УДК 004.9 **doi: 10.26906/SUNZ.2019.5.067**

Є. В. Мелешко, В. Д. Хох, О. С. Улічев

Центральноукраїнський національний технічний університет, Кропивницький

ДОСЛІДЖЕННЯ ВІДОМИХ МОДЕЛЕЙ АТАК НА РЕКОМЕНДАЦІЙНІ СИСТЕМИ З КОЛАБОРАТИВНОЮ ФІЛЬТРАЦІЄЮ

Об'єктом вивчення у статті є процес забезпечення інформаційної безпеки рекомендаційних систем. **Метою** є дослідження відомих моделей атак на рекомендаційні системи з колаборативною фільтрацією. **Завдання:** дослідити основні особливості відомих атак на рекомендаційні системи, а також методи виявлення та нейтралізації даних атак. Отримані такі **результати:** проведено дослідження основних моделей атак на рекомендаційні системи з колаборативною фільтрацією, їх впливу на результати роботи рекомендаційних систем, а також характерних особливостей даних атак, що можуть дозволити їх виявляти. **Висновки.** Проведене дослідження показало, що основним видом атак на рекомендаційні системи є атака ін'єкцією профілів. Даний вид атак може бути реалізований випадковою атакою, середньою атакою, атакою приєднання до більшості, популярною атакою, тощо. Дані атаки можуть використовуватися як для підвищення рейтингу цільового об'єкта, так і для пониження його рейтингу. Але існують спеціалізовані моделі атак, що ефективно працюють для пониження рейтингу, наприклад, атака любов/ненависть та атака обернена приєднанню до більшості. Усі ці атаки відрізняються одна від одної кількістю інформації, яку необхідно зібрати зловмиснику про систему. Чим більше у нього інформації, тим легше йому створити профілі ботів, які системі буде складно відрізнити від справжніх та нейтралізувати, але тим дорожче і довше підготовка до атаки. Для збору інформації про рекомендаційну систему та її базу даних може використовуватися атака зондом. Для захисту рекомендаційних систем від атак ін'єкцією профілів необхідно виявляти профілі ботів та не враховувати їх оцінки для формування списків рекомендацій. Виявити профілі ботів можна досліджуючи статистичні дані профілів користувачів рекомендаційної системи. Було розглянуто показники, які дозволяють виявляти профілі ботів та розпізнавати деякі типи атак.

Ключові слова: рекомендаційні системи, інформаційні атаки, інформаційна безпека, Інтернет-боти, накручування рейтингів, колаборативна фільтрація

Вступ

Рекомендаційні системи (РС) на основі колаборативної фільтрації (КФ) вразливі до атак зловмисників, які прагнуть змістити частоту потраплянь певного контенту чи товару у списки рекомендацій певної категорії користувачів, для чого вони накручують рейтинги деяким об'єктам системи [1, 2]. При використанні КФ розробники роблять припущення [1] – що користувачі використовують систему лише для одержання якісних рекомендацій для себе, а їх оцінки та коментарі, засновані на особистій думці та покращують рекомендації інших користувачів (напр., рейтинги продавців на сайті оголошень, дозволяють користувачам обирати чесних продавців та оминати шахраїв). Але, нажаль, у окремих користувачів системи можуть бути й інші цілі – відмінні від цілей власників веб-ресурсу та основної частини користувачів системи [2, 3]. Користувачі можуть прагнути підвищувати чи понижувати рейтинги певних об'єктів системи для просування своїх комерційних, політичних чи інших інтересів.

Основним типом атак на рекомендаційні системи з КФ для накручування рейтингів є атаки ін'єкцією профілів [1, 2, 4-10], які полягають у створенні групи профілів ботів для виконання узгоджених дій по зміщенню рейтингів об'єктів у системі. Оскільки результати роботи алгоритмів КФ залежать від дій користувачів системи, можна створювати підробні профілі, або платити справжнім користувачам за виконання дій, що будуть впливати потрібним зловмиснику чином на результати роботи системи. Для атак ін'єкцією профілів завжди буде цільовий об'єкт та об'єкти для наповнення профілю бота [1, 5]. Рейтинг цільового об'єкту зловмиснику треба збільшити, або зменшити, а нецільові об'єкти, будуть оцінюватися для наповнення профілю бота та намагання зробити його схожим на профіль справжніх користувачів атакованої системи. Зловмисник для здійснення впливу повинен досить точно імітувати дії звичайних користувачів, щоб не бути виявленим. А робастна до атак РС повинна працювати так, щоб результат від дій зловмисників був настільки малоефективним, щоб у них не було стимулів продовжувати атаки, а справжні користувачі продовжували одержувати релевантні невикривлені рекомендації.

