УДК 004.8:044.89

Т.Д. Польгуль
А.А. Яровий

ВИЗНАЧЕННЯ ШАХРАЙСЬКИХ ОПЕРАЦІЙ ПРИ ІНСТАЛЯЦІЇ МОБІЛЬНИХ ДОДАТКІВ З ВИКОРИСТАННЯМ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Вінницький національний технічний університет

Анотація

У даній роботі здійснено аналіз шахрайських способів встановлення мобільних додатків, сучасних методів їх визначення, запропоновано модель визначення шахрайських способів встановлення мобільних додатків з використанням інтелектуального аналізу даних.

Ключові слова: Fraud Detection, Click Spamming, Mobile Hijacking, Action Farming, Anomaly Detection, визначення шахрайства, визначення аномалій.

Abstract

A comparative analysis of mobile install fraud techniques was done, modern techniques of mobile install fraud techniques detection was done, fraud detection technique using data mining was offered.

Keywords: Fraud Detection, Click Spamming, Mobile Hijacking, Action Farming, Anomaly Detection.

Проблема визначення шахрайських інсталяцій є актуальною, оскільки розробники мобільних додатків витрачають великі кошти на компанії, які у свою чергу зобов'язані здійснити вказану кількість інсталяцій додатку. Проте, багато з цих компаній застосовують шахрайські способи, які показують замовнику, що інсталяція відбулася, але в дійсності це не так. Або ж інсталяція дійсно відбувається, проте через певний час видаляється, оскільки здійснювалась «своєю» людиною або певним технічним забезпеченням. У кінцевому результаті замовник констатує, що компанія здійснила замовлену кількість встановлень додатку, а через невеликий проміжок часу цієї кількості користувачів немає.

Метою даного дослідження є розробка моделі визначення шахрайських інсталяцій мобільних додатків з використанням інтелектуального аналізу даних.

Для вибору моделі визначення шахрайських операцій при інсталяції мобільних додатків, необхідно визначити природу існуючих та можливих шахрайських способів інсталяції додатків. Серед шахрайських способів інсталяцій додатків можна виділити: кліковий спам (Click Spamming) [1-3], мобільне викрадення (Mobile Hijacking), ферми дій (Action Farms) [1].

Серед сучасних та найбільш очевидних контрзаходів, які вже стали своєрідним стандартом у цій галузі, можна виділити: IP-фільтрацію, блокування видавця, зовнішню фільтрацію натиснень, виявлення стрибків при кліках або запитах інсталяції. Також, існують методи, які визначають співвідношення населення по геолокації, використовують аналіз дельти часу між подіями (такі відомі фірми як Adjust та Kochava), аналізують показники продуктивності для визначення шахрайства [1].

Зважаючи на вищевказані шахрайські способи інсталяції мобільних додатків та аналізуючи дані власного мобільного додатку, можна зробити висновок, що події, які відбуваються шахрайським способом, мають спільні ознаки. Використовуючи методи кластеризації, користувачів, яких було залучено шахрайським способом, можна віднести до одного кластеру, правильно визначивши ознаки, за якими здійснювати кластеризацію.

Проаналізувавши відомі методи кластеризації та класифікації [4-10], у даній роботі запропоновано математичну модель знаходження подібних користувачів [9-14]. Розроблена математична модель базується на модифікованому методі колаборативної фільтрації [9] і розв'язує багатокритеріальну задачу визначення подібних користувачів. Система складається з трьох модулів.

Отже, першим кроком необхідно зібрати інформацію, для цього у розробленій моделі є модуль збору інформації, який отримує на вхід «сиру», тобто необроблену, інформацію, структурує та розподіляє отримані дані у базу користувачів та базу дій кожного з користувачів

та зберігає їх у структурованому вигляді. Модуль збору інформації обробляє такі вхідні дані системи: дії користувача при встановленні мобільного додатку та після його встановлення, по кожній дії користувача доступна його геолокація, час дії, IP тощо.

Далі, необхідно визначити подібність користувачів у відповідному модулі, маючи по кожному з них набір інформації, тобто вектори зі значеннями по кожній з ознак. Для визначення подібності користувачів, маючи набір зібраної та структурованої нами інформації по кожному з них, використовуються різні коефіцієнти схожості, серед розглянутих коефіцієнтів було доведено покращення визначення подібності користувачів [11] при використанні комбінованої метрики схожості, яка формується на основі коефіцієнта косинусної схожості між двома векторами (1) та коефіцієнта Танімото (2) [13]. Коефіцієнт знаходить подібність між усіма парами користувачів.

Подібність користувачів визначається в модулі визначення подібності користувачів за допомогою коефіцієнта косинусної схожості між двома векторами [14].

$$k = \cos(a, b) = \frac{(a \cdot b)}{|a| \cdot |b|} \quad (1)$$

де A, B - вектори, елементами яких є частоти появи окремих ознак у заданому наборі інформації.

$$k = T(A, B) = \frac{N_c}{N_a + N_b - N_c} \quad (2)$$

де Na – кількість елементів у наборі даних користувача A,
Nb – кількість елементів у наборі даних користувача B,
Nc – кількість елементів в їх перетині.

