DOI: https://doi.org/10.33216/2222-3428-2022-23-4
УДК 004.9

Покришка С.А., Шумова Л.О.

ЗАБЕЗПЕЧЕННЯ СТІЙКОСТІ РЕКОМЕНДАЦІЙНИХ СИСТЕМ ДО ШИЛІНГОВИХ АТАК

У статті розглянуто проблеми підвищення точності пропозицій рекомендаційних систем користувачам контент-орієнтованих веб-ресурсів в умовах шилінгових атак. Проведений аналіз зовнішніх факторів, що можуть дестабілізувати роботу рекомендаційних систем, показав уразливість рекомендаційних систем до загроз інформаційних атак. Підвищення стійкості рекомендаційних систем до дії негативних факторів дозволить підвищити точність та інші показники якості роботи. Основним зовнішнім дестабілізуючим фактором у рекомендаційних системах є інформаційні атаки ін'єкцією – шилінгові атаки. Шилінгові зловмисники мотивовані різними цілями, що зумовлює розробку різних моделей шилінгових атак, які розрізняються в основному рівнем знань про об'єкти рекомендаційної системи та ступенем впливу на неї. Розглянуті мотивації та наслідки шилінгу. Зловмисники маніпулюють частотою рекомендацій цільових елементів фальсифікуючи профілі користувачів. Щоб вплинути на список рекомендацій рекомендаційних систем, зловмисники шилінгу впроваджують підроблені профілі користувача контенту. Деякі атаки можуть намагатися "штовхати" цільові об'єкти (push-атаки), інші можуть бути спрямовані на "підрив" деяких цільових об'єктів (nuke-атаки). Наведено класифікацію методів і моделей шилінгу та здійснено порівняльний аналіз їх негативного впливу на точність пропозицій рекомендаційних систем. Забезпечення стійкості рекомендаційних систем до шилінгових атак є важливою умовою для підвищення точності їх роботи. Досліджено методи виявлення інформаційних атак на рекомендаційні системи. Для аналізу, тренування та тестування було обрано найбільш впливові методи шилінгових атак, а саме: random, average і bandwagon атаки. Найбільшу точність пропозицій користувачам за умов загроз обраних моделей шилінгових атак забезпечив метод sgd_classifier.

Ключові слова: *рекомендаційні системи, інформаційна безпека, шилінгові атаки, веб-ресурси, машинне навчання.*

Актуальність дослідження. В інтернеті величезний об'єм даних, які постійно створюються, що ускладнює користувачам контент-орієнтованих веб-ресурсів пошук необхідної для них інформації. Рекомендаційна система - це особливий тип системи фільтрації інформації, адаптований онлайн-продавцями для надання рекомендацій своїм клієнтам на основі їхніх вимог [1, 2]. Однією з найбільш широко використовуваних рекомендаційних систем є спільна фільтрація, але практика доводить, що вона схильна до загроз зовнішніх дестабілізуючих факторів у комп'ютерних мережах таких як атаки шилінгу [3]. Такі атаки змінюють процес рекомендацій, щоб просувати чи знижувати певний продукт. Протягом багатьох років розроблено безліч моделей атак і, як протидія, досліджено і впроваджено засоби їх виявлення [4, 5]. Але процес протистояння триває. Таким чином, стрімке збільшення у комп'ютерних мережах кількості контент-орієнтованих веб-ресурсів, які використовують рекомендаційні системи, та їх уразливість до загроз шилінгових атак зумовлюють актуальність завдань підвищення стійкості рекомендаційних систем веб-сервісів до шилінгових атак.

Метою статті є дослідження методів вирішення проблеми підвищення точності пропозицій рекомендаційних систем в умовах інформаційних загроз на основі розробки та впровадження моделей виявлення шилінгових атак. Для виконання поставленої мети поставлені наступні завдання.

1. Аналіз методів і моделей шилінгу та порівняння ступеня їх негативного впливу на точність пропозицій рекомендаційних систем.
2. Дослідження методів та алгоритмів виявлення атак шилінгу в рекомендаційних системах.
3. Розробка моделей виявлення шилінгових аккаунтів і атак засобами машинного навчання.