Метою роботи є дослідження відомих моделей атак на рекомендаційні системи з КФ та виділення ознак, за якими їх можна ідентифікувати.

Загальні принципи атак на рекомендаційні системи ін'єкцією профілів

Атакою на РС будемо вважати узгоджені зусилля великої кількісті профілів щодо зміщення результатів її роботи таким чином, щоб деяка група користувачів або усі користувачі почали отримувати рекомендації, що суперечать їх потребам.

З точки зору зловмисника, найкраща атака проти рекомендаційної системи – це найбільший вплив на рейтинги за найменшу кількість зусиль з його боку. При атаці на рекомендаційну систему треба здійснювати два види зусиль:

1. Зусилля, пов'язані зі створенням профілів ботів. Реєстрація на веб-ресурсі деякої кількості профілів, навіть в автоматичному, може вимагати часу. Атаки, які потребують меншої кількості профілів, будуть більш привабливі для зловмисників.

2. Зусилля, пов'язані з наповненням профілів оцінками. Тут важливим є кількість знань зловмисника про систему. Чим більше знань у зловмисника

про розподіл оцінок у системі, тим більш реалістичними будуть виглядати профілі ботів, які він створить. Зловмиснику для атаки на РС з мінімальними витратами часу та коштів необхідно визначити мінімальну кількість профілів та оцінок, які йому треба додати до системи, щоб атака була ефективною. Принцип атак ін'єкцією профілів на рекомендаційну систему з КФ зображено на рис. 1.

Рис. 1. Принцип атак ін'єкцією профілів на рекомендаційні системи з колаборативною фільтрацією

На рис. 1 зображено приклад частини бази даних рейтингів рекомендаційної системи, на яку здійснюється атака. Позитивні оцінки об'єктам позначено символом «+», негативні символом «-», а відсутні, які система буде намагатися спрогнозувати – «?». В даному випадку системі треба спрогнозувати оцінки користувачів 5 та 6 для об'єкту f. Якщо система спрогнозує позитивні оцінки, то об'єкт f з великою ймовірністю потрапить у списки рекомендацій даним користувачам. Тому боти виставляють позитивні оцінки об'єкту f, а іншим об'єктам системи виставляють оцінки схожі на ті, які виставили користувачі 5 та 6. З огляду на те, що звичайні користувачі, схожі на користувачів 5 і 6, які оцінювали об'єкт f, ставили йому негативні оцінки, то без дій зловмисників рекомендаційна система спрогнозувала б низьку оцінку даному об'єкту, і він не потрапив би до рекомендацій. Для того, щоб нейтралізувати дану атаку треба визначити, які профілі є ботами, та не враховувати їх оцінки при формуванні списків рекомендацій. Досліджуючи профіль бота, можна розділити всі об'єкти системи на наступні множини (рис. 2):

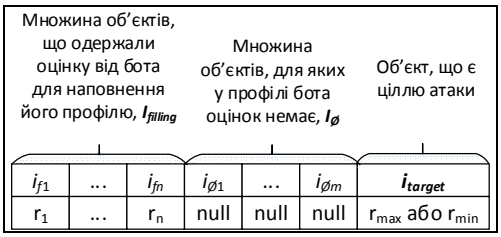

Рис. 2. Принцип оцінювання ботом об'єктів системи

Як видно з рис. 2, профіль бота містить наступні типи оцінок:

– оцінки об'єктам з множини $I_{filling}$ для імітації дій справжніх користувачів, оцінки даним об'єктам зловмисник змінювати не прагне, а навпаки намагається підібрати для них значення максимально схожі на справжні для цільової групи користувачів, на яких він прагне впливати;

– оцінка для цільового об'єкту i_{target}, це буде максимальна (чи близька до неї) оцінка у системі, якщо метою зловмисника є підвищення рейтингу даного об'єкту, або мінімальна (чи близька до неї) оцінка у системі, якщо метою є зниження рейтингу об'єкту.