У задачі, яку виконує модуль визначення подібності користувачів, залишається незрозумілим, як визначати подібність користувачів, маючи набір різнорідних ознак, серед яких є і геолокація, і час виконання різних дій користувачем, і IP тощо, та підібрати для кожної з ознак певний коефіцієнт значимості. Для вирішення цієї проблеми необхідно:
- оскільки набір даних містить як числові, так і дискретні дані, необхідно дискретні дані перевести у числові;
- далі, для вирішення проблеми з різнорідністю даних, застосовується модуль визначення схожості дій користувача, який використовує багатовимірне шкалювання [9], яке використовується саме для того, щоб зрозуміти, як різнорідні дані пов'язані між собою [9]. Алгоритм створює уявлення набору даних в просторі меншої розмірності, намагаючись по можливості зберегти вихідні відстані між елементами. Якщо мова йде про подання на екрані або на папері, то багатовимірний набір представляється у двовимірному просторі [9].

Отримавши усю зібрану та стуктуровану інформацію у двовимірному просторі, отриманий вектор подається у модуль подібності користувачів, який на вихід віддає відсортований вектор із подібними користувачами.

Розроблена модель для визначення подібності користувачів базується на модифікованому методі колаборативної фільтрації, оскільки методи колаборативної фільтрації використовувались у рекомендаційних системах таких великих компаній як Netflix, Amazon тощо. А як відомо, одним з етапів побудови рекомендаційних систем є знаходження найбільш подібних між собою користувачів.

Отже, у даній роботі проаналізовано шахрайські способи встановлення мобільних додатків, запропоновано та розроблено модель знаходження подібності користувачів на основі модифікованого методу колаборативної фільтрації з метою знаходження користувачів, створених при встановленні мобільних додатків шахрайськими способами та користувачів, створених при органічному встановленні додатку.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Our take on mobile fraud detection [Електронний ресурс]. – Режим доступу: http://geeks.jampp.com/data-science/mobile-fraud/ (дата звернення 10.11.2016)
2. Vacha Dave / ViceROI: Catching Click-Spam in Search Ad Networks. / Vacha Dave, Saikat Guha, Yin Zhang [Електронний ресурс]. – Режим доступу: http:// www.sysnet.ucsd.edu/ ~vacha/ccs13.pdf (дата звернення 09.11.2016)
3. Dave, V., Guha, S., and Zhang, Y. Measuring and Fingerprinting Click-Spam in Ad Networks. In Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM) (Helsinki, Finland, Aug. 2012), pp. 175–186.

4. MachineLearning.ru [Електронний ресурс]. – Режим доступу: http://www.machinelearning.ru
5. Varun Chandola / Anomaly Detection : A Survey. / Varun Chandola, Arindam Banerjee, Vipin Kumar [Електронний ресурс]. – Режим доступу: http://cucis.ece.northwestern.edu/ projects/DMS/publications/AnomalyDetection.pdf (дата звернення 20.01.2017)
6. Agarwal, D. 2005. An empirical bayes approach to detect anomalies in dynamic multidimensional arrays. In Proceedings of the 5th IEEE International Conference on Data Mining. IEEE Computer Society, Washington, DC, USA, 26–33.
7. Agarwal, D. 2006. Detecting anomalies in cross-classified streams: a bayesian approach. Knowledge and Information Systems 11, 1, 29–44.
8. Agrawal, R. and Srikant, R. 1995. Mining sequential patterns. In Proceedings of the 11th International Conference on Data Engineering. IEEE Computer Society, Washington, DC, USA, 3–14.
9. Сегаран Т. Программируем коллективный разум. / Т. Сегаран; пер. с англ. А. Слинкина – СПб: Символ-Плюс, 2008. – 368 с., ил. – ISBN 5-93286-119-3.
10. Савчук Т. О. / Кластеризація станів комп'ютерної техніки з використанням інфрпмаційної технології. Вісник Хмельницького національного технічного університету (серія: технічні науки), 149-152.
11. Кюльян А. Г. / Математична модель рекомендаційного сервісу на основі методі колаборативної фільтрації. / Кюльян А. Г., Польгуль Т. Д., Хазін М.Б. [Електронний ресурс]. – Режим доступу: http://ir.lib.vntu.edu.ua/bitstream/handle/123456789/7911/226-227.pdf?sequence= 1&isAllowed=y
12. Alexander T. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions / T. Alexander – IEEE Trans. On Knowledge and Data Engineering, vol. 17, no. 6, June 2005 , pp. 734-749.
13. Linden G.D. Collaborative Recommendations Using Item-to-Item Similarity Mappings/ G.D.Linden, J.A.Jacobi, E.A Benson – US Patent 6,266,649 (to Amazon.com), Patentand Trade-mark Office, Washington, D.C., 2001.
14. Sarwar B.M. Item-Based Collaborative Filtering Recommendation Algorithms / B.M. Sarwar, – 10th Int'l World Wide Web Conference, ACM Press, 2001, pp. 285-295.

Польгуль Тетяна Дмитрівна – аспірант кафедри комп'ютерних наук ВНТУ, Вінницький національний технічний університет, м. Вінниця, e-mail: tanapolg93@gmail.com
Науковий керівник: Яровий Андрій Анатолійович – д.т.н., професор, професор кафедри комп'ютерних наук ВНТУ, Вінницький національний технічний університет, м. Вінниця, e-mail: a.yarovyy@vntu.edu.ua

Tetiana D. Polhul – postgraduate student of the Computer Sciences Chair, Vinnytsia National Technical University, Vinnytsia, e-mail: tanapolg93@gmail.com
Scientific Supervisor: Andrii A. Yarovyi – Doctor Sc. (Eng), Professor, Professor of the Computer Sciences Chair, Vinnytsia National Technical University, Vinnytsia, e-mail: a.yarovyy@vntu.edu.ua