Методи і моделі шилінгу

Щоб вплинути на список рекомендацій рекомендаційних систем, зловмисники шилінгу впроваджують підроблені профілі користувача контенту. Деякі атаки можуть намагатися "штовхати" цільові об'єкти (push-атаки), інші можуть бути спрямовані на "підрив" деяких цільових об'єктів (nuke-атаки). Зловмисники маніпулюють частотою рекомендацій цільових елементів фальсифікуючи профілі користувачів. На основі мотивації та знань зловмисників за минулі роки було розроблено безліч моделей атак [6]. Всі ці атаки можуть бути класифіковані як атаки з високим рівнем знань про об'єкти рекомендаційної системи та атаки з низьким рівнем знань (рис. 1). Атаки з низьким рівнем знань (в основному це стандартні атаки) є більш практичними але мають менше шансів надати реальний вплив, і ефективність таких атак також низька. З іншого боку, атаки з високим рівнем знань (представники групи прихованих атак, obfuscated attacks) можуть вплинути на продуктивність рекомендаційних систем, але їх складніше здійснити.

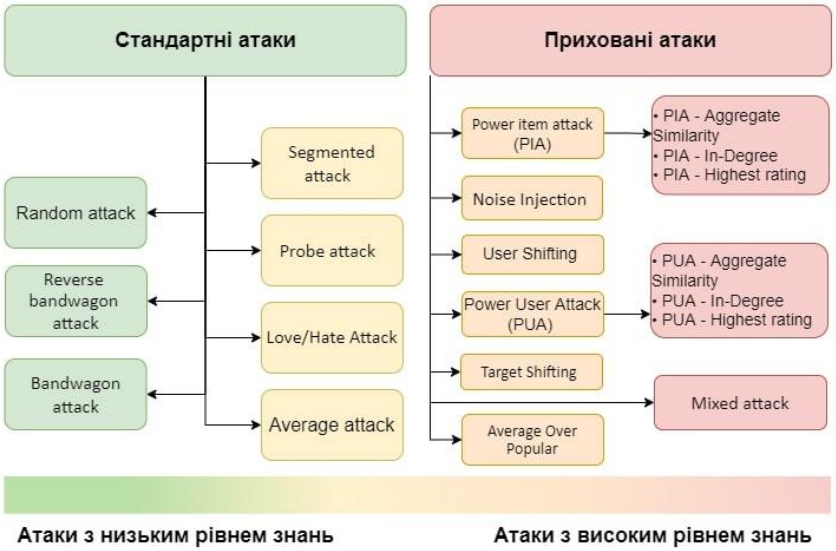

Рисунок 1 - Класифікація методів і моделей шилінгових атак та їх оцінка за рівнем знань про об'єкти рекомендаційної системи

Стандартні атаки (standart attacks) не роблять виняткових спроб залишитися непоміченими у рекомендаційній системі. Багато алгоритмів виявлення загроз мають вищі шанси виявити профілі атак шилінгу, впроваджені з допомогою цих атак. Розроблено наступні моделі стандартних атак: random attack, average attack, bandwagon attack, segmented attack, probe attack, Love/Hate Attack, reverse bandwagon attack (рис. 1).

Random attack є найпростішою формою атаки шилінгу. У цій моделі елементи, оцінені профілем атаки, вибираються випадковим чином, крім цільового елемента. Цільовий елемент отримує максимальний або мінімальний рейтинг залежно від того, чи він є атакою "push" або "nuke". Деякі атаки спрямовані на порушення надійності рекомендаційної системи - відомі як випадковий вандалізм. Метою випадкової атаки зазвичай є скоріше порушення роботи рекомендаційної системи, ніж просування цільового елемента. Будучи найпростішою атакою, вона дуже ефективна.

Average attack схожа на випадкову атаку щодо процесу випадкового вибору елементів. Ця атака здійсненна тільки в тому випадку, якщо атакуючий має суттєві знання про набір даних, на якому побудована рекомендаційна система. Ефективність цієї моделі пропорційна знанням атакуючого. Хоча різниця між випадковою атакою та середньою атакою полягає лише в рейтингах наповнювачів, ефективність середньої атаки набагато вища.