Для деякої множини об'єктів у системі $I_{\emptyset}$ оцінок профіль бота не містить. Досліджуючи довжину та склад множини $I_{\emptyset}$ також можна зробити певні корисні висновки під час перевірки профілю та ідентифікації атаки і її типу.

Базові моделі атак на рекомендаційні системи ін'єкцією профілів

Розглянемо базові моделі атак на рекомендаційні системи з колаборативною фільтрацією. Найперші моделі атак було запропоновано в [3], – це випадкові та середні моделі атак. Обидві ці моделі атак передбачають генерацію профілів ботів, що будуть випадковим чином виставляти оцінки об'єктам з множини $I_{filling}$. В роботах [1, 5, 6, 8-10] розглянуто також більш складні та інформаційно-ємкі атаки.

1. Випадкова атака У профілях ботів множина $I_{filling}$ буде заповнюватися оцінками для об'єктів, вибраних випадковим чином. Оцінки обраним об'єктам будуть підбиратися також випадковим чином, але так, щоб вони були близькі до глобальної середньої оцінки у системі, напр., буде використовуватися нормальний розподіл з математичним сподіванням рівним глобальній середній оцінці. Цільовому об'єкту буте ставитися максимальна r_{max} або мінімальна оцінка r_{min}, в залежності від цілей атаки. Знання та зусилля, необхідні для здійснення такої атаки, є досить мінімальними – глобальну середню оцінку у багатьох системах можна легко дізнатися напряму або за допомогою опосередкованих даних. Ця атака не є особливо ефективною.

2. Середня атака Використовує індивідуальні середні значення оцінок кожного об'єкту для створення множини $I_{filling}$. Інформації для даної атаки треба зібрати більше. Однак середня атака може бути успішною навіть при використанні невеликого набору елементів у $I_{filling}$, що дозволяє зменшити кількість необхідної для збору інформації. Але ціною такого зменшення необхідних даних буде велика кількість профілів з однаковими оцінками, що буде, звичайно, легко виявити. Ця атака більш ефективна, ніж випадкова. Але вона практично неефективна для алгоритмів КФ типу item-based. Середня атака вимагає відносно великої кількості знань про статистику дій справжніх користувачів у системі. Розумний захист РС від таких атак буде ускладнювати нападнику збір необхідних даних. Для обходу такого захисту використовуються інші атаки, для яких вимоги до кількості знань значно нижчі.

Розглянемо існуючі атаки, що вимагають менше знань, ніж середня атака, але працюють ефективніше, ніж випадкова атака.

3. Атака приєднання до більшості Мета цієї атаки – асоціювати атакований об'єкт з невеликою кількістю об'єктів, які часто оцінюються користувачами (назвемо їх широковідомими). Зловмисник створює профілі ботів, що містять у $I_{filling}$ оцінки широковідомим об'єктам. Такі профілі мають високу

ймовірність бути схожими на велику кількість користувачів, оскільки широковідомі об'єкти – це ті, які оцінили багато користувачів. Дані для такої атаки одержати досить легко. Отже, серед широковідомих об'єктів випадковим чином обирається декілька. Цим об'єктам ставляться максимальні оцінки разом із цільовим об'єктом. Деякій частині об'єктів у $I_{filling}$ можуть ставитися випадкові оцінки, напр., як у випадковій атаці для того, щоб урізноманітнити профілі ботів. Досить ефективна атака, але, як і середня, стає неефективною при використанні проти item-based КФ.

4. Сегментна атака Основна ідея даної атаки полягає у тому, щоб змінювати рейтинг об'єкту у цільовій групі користувачів з відомими або легко передбачуваними вподобаннями. Тобто цільовому об'єкту рейтинг буде накручуватися тільки у певному сегменті користувачів, щоб він потрапляв у рекомендації тільки ним. Інакше, якщо цільовий об'єкт потрапить у рекомендації користувачам з інших сегментів він може почати отримувати від них низькі оцінки, яких буде більше, ніж накручених оцінок. Щоб здійснити таку атаку треба знайти реальних користувачів, які належать до цільового сегменту та зібрати дані про оцінки, які вони зазвичай виставляють об'єктам системи. Як і в атаці приєднання до більшості, зазвичай визначається, які об'єкти в цільовому сегменті є широковідомими. Цим об'єктам присвоюється максимальне значення оцінки разом із цільовим об'єктом. Щоб забезпечити максимальний ефект від атаки, деякі об'єкти для $I_{filling}$ обираються випадково та одержують мінімальні оцінки, що дозволяє зробити профілі ботів різними. Дана атака є ефективною проти алгоритмів КФ типу item-based. Слід зазначити, що усі розглянуті вище моделі атак можуть використовуватися для пониження рейтингу об'єкту, але існують спеціалізовані атаки, які працюють краще, ніж інші саме для пониження рейтингів. Розглянемо моделі атак призначені для пониження рейтингу об'єкту.

