which was rated.

Keywords: recommendation systems; information attacks; information security; Internet bots; neural networks; data clustering.

Introduction

Recommendation systems that use feedback from users, such as ratings, views, comments, etc., to create recommendations are vulnerable to information attacks aimed at changing ratings of certain items [1, 2].

The main type of information attacks on recommendation systems is profile-injection attacks [1-9], which are to create a network of bots to perform concerted actions to change the ratings of target items.

In this work, a bot should be understood as a program that automatically performs actions that mimic user activity of a recommendation system and performs target actions in parallel, in order to change ratings of target objects and/or to collect statistics about authentic users.

Bots should very accurately simulate actions of authentic users of a recommendation system, otherwise, they will be detected by the system, and their actions will be neutralized. However, they will never have all statistic data of authentic users without access to a website at the administrator level, so they will not be able to perfectly simulate the actions of authentic users. Therefore, creators of such bots should seek the compromise between the obscurity of bots and the information amount that must be collected for an attack. Attackers can collect information to attack in two ways:

1. Parsing HTML-text of open data on a website to get user activity statistics (such as global average rating, the occurrence frequency of different ratings, average ratings of target items, lists of popular items in target user segments, etc.).

2. Performing Probe attack [1], which consists of creating bot profiles with characteristics of users from target segments and collecting statistics about preferences of such users based on recommendation lists generated by a system for these bot profiles. These methods can be used individually or together.

At the same time, a robust recommendation system should work so that attackers' actions are so ineffective that their results will not give incentives to continue attacks and authentic users will continue to receive relevant undistorted recommendations.

To protect recommendation systems from profile injection attacks, can use the following steps:

1. Models of possible information attacks on a specific recommendation system are created.

2. Methods of identification bot profiles on the basis of created information attack models are being developed. Typically, these methods are based on data classification and clustering methods and allow divide all profiles into two groups: authentic users and bots.

3. The detected bot profiles are not taken into account in the formation of recommendation lists, their data (ratings, actions, etc.) are removed from the database of a recommendation system.

The purpose of this work is to develop a method of identifying bot profiles in recommendation systems to prevent information attacks of item ratings' change. The developed method is to identify the bot profiles that try to change the ratings of items in the system so that to change the appearing probability of target items in recommendation lists.

Removing detected bot profiles from a recommendation system database before calculating recommendations will greatly increase its robustness to information attacks.

1. Models of information attacks on recommendation systems

An information attack on a recommendation system based on the profile injection method - is a concerted effort by a bot network to put certain ratings to some items in order to change their frequency of appearance in recommendation lists for all or a specific segment of users.

Attack on recommendation systems usually consists of two steps:

1. Creating profiles of bots. Attacks that require fewer bot profiles will be more simpler to attackers.

2. Filling in the bot profiles with ratings. To do this, attackers need to collect some statistics from a system. The more knowledge attackers have about the distribution of ratings in a system, the more realistic bot profiles they will create.

The bot profile model is shown in Fig. 1.

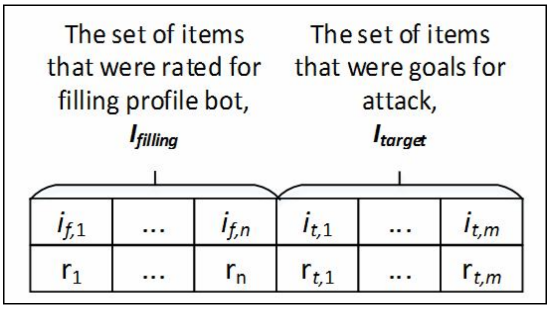
The set of items that were rated for filling profile bot, $I_{filling}$
The set of items that were goals for attack, I_{target}
$i_{f,1}$
…
$i_{f,n}$
$i_{t,1}$
…
$i_{t,m}$
r_1
…
r_n
$r_{t,1}$
…
$r_{t,m}$

Fig. 1. Model of bot profile

As can be seen from the figure, a bot profile contains the following types of ratings:

– ratings of the set $I_{filling}$ to simulate the actions of real users, attackers try to choose values of these ratings as closely as possible to true ratings for the target user group, which they want to influence;

– ratings of the set I_{target} for target items, this is usually maximum ratings (or close to maximum) if an attacker aims to raise the rating of target items, or minimum ratings (or close to minimum) if the goal is to lower the rating of target items.

In this work, we have modeled attacks on object rating increase.