Bandwagon attack - це тип атаки, коли профілі, створені зловмисниками, заповнюються популярними елементами з високим рейтингом. Цільовому об'єкту надається найвищий рейтинг. Цю атаку можна додатково розділити на bandwagon-random та bandwagon-average залежно від схеми рейтингу, що використовується для заповнення елементів. Bandwagon належить до категорії атак з низьким рівнем знань, оскільки атакуючі потребують лише загальнодоступних даних.

Segmented attack спрямована на певну групу користувачів, які швидше за все придбають цільовий товар в електронній комерції. Сегментні атаки зазвичай застосовуються при колективній фільтрації на основі елементів. Товари та рейтинги, що оцінюються, ґрунтуються на знаннях атакуючого про сегмент. Істотна перевага цього перед іншими полягає у можливості охопити потенційних покупців. Оскільки атака розгорнута лише у сегменті системи, її вплив великий.

Probe attack - це та атака, яку можна направити на будь-яку систему. Деякі рекомендаційні системи прогнозують рейтинговий бал для кожного предмета. Зловмисник використовує цю інформацію для оцінки предметів, що дозволяє йому бути схожим на інших користувачів. Зловмисник виставляє справжні оцінки деяким початковим предметам. Потім, коли рекомендатор пропонує інші елементи, зловмисник формує список елементів, що оцінюються на основі цих елементів. Ця схема гарантує, що профілі атаки залишаються близькими до своїх сусідів. Вона також дозволяє зловмиснику більше дізнатися про систему.

В моделі Love/Hate Attack атакуючий випадковим чином вибирає елементи-заповнювачі та надає їм найвищий та найменший рейтинг. Незважаючи на простоту цієї моделі, її ефективність досить висока.

Reverse bandwagon attack є точною зворотною стороною атаки bandwagon. Ця атака використовується для знищення цільового продукту шляхом присвоєння низьких оцінок товарам з великою кількістю негативних відгуків та надання найменшої оцінки цільовому товару. Це також атака з низьким рівнем знань, як і атака bandwagon. Хоча вона дуже схожа на атаку bandwagon, ефективність зворотної атаки bandwagon трохи вища.

Приховані атаки (obfuscated attacks) націлені на те, щоб залишитись непоміченими алгоритмами виявлення. Для цього зловмисники намагаються замаскувати сигнатури своїх атак. Багато моделей включають невеликі модифікації стандартних методів атак досягнення обфускации. Хоча обфускація може зменшити вплив атаки, це краще, ніж бути виявленим. Можна виділити такі моделі прихованих атак: Power item attack, Noise Injection, User Shifting, Power User Attack, Target Shifting, Average Over Popular, Mixed attack (рис. 1).

Power item attack (PIA) використовує потужні елементи, що вибираються на основі трьох методів. Потужні елементи визначаються як набір елементів, які можуть впливати найбільшу групу елементів. Ці елементи ефективно змінюють рекомендації іншим користувачам. У PIA-AS як сильні елементи вибираються N елементів з найбільшою сумарною подібністю. Така подібність можлива лише в тому випадку, якщо значну кількість користувачів оцінили два однакові предмети. У PIA-ID критерієм вибору елементів влади є центральність In-Degree. Подібність кожної пари елементів розраховується за допомогою виваженої значущості та вибирається топ-N кожного елемента. PIA-NR вибирає елементи з найбільшим числом користувачів як потужні елементи.

Noise Injection додає до кожного рейтингу підмножини профілів, що зазнали ін'єкції, випадкове число з розподілом Гауса, помножене на константу. Ступінь обфускації залежить від значення цієї константи. Цей метод може бути ефективно застосований до всіх стандартних методів атак для обфускування їхньої сигнатури.

User Shifting - це тактика обфускації, коли змінюється підмножина елементів рейтингу кожного інжектованого профілю.