5. Атака любов/ненависть Ця атака дуже проста – без вимог до знань. Цільовому об'єкту присвоюється мінімальна оцінка r_{min}, а об'єкти для $I_{filling}$ одержують максимальні оцінки r_{max}. Незважаючи на надзвичайну простоту, це одна з найефективніших атак на пониження рейтингу проти user-based алгоритмів.

6. Атака обернена приєднанню до більшості Це варіант атаки приєднання до більшості, описаний вище, в якому для $I_{filling}$ вибираються широковідомі об'єкти, яким переважна більшість користувачів ставить низькі оцінки. Цим об'єктам у профілях ботів присвоюються низькі оцінки, а також низька оцінка присвоюється цільовому об'єкту. Таким чином, цільовий об'єкт починає асоціюватися з об'єктами, що не подобаються великій кількості користувачів, і це збільшує ймовірність того, що для об'єкта будуть прогнозуватися низькі оцінки і він не буде потрапляти у списки рекомендацій. Хоча ця атака не є настільки ефективною, як середня атака з великою кількістю знань для user-based систем, вона є дуже ефективною атакою на пониження рейтингів проти item-based систем.

Атаки з низьким рівнем знань використовують широковідомі об'єкти для наповнення профіля бота оцінками для них. Таким чином зловмисник може створити профіль схожий на середньостатистичного користувача, дослідивши оцінки лише широковідомих об'єктів. Якщо зловмисник знає, який саме алгоритм використовує рекомендаційна система, він може зібрати більше інформації для атаки. Таким чином атаки можна класифікувати на: атаки з малою кількістю знань (цей тип атак не потребує детальних знань про розподіли оцінок у системі, він вимагає системно-незалежних знань, які легко можна отримати за допомогою публічних джерел інформації); атаки з великою кількістю знань (зловмиснику потрібно мати якнайбільше знань про алгоритми системи та розподіли оцінок у об'єктів системи).

Прикладом атаки з великою кількістю знань є популярна атака.

7. Популярна атака Припустимо, що система використовує стандартний user-based алгоритм КФ, де подібність між користувачами визначається за допомогою кореляції Пірсона. Аналогічним чином, як і в атаці приєднання до більшості, $I_{filling}$ заповнюється з використанням широковідомих об'єктів системи. Однак це не гарантує високої схожості між профілем бота та справжніми профілями. Тому популярна атака використовує середні значення оцінок та оцінює об'єкти для $I_{filling}$ (r_{min} + 1) або r_{min}, залежно від того, чи є середня оцінка для об'єкту вищою чи нижчою. Така стратегія призведе до позитивних кореляцій між профілями ботів та автентичними профілями. Для визначення широковідомих об'єктів не потрібно багато знань, але для визначення середніх оцінок обраних об'єктів треба зібрати багато інформації. Популярну атаку можна легко налаштувати також для атак на пониження рейтингу. Популярну атаку можна виявляти, якщо порівнювати профілі у системі – профілі ботів будуть сильно схожими.

8. Атака зондом для зібрання інформації Профілі ботів тим важче розпізнати, чим більш схожі їх оцінки на оцінки справжніх користувачів. Знання про реальні вподобання різних сегментів користувачів можна отримати із самої системи через атаку зондом. Для здійснення цієї атаки зловмисник створює насіннєвий профіль, а потім використовує його для одержання рекомендацій з системи. Ці рекомендації формуються на основі інформації системи про реальних користувачів, тому використання одержаних рекомендацій дозволить створити профілі ботів більш схожі на справжніх користувачів. Можна здійснювати зондування невеликої частини користувачів, щоб потім вплинути на малу групу, як у сегментній атаці, або великої частини – щоб одержати інформацію, напр., для середньої атаки. Зловмиснику потрібно використовувати лише невелику кількість насіннєвих профілів для того щоб рекомендаційна система сама надала йому потрібну інформацію у вигляді рекомендацій.