The most common models of attacks on recommendation systems are the following – Random attack,

Average attack and Popular attack [1-9]. Let's look at these models of attacks in more detail.

Random attack. In bot profiles, the set $I_{filling}$ will be filled with random ratings for items selected at random. Ratings are generated close to the global average in a system. The target item will be rated a maximum rating r_{max}. To perform such an attack, one only needs to know the global average rating in a recommendation system.

Average attack. The set $I_{filling}$ filled with random items that get ratings close to their individual average values of ratings in a system. For this attack, an attacker needs to collect data about the average values of all or some ratings in a system. This attack is more inconspicuous than Random attack, bot profiles will be very similar to a large number of user authentic profiles.

Popular attack. The set $I_{filling}$ filled with popular items that get ratings that are equal to their average ratings in a system. Such strategy will lead to positive correlations between bot profiles and authentic profiles. Therefore, bot profiles will be extremely difficult to detect.

These attack models were used to model the data for bot profiles in this work.

2. The method of identification bot profiles in recommendation systems

To identify bot profiles, the multilayer feedforward neural network was developed (Fig. 2).

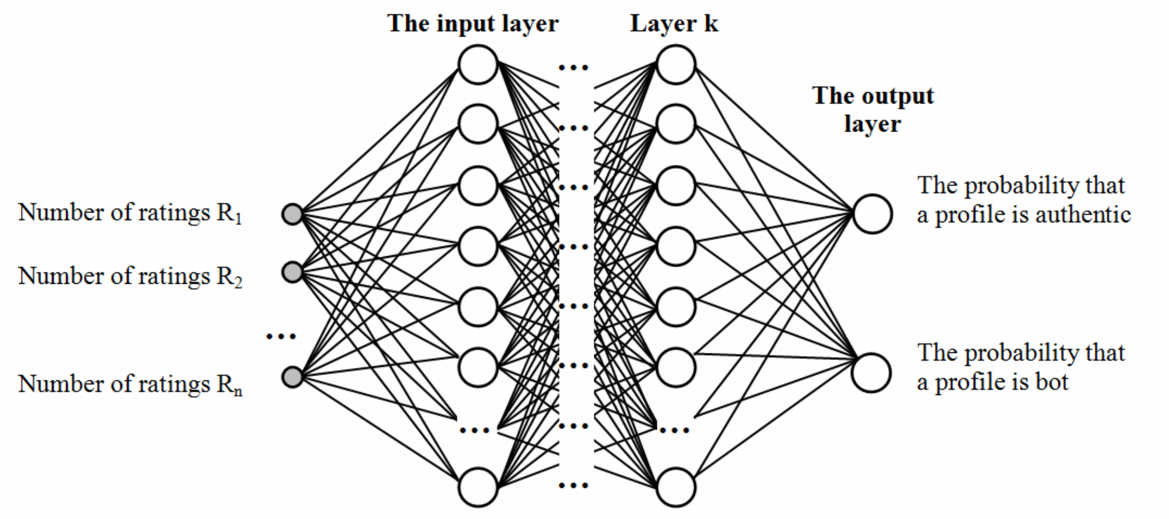

Fig. 2. General architecture of the artificial neural network for identifying bots

As the input data for the artificial neural network were selected the number of each different ratings in a user profile. At the output, the neural network gives probabilities that the profile belongs to an authentic user or bot. For the experiments, the data of authentic users was taken from MovieLens Datasets [10], and the data of bots was generated based on discussed above models of information attacks on recommendation systems.

The TensorFlow library [11] and the Python programming language were used to implement the neural network.

It was experimentally found that the balance between accuracy and complexity of the neural network can be achieved with the following parameters, which are shown in table 1.

Table 1 – **Parameters of the developed neural network**

Layer	**Type**	**Number of neurons**	**Number of inputs on each neuron**	**Activation function**
1	Input	100	10	Sigmoid
2	Hidden	100	100	Sigmoid
3	Hidden	100	100	Sigmoid
4	Output	2	100	Softmax

Because in MovieLens Datasets there are ten variants of different ratings from 0.5 to 5, then there should be 10 inputs in the neural network for this case. It was decided to create 2 hidden layers in the neural network.

The sigmoid activation function was used in the input layer and hidden layers. The output layer consists of 2 neurons with softmax activation function and showing the probabilities that the profile is authentic or bot.

The Adam algorithm [12] was used to train the neural network, it is a modification of the stochastic gradient descent algorithm, which has been widely used in systems with deep learning.

