УДК 004.62
В. Грицюк, М. Стадник
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

КЛАСТЕРИЗАЦІЯ СПАМ-ДОМЕНІВ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

UDC 004.62
V. Hrytsiuk, M. Stadnyk

SPAM DOMAINS CLUSTERIZATION BY USING MACHINE LEARNING METHODS

Covid-2019 спричинив значний поштовх для ще більшого розвитку електронної комерції. Кількість людей, що замовляла одяг, побутові товари чи товари першої необхідності через інтернет тільки зростала. Водночас зростав і попит на різноманітні мобільні додатки та веб-інтерфейси для успішного ведення бізнесу та представлення бренду в мережі. Після завершення пандемії тенденція замовляти товари із супермаркетів чи з локальних брендів зберіглася. Відповідно власник торгового маркетплейсу чи брендованого магазину на Shopify чи іншій платформі повинен задуматись над безпекою транзакції та над стабільністю веб-додатку. Звичайно існують системи захисту банківських транзакцій, багатофакторна аутентифікація, Google reCaptcha v2 чи v3, проте зловмисники вигадають нові способи зробити електронний ресурс недоступним. Задача виявлення зловмисника або ресурсу, що здійснює DDoS атаку, ідентифікувати його як спам-домен є актуальною [1].

Спам-доменом називають домен, який попав у спам-список. Ідентифікація спам домена є попереднім етапом, перед тим як спам домен попадає у чорний список. При надсиланні листа чи при доступі до ресурсу веб додатку сервер розпізнає запит від вказаної IP адреси, перевіряє наявність її у спам списку і лише тоді здатен виконати запит. Звичайно, якщо IP адрес не вказаний у спам списку, сервер буде багатократно здійснювати свою роботу таким чином відбувається навантаження на систему та врешті – відмова у доступі. Формування актуального спам-списку сприяє вчасному виявленню зловмисних дій.

Для актуалізації спам-списку, що включає як новостворені домени, так і раніше сформовані використовують кілька технік. Найпростішою методикою є виявлення спам-домена на основі його дій, тобто пост-фактум, при цьому зловмисник уже домігся своєї цілі.
Іншою технікою є збір параметрів про IP адресу та з використанням отриманої інформації відбувається класифікація IP адреси з використанням методів машинного навчання чи штучного інтелекту. Для прикладу сервіс Alexa таким чином рангує домени, а з метою захисту користувачів stop-forumspam.org висвітлює «токсичні» домени [2].

Однією проблемою при виконанні класифікації чи кластеризації є поява значної кількості нових доменів щоденно. Зважаючи на таку тенденцію необхідно виконувати агрегацію новостворенних доменних імен і в зазначеному періоді здійснювати кластеризацію. В роботі було використано метод k-найближчого сусіда, дерево рішень, алгоритми на основі графів. За результатами оцінок метод k-найближчих сусідів є найбільш оптимальним для поставленої задачі.

Література

1. Chio, C.; Freeman, D. Machine Learning and Security, O'Relly Media, 2018, 125–180.
2. Webb, S., Caverlee, J. «Characterizing Web Spam Using Content and HTTP Session Analysis». The 4th Conference on Email and AntiSpam. Aug. 2007. Mountain View, CA.