Способи виявлення профілів ботів у рекомендаційних системах

Для того, щоб здійснити атаку, зловмисник намагається отримати доступну інформацію про систему, зокрема, про розподіл оцінок, але він не може дізнатися всю інформацію. Тому профілі ботів, завжди будуть відрізнятися від справжніх користува-

чів. Для виявлення профілів ботів можна використовувати наступні загальні ознаки [6, 8, 9]:

1. Відхилення оцінок від середньої угоди (RDMA) можна використати для знаходження профіля зловмисника:

$$RDMA_u = \sum_{i=0}^{n_u} \left| \left(r_{u,i} - \overline{r}_i \right) / l_i \right| / n_u \ , \qquad (1)$$

де n_u – кількість об'єктів, які оцінив користувач u; r_u – оцінка, яку поставив користувач u елементу i; l_i – кількість оцінок, виставлених об'єкту i всіма користувачами; $\overline{r}_i$ – середнє значення усіх оцінок об'єкту i.

2. Зважене відхилення від середньої угоди (WDMA), можна використати для знаходження профілів користувачів, що вносять значний вплив на зміну рейтингів деяких об'єктів:

$$WDMA_u = \sum_{i=0}^{n_u} \left| \left(r_{u,i} - \overline{r}_i \right) / l_i^2 \right| / n_u \ . \qquad (2)$$

3. Ступінь подібності з топ-сусідами (DegSim), дозволяє виявляти групи ботів, оскільки профіль бота буде сильніше схожий на профілі найбільш схожих на нього користувачів, ніж це відбувається з профілями справжніх користувачів:

$$DegSim_u = \sum_{v=1}^{k} sim_{u,v} / k \ , \qquad (3)$$

де $sim_{u,v}$ – коефіцієнт подоби між користувачами u, v.

4. Відхилення у кількості оцінок. У системі, що має дуже велику базу даних об'єктів, справжні користувачі навряд чи будуть оцінювати великий процент об'єктів системи, оскільки це довго та недоцільно на практиці. А профілі ботів можуть виставляти значно більше оцінок, оскільки вони роблять це автоматично. Тому можна спробувати визначити профілі ботів за допомогою відхилення кількості оцінок в певному профілі від середньої кількості оцінок в профілях користувачів у базі даних системи. Цей показник можна обчислити за формулою:

$$QD_u = \left| n_u - \overline{n} \right| / \sum_{k \in U} (n_k - \overline{n})^2 \ , \qquad (4)$$

де $\overline{n}$ – середня кількість оцінок у профілях користувачів системи; n_u – те ж саме для користувача u.

Можна також врахувати, що кожна модель атаки має свої специфічні та характерні особливості.

Показники, що можуть допомогти виявити деякі з моделей атак [6, 9, 10]:

1. Середнє відхилення значень оцінок (MeanVar) використовується для виявлення середніх атак на підвищення чи пониження рейтингу та визначення цільових об'єктів зловмисника:

$$\text{MeanVar} = \sum_{i \in P_t} (r_{i,u} - \overline{r}_i)^2 / n_u \ , \qquad (5)$$

де P_t – множина об'єктів, серед яких гіпотетично є цільові; $r_{i,u}$ – оцінка користувача u об'єкту i; $\overline{r}_i$ – середня оцінка об'єкту i серед усіх користувачів; n_u – кількість оцінок у профілі користувача u.

Необхідно обчислювати *MeanVar* для кожного можливого p_t у профілі користувача u, де p_t вибирається з об'єктів, які отримують оцінку r_t (максимальну – для атаки підвищення рейтингу або мінімальну – для атаки пониження рейтингу).

2. Різниця між середніми значеннями оцінок гіпотетичних цільових об'єктів та об'єктів для наповнення профілю (FMTD). Цей показник використовується для виявлення атаки приєднання до більшості, атаки зворотної приєднанню до більшості, сегментної атаки. У множину P_t потрапляють всі об'єкти, яким користувач u поставив цільову оцінку r_t. У множину P_f потрапляють усі інші об'єкти, оцінені користувачем u.

$$FMTD_u = \left| \sum_{i \in P_t} r_{u,i} / n_t - \sum_{k \in P_f} r_{u,k} \ / n_f \right| , \qquad (6)$$

де P_t – множина об'єктів у профілі користувача u, які мають цільові оцінки від нього; P_f – усі інші об'єкти у профілі користувача u; n_u – кількість цільових оцінок у профілі користувача u; n_f – кількість всіх інших оцінок.