Adam algorithm differs from the classic gradient descent in that it calculates individual adaptive speeds of descent for different neural network weights.

3. The experiment to test the efficiency of the developed method

The series of experiments were conducted, the results of which are given in Table 2. The designations used in Table 2:

RA – Random attack,
AA – Average attack,
PA – Popular attack.

For each considered attack model, one training data set and eight testing data set were created. Each data sets contained 3000 users. In training data sets, half of profiles were bots. In testing data set, bots were between 5% and 30% of profiles.

Also, data sets for testing with different numbers of targets for bots – with 1 and 10 target items were created.

The identical neural networks, one for each considered attack model, were created, each of which was trained on the corresponding training data set and tested on the corresponding testing data set.

In the table 2 shows the results of the experiments.

Table 2 – **Test results of the developed method**

Experiment number	**Number of bots, %**	**The number of goals for each bot, grand**	**Type I errors, %**			**Type II errors, %**			**Precision, %**			**Recall, %**			**F-score, %**		
			RA	**AA**	**PA**	**RA**	**AA**	**PA**	**RA**	**AA**	**PA**	**RA**	**AA**	**PA**	**RA**	**AA**	**PA**
1	5	1	0.001	0.002	0.002	0.023	0.049	0.047	0.95	0.12	0.25	0.52	0.006	0.006	0.67	0.013	0.013
2	10	1	0.0006	0.001	0.0006	0.036	0.099	0.099	0.98	0.25	0.33	0.63	0.003	0.003	0.77	0.006	0.006
3	20	1	0.002	0.001	0.002	0.080	0.199	0.197	0.98	0.28	0.14	0.59	0.003	0.001	0.74	0.006	0.003
4	30	1	0.0006	0.002	0.002	0.116	0.297	0.296	0.99	0.25	0.30	0.61	0.002	0.003	0.75	0.004	0.006
Mean values:			**0.001**	**0.001**	**0.001**	**0.063**	**0.161**	**0.159**	**0.97**	**0.22**	**0.25**	**0.58**	**0.003**	**0.003**	**0.73**	**0.007**	**0.007**
5	5	10	0.002	0.001	0.001	0.001	0.024	0.009	0.96	0.93	0.97	0.98	0.50	0.81	0.97	0.65	0.88
6	10	10	0.001	0.001	0.002	0.002	0.045	0.014	0.98	0.98	0.96	0.97	0.55	0.85	0.98	0.70	0.90
7	20	10	0.001	0.001	0.001	0.005	0.047	0.013	0.99	0.99	0.99	0.97	0.76	0.93	0.98	0.86	0.96
8	30	10	0.001	0.001	0.002	0.005	0.024	0.006	0.99	0.99	0.99	0.98	0.92	0.97	0.98	0.95	0.98
Mean values:			**0.001**	**0.001**	**0.001**	**0.003**	**0.035**	**0.010**	**0.98**	**0.97**	**0.97**	**0.97**	**0.68**	**0.89**	**0.97**	**0.79**	**0.93**

The following metrics to evaluate the quality of the neural network work were selected:

1. **Type I errors** – "false alarm" when an authentic user is identified as a bot.

2. **Type II errors** – "missed goal", when a bot is identified as an authentic user.

3. **Precision** – the measure that characterizes how many positive predictions of the neural network are correct. It was calculated by the formula:

$$Precision = \frac{tp}{tp + fp}, \qquad (1)$$

where *tp* – positive predictions from the neural network, which turned out to be correct;

fp – positive predictions of the neural network, which turned out to be wrong.

4. **Recall** (also known as Sensitivity) – the measure that characterizes the ability of the neural network to generate correct positive predictions. It was calculated by the formula:

$$recall = \frac{tp}{tp + fn}, \qquad (2)$$

where *tp* – positive predictions from the neural network, which turned out to be correct;

fn – negative predictions from the neural network, which turned out to be incorrect.

5. **F-score** – is the harmonic mean of the precision and recall. It was calculated by the formula:

$$F = 2 \cdot \frac{tprecision \cdot recall}{tprecision + recall}. \qquad (3)$$

As the experiments showed, the developed neural network can quite accurately identify bots, which use Random attack, regardless of the number of their targets, with an average precision of 0.97.

But it does have significant problems identify bots that use Average and Popular attacks.

The neural network can only identify such bots if they have several goals.

Also, the results of the experiments show that type I errors (when the system identifies authentic users as bots) is very rarely appear in the developed method.

