

ОЦІНКА СЛАБКОСТІ ПАРОЛІВ НА ОСНОВІ ВІДКРИТОЇ ІНФОРМАЦІЇ

К. О. Стецюк¹, О. М. Барановський¹

¹Національний технічний університет України «Київський політехнічний інститут»

Анотація

В даній роботі розглянуто та представлено рішення проблеми злому паролів на основі відкритої інформації користувача. Проведено аналіз даних та створено програму, яка визначає стійкість паролю на основі відкритих даних. Робота програми базується на методах машинного навчання з вчителем.

Ключові слова: пароль, машинне навчання, класифікація, атака, злом

Вступ

У наш час надзвичайно важливим є захист особистих даних, а саме забезпечення цілісності, конфіденційності та доступності інформації. Компанії витрачають велику кількість ресурсів на побудову якісних систем інформаційної безпеки. У кожній людини є приватна інформація, ділитися якою вона не вважає потрібним. Зловмисники найрізноманітнішими способами намагаються викрасти, спотворити приватну інформацію. Існують різні методи захисту інформації, починаючи від простої схованки, сейфу і закінчуючи біометричними методами захисту інформації. Найбільш популярний метод автентифікації – це використання паролю.

1. Методи злому паролів

Всі методи злому базуються на використанні принципів: щось знаю, щось маю і чимось являюсь. Найвідоміші з методів: атаки за допомогою словника, атаки повного перебору (англ. *brute force*), використання райдужних таблиць, фішинг, соціальна інженерія, метод «павуків» та шкідливе програмне забезпечення.

Метод атаки за допомогою словника використовує файл, який містить великий об'єм слів зі словника. Тут виконується простий перебір усіх слів, і якщо не вдається знайти пароль одним словом, то використовуються комбінації слів.

В основі методу райдужних таблиць є список попередньо обчислених хеш-значень (числових значень зашифрованих паролів), які використовуються більшістю сучасних систем. Таблиця включає в себе хеш-значення всіх можливих комбінацій паролів для будь-якого виду алгоритму хешування. Час, необхідний для злому пароля за допомогою райдужної таблиці, зводиться до того часу, який потрібен, щоб знайти хеш-значення паролю у списку.

Фішинг використовує надсилання повідомлення, яке переводить користувача, який нічого не підозрює на підроблені сайти онлайн-банкінгу, платіжних систем або інші сайти, на яких потрібно обов'язково ввести особисті дані. Метод соціальної інженерії до-

тримується тієї ж концепції, що і фішинг – «запитати у користувача пароль», але не за допомогою поштової скриньки, а в реальному світі.

Шкідливе програмне забезпечення здійснює перехват всієї інформації, яку вводить користувач з клавіатури, або яка виводиться на екран.

Метод «павуків» базується на тому, що створюється список слів будь-якої тематики, який потім використовується для злому методом грубої сили. Зловмисники автоматизували процес і запускають «павутинні» додатки, аналогічні тим, які застосовуються провідними пошуковими системами, щоб визначити ключові слова, зібрати і обробити списки для злому.

Здійснити атаку на паролі, які використовують відкриту інформацію користувачів можна використовуючи або метод соціальної інженерії, або метод «павуків». Відомо, що більшість користувачів використовують приватну інформацію при створенні паролю. Таку інформацію дуже часто можна знайти в соціальних мережах на сторінках користувачів, в блогах та подібних ресурсах. Таким чином, використання в паролі приватної інформації призводить до легкого злому зловмисниками, використовуючи методи соціальної інженерії – випитавши, або знайшовши приватну інформацію користувача, або використовуючи метод «павуків», створивши список слів, які пов'язані з користувачем.

На основі чисельних опитувань та досліджень виявилось, що більшість людей використовують пароль, який базується на особистій інформації користувача, такий як ім'я, прізвище, дата народження, інтереси. Модифікуючи цю інформацію, змінюючи порядок слів, використовуючи різні заміни букв на цифри, використовуючи скорочення слова, користувач формує свій власний пароль, який може легко запам'ятати.

Відомим способом для відновлення паролю є встановлення підказок користувачем. Так він обирає питання, на яке має знати відповідь лише він – і ця відповідь буде підказкою, щоб згадати пароль у випадку, якщо користувач не може згадати пароль. При аналізі інформації, отриманої з облікових запи-

сів користувачів, з 7 365 869 підказок до паролів було виокремлено найбільш популярні: у 559 358 користувачів стоїть підказка «собака», у 479 828 – «ім'я», у 242 150 – «я», у 183 943 – «син», у 181 047 – «донька», далі йдуть «робота», «дата народження» і інші подібні слова, які виражають персональну інформацію користувача [1]. Більшість таких даних можна знайти в соціальних мережах, де відображається персональна інформація, місце проживання, сини, доньки, брати та сестри. Оскільки більшість людей створюють паролі самостійно, а не використовують для цього спеціальні генератори, ці паролі часто створені на основі приватної інформації користувача. За рахунок того, що люди розміщують багато інформації про себе в соціальних мережах та блогах, дана інформація являється публічною. Це продемонстровано в аналізі підказок для відновлення паролів, де використовувались імена користувачів та їх близьких людей, дата народження і подібна інформація, доступ до якої може отримати кожен.