Методи захисту від атак на РС поділяються на методи підвищення надійності самих рекомендаційних систем використанням більш робастних до атак алгоритмів (зокрема, надання переваги алгоритмам заснованим на моделях, замість алгоритмів заснованих на пам'яті, додаванні коефіцієнтів довіри у систему для оцінки користувачів), а також створенням системи детектування та нейтралізації профілів ботів у РС.

Висновки

Проведене дослідження показало, що основним видом атак на рекомендаційні системи є атаки ін'єкцією профілів. Даний вид атак може бути реалізований випадковою атакою, середньою атакою, атакою приєднання до більшості, популярною атакою, тощо. Дані атаки можуть використовуватися як для підвищення рейтингу цільового об'єкта, так і для пониження його рейтингу. Але існують спеціалізовані моделі атак, що ефективно працюють для пониження рейтингу. Усі ці атаки відрізняються одна від одної кількістю інформації, яку необхідно зібрати зловмиснику про систему. Чим більше у нього інформації, тим легше йому створити профілі ботів, які системі буде складно відрізнити від справжніх та нейтралізувати, але тим дорожче і довше підготовка до атаки. Для збору інформації про рекомендаційну систему та її базу даних може використовуватися атака зондом.

Для захисту рекомендаційних систем від атак ін'єкцією профілів необхідно виявляти профілі ботів та не враховувати їх оцінки для формування списків рекомендацій. Виявлення ботів можна здійснювати на основі аналізу статистичних даних користувачів системи – профілі ботів та звичайних користувачів будуть відрізнятися, що можна виявити за допомогою ряду розглянутих метрик. Також кожна з досліджених атак має свої характерні особливості, при виявленні яких можна ідентифікувати профілі як зловмисні.

Також для створення надійних рекомендаційних систем слід враховувати, що різні алгоритми колаборативної фільтрації мають різну робастність до атак ін'єкцією профілів, зокрема, item-based алгоритми надійніші, ніж user-based алгоритми, а алгоритми засновані на моделях надійніші, ніж алгоритми засновані на пам'яті.

СПИСОК ЛІТЕРАТУРИ

1. Recommender Systems Handbook / Editors Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor. – 1st edition. – New York, NY, USA: Springer-Verlag New York, Inc. – 2010. – 842 c.
2. Lam S.K., Riedl J. Shilling recommender systems for fun and profit // In Proceedings of the 13th International World Wide Web Conference. – 2004. – pp. 393–402.
3. Улічев О.С., Мелешко Є.В. Програмне моделювання поширення інформаційно-психологічних впливів у віртуальних соціальних мережах // Сучасні інформаційні системи. – 2018. – Т. 2, № 2. – С. 35-39.
4. O'Mahony M.P., Hurley N.J., Silvestre G.C.M. Promoting recommendations: An attack on collaborative filtering // from book Database and Expert Systems Applications: 13th Int. Conf., DEXA Aix-en-Provence, France. – 2002. – pp. 494-503.
5. A.Williams C., Mobasher B., Burke R. Defending recommender systems: detection of profile injection attacks // Service Oriented Computing and Applications. – 2007. – pp. 157–170.
6. Chirita P.A., Nejdl W., Zamfir C. Preventing shilling attacks in online recommender systems // In Proceedings of the ACM Workshop on Web Information and Data Management. – 2005. – pp. 67–74.
7. Zhou W., Wen J., Qu Q., Zeng J., Cheng T. Shilling attack detection for recommender systems based on credibility of group users and rating time series // PLoS ONE 13(5): e0196533. – 2018. – DOI: https://doi.org/10.1371/journal.pone.0196533
8. Kumari T., Punam B. A Comprehensive Study of Shilling Attacks in Recommender Systems // IJCSI International Journal of Computer Science Issues, Volume 14, Issue 4. – 2017. – URL: https://www.ijcsi.org/papers/IJCSI-14-4-44-50.pdf
9. Mobasher B., Burke R., Bhaumik R.,Williams C. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness // ACM Transactions on Internet Technology, Vol. 7(4). – 2007. – 41 p.
10. Mobasher B., Burke R., Bhaumik R., Williams C. Effective attack models for shilling item-based collaborative filtering system // In Proceedings of the WebKDD Workshop . – 2005. – 8 p.