Conclusions

The method of identification bot profiles in recommendation systems based on the multilayer feedforward neural network was developed. The experiments have been testing the efficiency of the developed method.

The results of the experiment showed that the easiest task is to identify bots which implement Random attack, they can be detected even if bots have one target item. It is much more difficult task to detect Average and Popular attacks.

With such types of attacks, the developed method can only detect bots that have several target items.

To improve the accuracy of the neural network, there can add to analysis also other data of user profiles, such as the timestamp of each rating and as segments (clusters) of items, which was rated.

References

1. (2010), *Recommender Systems Handbook*, Editors Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor, 1st edition, Springer-Verlag New York Inc., New York, USA, 842 p.
2. Lam, S.K.,and Riedl, J. (2004), “Shilling recommender systems for fun and profit”, *Proceedings of the 13th International World Wide Web Conference*, pp. 393–402.
3. O’Mahony, M.P., Hurley, N.J. and Silvestre, G.C.M. (2002), “Promoting recommendations: An attack on collaborative filtering”, *Database and Expert Systems Applications: 13th International Conference, DEXA Aix-en-Provence*, France, pp. 494-503.
4. Williams, C., Mobasher, B. and Burke, R. (2007), “Defending recommender systems: detection of profile injection attacks”, *Service Oriented Computing and Applications*, pp. 157–170.
5. Chirita, P.A., Nejdl, W. and Zamfir C. (2005), “Preventing shilling attacks in online recommender systems”, *Proceedings of the ACM Workshop on Web Information and Data Management*, pp. 67–74.
6. Zhou W., Wen J., Qu Q., Zeng J. and Cheng T. (2018), “Shilling attack detection for recommender systems based on credibility of group users and rating time series”, PLoS ONE 13(5): e0196533, DOI: https://doi.org/10.1371/journal.pone.0196533
7. Kumari, T. and Punam, B.A (2017), “Comprehensive Study of Shilling Attacks in Recommender Systems”, *IJCSI International Journal of Computer Science Issues*, Volume 14, Issue 4, URL: https://www.ijcsi.org/papers/IJCSI-14-4-44-50.pdf
8. Mobasher, B., Burke, R., Bhaumik, R. and Williams C. (2007), “Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness”, *ACM Transactions on Internet Technology*, Vol. 7(4), pp. 1–41.
9. Mobasher, B., Burke, R., Bhaumik R. and Williams C. (2005). “Effective attack models for shilling item-based collaborative filtering system”, *Proceedings of the WebKDD Workshop*, pp. 1–8.
10. Harper, F.M. and Konstan J.A. (2016), “The MovieLens Datasets: History and Context”, *ACM Transactions on Interactive Intelligent Systems (TiiS)*, available at: https://doi.org/10.1145/2827872
11. (2020), *TensorFlow tutorials*, URL: https://www.tensorflow.org/tutorials/
12. (2017), “The gentle introduction to Adam optimization algorithm for deep learning”, *Blog about Machine Learning, Neural Networks*, Artificial Intelligence, URL: https://www.machinelearningmastery.ru/adam-optimization-algorithm-for-deep-learning (in Russian).

Надійшла (received) 21.02.2020
Прийнята до друку (accepted for publication) 29.04.2020

Відомості про авторів / About the Authors

Мелешко Єлизавета Владиславівна – кандидат технічних наук, доцент, докторант кафедри кібербезпеки та програмного забезпечення, Центральноукраїнський національний технічний університет, Кропивницький, Україна;

Yelyzaveta Meleshko – Candidate of Technical Sciences, Associate Professor, Doctoral Student of Cybersecurity and Software Department, Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine;
e-mail: elismeleshko@gmail.com; ORCID ID: https://orcid.org/0000-0001-8791-0063.

Дрєєв Олександр Миколайович - кандидат технічних наук, доцент кафедри кібербезпеки та програмного забезпечення, Центральноукраїнський національний технічний університет, Кропивницький, Україна;
Oleksandr Drieiev – Candidate of Technical Sciences, Associate Professor of Cybersecurity and Software Department, Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine;
e-mail: drey.sanya@gmail.com; ORCID ID: https://orcid.org/0000-0001-6951-2002.

Дрєєва Ганна Миколаївна - аспірант кафедри кібербезпеки та програмного забезпечення, Центральноукраїнський національний технічний університет, Кропивницький, Україна;
Hanna Drieieva – Graduate student of Cybersecurity and Software Department, Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine;
e-mail: gannadreeva@gmail.com; ORCID ID: https://orcid.org/0000-0002-8557-3443.