Power User Attack, подібно до PIA, вибирає набір користувачів, що мають максимальний вплив на найширшу групу користувачів. У PUA-ID як впливові користувачі вибираються користувачі, які беруть участь у найбільшій кількості сусідств, на основі концепції центральності In-Degree. PUA-NR сильними користувачами є користувачі із найбільшою кількістю оцінок у профілі.

Target Shifting зміщує рейтинг цільового елемента на один рівень нижче, ніж максимально можливий під час push-атаків. При nuke-атаках рейтинг мети зміщується на один рейтинг вищим, ніж мінімально можливий. Ця стратегія є особливо корисною для ухилення від методів виявлення, які карають користувачів, які дають екстремальні оцінки предметам. Якщо цільовий елемент вже популярний, його складніше просунути, застосовуючи обфускацію зі зсувом цілей. У разі слід використовувати інші методи обфускации.

Average Over Popular - техніка, що використовується для маскування атак усереднення. Тут елементи-заповнювачі вибираються з X% найпопулярніших елементів з ймовірністю. Цей метод набагато ефективніший, ніж випадковий вибір із усієї колекції елементів. Вибір X впливає на виявлення атаки.

Mixed attack здійснюється шляхом одночасного використання випадкової, середньої, смугової та сегментованої атак у рівних пропорціях. Різні методи атаки використовуються для на той же цільовий елемент. Це допомагає уникнути кількох методів виявлення.

Знання методів і моделей шилінгових атак дозволяють проводити успішні дослідження з розробки та вдосконалення засобів ефективної протидії їм.

Методи виявлення загроз шилінгу

Вразливість рекомендаційних систем до шилінгових атак обумовило дослідження підходів підвищення стійкості рекомендаційних систем веб-сервісів до різноманітних шилінгових атак та розробку засобів виявлення цих загроз [7, 8]. Ці підходи можна розділити на контрольовані та неконтрольовані методи виявлення. Контрольовані підходи на відміну від неконтрольованих вимагають позначки даних у процесі навчання. У наборах даних рекомендаційних систем доступність мічених даних мінімальна. Цей недолік призвів до того, що останнім часом неконтрольовані підходи використовуються частіше, ніж контрольовані.

Можна виділити такі контрольовані підходи: метод опорних векторів (SVM, support vector machine);k-найближчих сусідів (k-nearest neighbors algorithm, k-NN); багаторівнева perceptron мережа (MLP); Random Forest; Naive Bayes (Gaussian). До неконтрольованих відносяться UD-HMM (непідконтрольний метод виявлення шилінгових атак на основі прихованої моделі Маркова) та метод ієрархічної кластеризації.

Моделі виявлення шилінгових аккаунтів і атак засобами машинного навчання

Для машинного навчання і виявлення шиллінгових аккаунтів і атак використана бібліотека scikit learn, яку ще називають sklearn python, яка побудована на базі інших популярних Python-бібліотек - NumPy, SciPy та matplotlib. У цій бібліотеці, як і багатьох інших Python-пакетів, вихідний код повністю відкритий.

Бібліотека scikit learn підходить для вирішення задач попередньої обробки даних; зменшення розмірності; вибору, навчання та отримання передбачення для моделі; регресії; класифікації та кластерного аналізу.

Для аналізу, тренування та тестування було обрано найбільше впливові методи шилінгових атак, а саме: random, average і bandwagon атаки. Ці методи можуть нанести найбільшу шкоду рекомендаційній системі.

В роботі використано базу відміток користувачів, яка була згенерована з використанням алгоритмів шиллінгових атак. З використанням бібліотеки створені функції для побудови моделей, навчання та тестування. Використані такі моделі: sgd_classifier, logistic_regression, svm_classifier, random_forest, knn_classifier, gaussian_naive_bayes. Тестування проведено в три етапи за видом атак: 1-random; 2-bandwagon; 3-average, отримані результати наведені відповідно до генерованих атак в таблиці 1, таблиці 2, таблиці 3.