Використання підказок, які створенні, використовуючи публічну інформацію, означає, що такі ж самі дані користувачі можуть застосовувати при створенні паролю, що призводить до швидкого його злому і, як наслідок, до порушення конфіденційності, цілісності та доступності інформації. Для перевірки цієї гіпотези було проведено аналіз особистих даних користувачів та їх паролів.

2. Аналіз даних користувачів

Проведено аналіз баз даних облікових записів користувачів відомого ресурсу «mail.ru», які були опубліковані зловмисниками в мережі. На основі дослідження виявлено, що у 70 % користувачів пароль включає в себе ім'я, прізвище або дату народження. Ці дані лише можуть доповнюватися чергуванням маленьких та великих букв, заміною букв «і» на «1» і подібними перетвореннями. І очевидно, що вся ця інформація є доступною для більшості користувачів інтернету, адже кожен може отримати цю інформацію з різних профілів, особливо з соціальних мереж.

Для аналізу паролів було використано відстань Левенштейна (відстань редагування) – це міра відмінності між двома послідовностями символів [2]. Так для кожного облікового запису оцінено відстань Левенштейна між паролем та різними іменами. В якості словника імен було взято найбільші вибірки як українських, російських імен, так і англійських. Також було сформовано вибірку дат у вигляді року, місяця, дня народження та утворено різні їх варіації.

Результат цього аналізу показує, що у 24 % користувачів паролем є ім'я у чистому вигляді (без заміни деяких букв на цифри та без додавання цифр чи слів до імені). У 47 % користувачів пароль містить ім'я, яке видозмінено лише однією або двома модифікаціями букв. У 12 % користувачів пароль складається з імені та дати народження. І лише у 17 % користувачів пароль не містить жодної інформації пов'язаної з ім'ям та датою народження. За результатами цього аналізу, сформовано гіпотезу, що користувачі вико-

ристовують саме власні імена та дати при створенні паролю.

Для перевірки цієї гіпотези було проведено аналіз іншої бази даних, яка містить не лише паролі користувачів, а особисту інформацію і пароль. В ході роботи використано таку особисту інформацію користувача, як поштова адреса, ім'я, прізвище, логін та дата народження.

Для оцінки використання особистої інформації в паролі було створено коефіцієнт, який обчислюється за формулою:

$$\text{coef}(d, p) = \frac{\text{levenshtein}(d, p)}{\max(d.\text{length}, p.\text{length})}$$

де: d – дані користувача,

p – пароль користувача,

$d.\text{length}, p.\text{length}$ – довжина поля даних користувача та паролю відповідно,

$\text{levenshtein}(d, p)$ – відстань Левенштейна (відстань редагування) між даними користувача і паролем.

Даний коефіцієнт буде знаходитися завжди в межах від 0 до 1. Чим більший буде коефіцієнт, тим менше інформації про користувача використовується у паролі. В результаті аналізу встановлено, що 7 % користувачів використовують особисті дані у чистому вигляді в паролі (коефіцієнт рівний 0), 25 – 35 % використовують пароль, у яких більше ніж $\frac{1}{2}$ паролю співпадає з одними з особистих даних (коефіцієнт в межах 0.0 – 0.5), у 20 % користувачів більше ніж $\frac{1}{3}$, але менше ніж $\frac{1}{2}$ паролю містить особисті дані. І решта створюють паролі, де менше, ніж в $\frac{1}{3}$ частині фігурує особиста інформація.

На діаграмі (рис. 1). показано відсоткове відношення між даними, які користувачі використовують в паролях.

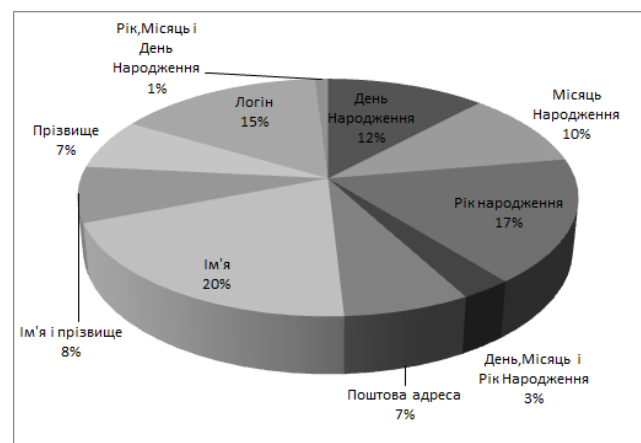


Рис. 1. Розподіл використання даних у паролі

З діаграми бачимо, що найбільше у паролях використовується ім'я, рік народження та логін користувача.