Метод ідентифікації профілів ботів на основі нейронних мереж у рекомендаційних системах

Є. В. Мелешко, О. М. Дрєєв, Г. М. Дрєєва

Анотація. Об'єктом дослідження цієї роботи є процес підвищення інформаційної безпеки рекомендаційних систем. Метою роботи є розробка методу ідентифікації акаунтів ботів у рекомендаційній системі. У цій роботі було розглянуто основні моделі інформаційних атак ін'єкцією профілів на рекомендаційні системи, розроблено метод ідентифікації профілів ботів у рекомендаційних системах за допомогою багатошарової нейронної мережі прямого поширення та проведені експерименти для перевірки якості його роботи. Розроблений метод полягає у виявлені профілів ботів, які намагаються змінити рейтинги об'єктів у рекомендаційній системі з метою підвищення потрапляння цільових об'єктів до списків рекомендацій усім автентичним користувачам, або певним сегментам автентичних користувачів. Вилучення виявлених профілів ботів з бази даних рекомендаційної системи перед обчисленням рекомендацій значно підвищує точність її роботи та достовірність рекомендацій, а також захищає користувачів системи від інформаційних атак. Для моделювання атаки на рекомендаційну систему було використано випадкову, середню та популярну атаки. Для розпізнавання ботів аналізувалися оцінки, які вони виставляли об'єктам системи. Як показали проведені експерименти, нейронна мережа, що аналізує лише кількість різних оцінок у профілі, з високою точністю виявляє профілі ботів, які використовують випадкову атаку незалежно від кількості цільових об'єктів у кожного боту. В той же час розроблена нейронна мережа може виявляти ботів, що використовують середню та популярну атаку, тільки тоді, коли вони мають декілька цілей. Також з результатів експериментів видно, що у розробленому методі досить рідко виникають помилки першого роду – коли система ідентифікує звичайних користувачів як ботів. Для підвищення точності роботи нейронної мережі, можна враховувати й інші параметри профілів користувачів, зокрема, час виставлення кожної оцінки, а також те, до яких сегментів відносяться оцінені у профілі об'єкти.

Ключові слова: рекомендаційні системи; інформаційні атаки; інформаційна безпека; Інтернет-боти; нейронні мережі; кластеризація даних.

Метод идентификации профилей ботов на основе нейронных сетей в рекомендательных системах

Е. В. Мелешко, А. Н. Дреев, А. Н. Дреева

Аннотация. Объектом исследования данной работы является процесс повышения информационной безопасности рекомендательных систем. Целью работы является разработка метода идентификации аккаунтов ботов в рекомендательной системе. В данной работе были рассмотрены основные модели информационных атак инъекцией профилей на рекомендательные системы, разработан метод идентификации профилей ботов в рекомендательных системах с помощью многослойной нейронной сети прямого распространения и проведены эксперименты для проверки качества его работы. Разработанный метод заключается в выявлении профилей ботов, которые пытаются изменить рейтинги объектов в рекомендательной системе с целью повышения попадания целевых объектов в списки рекомендаций всем аутентичным пользователям или определенным сегментам аутентичных пользователей. Изъятие выявленных профилей ботов из базы данных рекомендательной системы перед вычислением рекомендаций значительно повышает точность ее работы и достоверность рекомендаций, а также защищает пользователей системы от информационных атак. Для моделирования атаки на рекомендательную систему было использовано случайную, среднюю и популярную атаки. Для распознавания ботов анализировались оценки, которые они выставляли объектам системы. Как показали проведенные эксперименты, нейронная сеть, которая анализирует только количество различных оценок в профиле пользователя, с высокой точностью выявляет профили ботов, которые используют случайную атаку независимо от количества целевых объектов у каждого бота. В то же время разработана нейронная сеть может обнаруживать ботов, использующих среднюю и популярную атаку, только тогда, когда они имеют несколько целей. Также по результатам экспериментов видно, что в разработанном методе очень редко возникают ошибки первого рода – когда система идентифицирует обычных пользователей как ботов. Для повышения точности работы нейронной сети, можно учитывать и другие параметры профилей пользователей, в частности, время выставлении каждой оценки, а также то, к каким сегментам относятся оцененные в профиле объекты.

Ключевые слова: рекомендательные системы; информационные атаки; информационная безопасность; Интернет-боты; нейронные сети; кластеризация данных.