Таблиця 1

Результати протидіїгRandom атаки

Модель	Точність
knn_classifier	91.65%
logistic_regression	100.00%
svm_classifier	100.00%
random_forest	100.00%
sgd_classifier	100.00%
gaussian_naive_bayes	95.75%

Таблиця 2

Результати Bandwagon атака

Модель	Точність
knn_classifier	80.63%
logistic_regression	99.97%
svm_classifier	100.00%
random_forest	99.93%
sgd_classifier	100.00%
gaussian_naive_bayes	93.48%

Таблиця 3

Результати Average атака

Модель	Точність
knn_classifier	80.63%
logistic_regression	99.97%
svm_classifier	99.97%
random_forest	99.97%
sgd_classifier	100.00%
gaussian_naive_bayes	92.72%

У всіх трьох експериментах з тестовим набором даних найкраще показав себе модель sgd_classifier. Метод sgd_classifier виявив 100% шилінгових акаунтів. Тому даний метод підійде для швидкого використання в боротьбі з простими типами атак. Інші атаки є більш складними в реалізації для атакуючого, та мають менший вплив на всю систему рекомендацій. Подальша робота має бути продовжена вже виходячи з конкретного типу шилінгової атаки, яка діє на рекомендаційну систему, так як для успішного проведення атаки потрібно міксувати різні типи атак та якнайбільше маскувати шилінгові атаки, що в свою чергу потребує дуже конкретно приспособлений метод виявлення цієї атаки на конкретному ресурсі (рекомендаційній системі)

Висновки. Досліджено методи вирішення проблеми підвищення точності пропозицій рекомендаційних систем в умовах шилінгових атак. Проведено аналіз методів і моделей шилінгу та порівняння ступеня їх негативного впливу на точність пропозицій рекомендаційних систем, виділені як найбільш впливові три методи шилінгових атак, а саме: random, average і bandwagon. Розглянуто основні методи протидії шилінговим атакам та проведено модельні експерименти що до забезпечення стійкості рекомендаційних систем до загроз шилінгових атак з використанням цих методів. Найбільшу точність пропозицій користувачам за умов загроз обраних моделей шилінгових атак забезпечив метод sgd_classifier.

Результати даних досліджень у подальшому будуть спрямовані на розробку програмної моделі рекомендаційної системи для тестування її стійкості до відомих інформаційних атак.

Література

1. Ricci F., Rokach L., Shapira B., Kantor P.B. (Editors) Recommender Systems Handbook // Boston: Springer. – 2011. – 842 p. – [Electronic resource] – Access mode: https://doi.org/10.1007/978-0-387-85820-3
2. Aggarwal C. Recommender Systems: The Textbook, New York: Springer. – 2017. - 498 p. - [Electronic resource] – Access mode: https://www.academia.edu/42933732/Recommender_Systems_The_Textbook
3. Покришка С.А., Шумова Л.О. Удосконалення рекомендаційної веборієнтованої системи з використанням колаборативної фільтрації. Вісник Національного технічного університету «ХПІ». Збірник наукових праць. Серія: Інформатика та моделювання. – Харків : НТУ «ХПІ». – 2021. – № 1 (5). – С. 115 – 123.
4. Kabir E, Hu J, Wang H, Zhuo G. A novel statistical technique for intrusion detection systems // Future Generation Computer Systems, Vol.79, Part 1. – 2018. P. 303-318.
5. Gunes I., Kaleli C., Bilge A., Polat H. Shilling attacks against recommender systems: a comprehensive survey // Artificial Intelligence Review, Vol. 42. – 2014. – P. 767-799. – [Electronic resource] – Access mode: https://doi.org/10.1007/s10462-012-9364-9
6. Мелешко Є.В., Хох В.Д., Улічев О.С. Дослідження методів підвищення робастності рекомендаційних систем до інформаційних атак // Матеріали VI Міжнародної науково-практичної конференції «Актуальні питання забезпечення кібербезпеки та захисту інформації», 19-22 лютого 2020 р. – м. Київ: Вид-во Європейського університету, 2020. – С. 65-70.
7. Zhou W., Wen J., Qu Q., Zeng J., Cheng T. Shillihg attack detection for recommender systems based on credibility of group users and rating time series. – 2018. - [Electronic resource] – Access mode: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196533
8. Chala O., Novikova L., Chernyshova L. Method for detecting shilling attacks in e-commerce systems using weighted temporal rules // EUREKA: Physics and Engineering, Vol. 5. – 2019. – P. 29-36.