Люди все ще ігнорують поради спеціалістів з безпеки інформації не використовувати особисті дані при створенні паролів.

3. Методи вирішення проблеми

В ході аналізу даних було встановлено, що більшість користувачів створюють паролі за допомогою відкритих особистих даних. Проте, чіткого алгоритму створення таких паролів не знайдено. Це пов'язано з тим, що можна використовувати різні інтерпретації відкритих даних та по різному компонувати їх. Вирішити проблему використання таких паролів можна створивши модуль в системах реєстрації, який буде попереджувати користувача при створенні паролю, що він використовує публічну інформацію і тому є слабким (легким до злому). Оскільки немає чіткого алгоритму, яким користується кожен при створенні паролю з відкритими даними, то створення такого модулю можна розробити використовуючи методи машинного навчання, створивши класифікатор, який за вхідними даними користувача буде виявляти чи є пароль стійким до злому.

Машинне навчання – великий підрозділ штучного інтелекту, який вивчає методи побудови алгоритмів, здатних навчатися.

Машинне навчання знаходиться на стику математичної статистики, методів оптимізації та класичних математичних дисциплін, але має також і власну специфіку, пов'язану з проблемами обчислювальної ефективності і перенавчання. Багато методів індуктивного навчання розроблялися як альтернатива класичним статистичним підходам. Багато методів тісно пов'язані з отриманням інформації та інтелектуальним аналізом даних.

4. Результат використання методів машинного навчання

У роботі було використано алгоритми машинного навчання, які засновані на техніці «навчання з вчителем». Техніка базується на тому, що алгоритм приймає вхідні дані (ознаки об'єкту) та вихідні дані (належність об'єкта до визначеного класу). В процесі навчання для кожного набору з вхідних даних вказуються «вчителем» правильні відповіді. Таким чином алгоритм навчається і в результаті сам за вхідними даними вказує правильну відповідь (вихідні дані). У даній роботі вхідними даними були мінімальний коефіцієнт Левенштейна по всіх даних, коефіцієнт для кожного поля даних та довжина паролю. Є два варіанти для вихідних даних: слабкий та стійкий пароль.

Було використано такі алгоритми машинного навчання: логістична регресія [3], градієнтний бустинг [4], метод опорних векторів [5] та просте голосування [6]. Для перевірки якості навчання класифікатора було використано метод крос-валідації. Цей метод оцінює достовірність математичної моделі для незалежних даних. Тобто вибірка даних, на яких проводиться навчання розділяється на декіль-

ка частин (в даній роботі здійснювався розподіл на 5 частин), і кожна частина по черзі виступає в ролі тестової вибірки. Це означає, що навчання відбувається на чотирьох частинах, а на п'ятій перевіряється достовірність класифікації. Після крос-валідації можна отримати точність алгоритму класифікації. В результаті для кожного методу отримано такі значення точності класифікації: логістична регресія – 87 %, градієнтний бустинг – 85 %, метод опорних векторів – 86 % і метод простого голосування (в даній роботі це змішування методів логістичної регресії та градієнтного бустингу), який виявився найточнішим і показав 92 % точності.

Даний результат означає, що було створено класифікатор, який з 92 % точністю показує чи є пароль слабким на основі відкритої інформації. В подальшій роботі, за даною моделлю буде створено модуль, який при заповненні реєстраційної форми особистими даними та встановленням паролю, повідомляє чи є пароль слабким до злому на основі відкритої інформації.

Висновки

Незважаючи на поради експертів не створювати пароль, який пов'язаний з ім'ям, датою народження, іменами близьких людей, іменами домашніх улюбленців та з іншою, пов'язаною з особою, інформацією, люди не перестають створювати саме такі паролі. Так, вони легкі до запам'ятовування, проте є найуразливішими до атак та злому. Тому важливо постійно нагадувати користувачам про те, що паролі не повинні містити відкритої інформації користувачів, тому що це призводить до втрат величезної кількості важливої інформації. Реалізація такого виду нагадування можлива через встановлення на сайтах модулю, що оснований на класифікації методом машинного навчання.

Перелік використаних джерел

1. Troy Hunt. Adobe credentials and the serious insecurity of password hints. — 2014. — URL: <https://www.troyhunt.com/adobe-credentials-and-serious/>.
2. Manning Christopher D. Introduction to Information Retrieval. — 2008.
3. David W. Hosmer Jr. Stanley Lemeshow Rodney X. Sturdivant. Applied Logistic Regression. — 2013.
4. Freadman Jerome H. Greedy Function Approximation: A Gradient Boosting Machine. — 2001.
5. Wang Lipo. Support Vector Machines: Theory and Applications. — 2005.
6. Bishop Christopher M. Pattern Recognition and Machine Learning. — 2006.