Рецензент: д-р техн. наук, проф. С. Г. Семенов,
Національний технічний університет «ХПІ», Харків
Received (Надійшла) 10.07.2019
Accepted for publication (Прийнята до друку) 18.09.2019

Исследование известных моделей атак на рекомендательные системы с коллаборативной фильтрацией

Е. В. Мелешко, В. Д. Хох, А. С. Уличев

Объектом изучения в статье является процесс обеспечения информационной безопасности рекомендательных систем. **Целью** является исследование известных моделей атак на рекомендательные системы с коллаборативной фильтрацией. **Задачи:** исследовать основные особенности известных атак на рекомендательные системы, а также методы выявления и нейтрализации данных атак. Получены следующие **результаты:** проведено исследование основных моделей атак на рекомендательные системы с коллаборативной фильтрацией, их влияние на результаты работы рекомендательных систем, а также характерные особенности данных атак, которые могут позволить их выявлять. **Выводы.** Проведенное исследование показало, что основным видом атак на рекомендательные системы является атака инъекцией профилей. Данный вид атак может быть реализован случайной атакой, средней атакой, атакой присоединения к большинству, популярной атакой, и др. Данные атаки могут использоваться как для повышения рейтинга целевого объекта, так и для понижения его рейтинга. Но существуют специализированные модели атак, которые эффективно работают для понижения рейтинга, например, атака любовь/ненависть и атака обратная присоединению к большинству. Все эти атаки отличаются друг от друга количеством информации, которую необходимо собрать злоумышленнику о системе. Чем больше у него информации, тем легче ему создать профили ботов, которые системе будет сложно отличить от настоящих и нейтрализовать, но тем дороже и дольше подготовка к атаке. Для сбора информации о рекомендательной системе и ее базе данных может использоваться атака зондом. Для защиты рекомендательных систем от атак инъекцией профилей необходимо выявлять профили ботов и не учитывать их оценки для формирования списков рекомендаций. Выявить профили ботов можно, исследуя статистические данные профилей пользователей рекомендательной системы. Были рассмотрены показатели, позволяющие выявлять профили ботов и распознавать некоторые типы атак.

Ключевые слова: рекомендательные системы, информационные атаки, информационная безопасность, Интернет-боты, накручивания рейтингов, колаборативних фильтрация.

The research to known attack models for recommendation systems with collaborative filtering

Ye. Meleshko, V. Khokh, O. Ulichev

The **subject matter** of the article is the process of procuring information security of recommendation systems. The **goal** is to investigate known attack models for recommendation systems with collaborative filtering. The **tasks** to be solved are: to study the main features of known attacks on recommendation systems, as well as methods for identifying and neutralizing these attacks. The following **results** were obtained: The researches to the main attack models on recommendation systems with collaborative filtering, their impact on the work results of the recommendation systems, as well as the characteristic features of these attacks that can allow them to be detected were conducted. **Conclusions.** The study found that the main type of attack on recommendation systems is the profile-injection attack. This type of attack can be implemented by the random attack, the average attack, the bandwagon attack, the popular attack, etc. These attacks can be used to increase the rating of a target object or to decrease its rating. However, there are specialized attack models that work effectively to decrease ratings, for example, a love/hate attack and reverse bandwagon attack. All these attacks differ from each other in the amount of information that an attacker needs to collect about the system. The more information he has, the easier it is for him to create bot profiles, which for a system will be difficult to distinguish from real ones and neutralize, but the more expensive and longer the preparation for an attack. The probe attack can be used to collect information about the recommendation system and its database. To protect recommendation systems from the profile-injection attacks, it is necessary to identify bot profiles and not take into account their ratings for the formation of recommendation lists. Identify bot profiles can be examining the statistics of user-profiles of the recommendation system. The study metrics that allow identifying bot profiles and recognizing some types of attacks was conducted.

Keywords: recommendation systems, information attacks, information security, Internet bots, shilling attacks, collaborative filtering.