References

1. Ricci F., Rokach L., Shapira B., Kantor P.B. (Editors) Recommender Systems Handbook // Boston: Springer. - 2011. - 842 p. – [Electronic resource] – Access mode: https://doi.org/10.1007/978-0-387-85820-3
2. Aggarwal C. Recommender Systems: The Textbook, New York: Springer. - 2017. - 498 p. - [Electronic resource] - Access mode: https://www.academia.edu/42933732/Recommender_Systems_The_Textbook
3. Pokrishka S.A., Shumova L.O. Improved recommendatory web-based system with various collaborative filtering. Bulletin of the National Technical University "KhPI". Collection of scientific works. Series: Informatics and modeling. – Kharkiv: NTU “KhPI”. - 2021. - No. 1 (5). - S. 115 - 123.
4. Kabir E, Hu J, Wang H, Zhuo G. A novel statistical technique for intrusion detection systems // Future Generation Computer Systems, Vol. 79, Part 1. - 2018. P. 303-318.
5. Gunes I., Kaleli C., Bilge A., Polat H. Shilling attacks against recommender systems: a comprehensive survey // Artificial Intelligence Review, Vol. 42. - 2014. - P. 767-799. – [Electronic resource] – Access mode: https://doi.org/10.1007/s10462-012-9364-9
6. Meleshko E.V., Khokh V.D., Ulichev O.S. Follow-up methods for improving the robustness of recommender systems before information attacks // Proceedings of the VI International Scientific and Practical Conference "Actual nutritional security of cybersecurity and protection of information", February 19-22, 2020. - m. Kiev: View of the European University, 2020. - P. 65-70.
7. Zhou W., Wen J., Qu Q., Zeng J., Cheng T. Shillihg attack detection for recommender systems based on credibility of group users and rating time series. - 2018. - [Electronic resource] - Access mode: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196533
8. Chala O., Novikova L., Chernyshova L. Method for detecting shilling attacks in e-commerce systems using weighted temporal rules // EUREKA: Physics and Engineering, Vol. 5. - 2019. - P. 29-36.

The article deals with the problems of improving the accuracy of recommender system offers to users of content-oriented web resources in the context of shilling attacks. An analysis of external factors that can destabilize the work of recommender systems has shown the vulnerability of recommender systems to the threats of information attacks. Increasing the resistance of recommender systems to the action of negative factors will increase the accuracy and other performance indicators. The main external destabilizing factor in recommender systems is an information injection attack - shilling attacks. Shilling attackers have different goals, which leads to the development of various models of shilling attacks, which differ mainly in the level of knowledge about the objects of the recommender system and the degree of impact on it. The motives and consequences of the shilling are considered. Attackers manipulate the recommendation rate of target elements by falsifying user profiles. In order to influence the recommendation list of recommender systems, shilling attackers plant fake user-generated content profiles. Some attacks may attempt to "push" targets, others may aim to "nuke" some targets. The classification of shilling methods and models is given. A comparison of the negative impact of shilling attacks on the accuracy of recommendations of recommender systems is carried out. Ensuring the resistance of recommender systems to shilling attacks is an important condition for improving the accuracy of their work. Methods for detecting information attacks on recommender systems have been studied. Three methods of shilling attacks were chosen for analysis, training and testing (random, average and bandwago). The sgd_classifier method provided the highest accuracy in prompting users in the face of shilling threats.

Keywords: *recommendation systems, information security, shilling attacks, web resources, machine learning.*

Покришка С.А. аспірант кафедри комп’ютерних наук та інженерії Східноукраїнського національного університету імені Володимира Даля, e-mail: xakermans@gmail.com

Шумова Л.О. к.т.н., доцент кафедри комп’ютерних наук та інженерії Східноукраїнського національного університету імені Володимира Даля, e-mail:shumova@snu.edu.ua