# Метричний аналіз та обчислювальна стійкість цільових словників паролів

УДК 004.056.52 (519.1)

Сергій Бабич[1], Андрій Сидор[2], Петро Голуб[3]

*Національний університет водного господарства та природокористування,*
*[1]s.v.babych@nuwm.edu.ua, [2]a.i.sydor@nuwm.edu.ua, [3]p.p.holub@nuwm.edu.ua*

Дослідження спрямовані на перетворення хаотичної генерацію варіантів у керований і обчислювально ефективний процес [1] синтезу ефективних цільових словників. Аби дослідження тримати в межах актуальності, необхідно оцінювати ефективність перетворення даних з «відомостей про сутність» (вхідні дані у форматі OSINT – дослідження) у подальші токени (стадія 1), паролі (стадія 2) та їх об'єднання – безпосередньо словник (стадія 3).

Центральним об'єктом запропонованої моделі є парольний токен — мінімальна семантично-персональна одиниця, вилучена з OSINT-профілю [2]. Наукова новизна підходу полягає у відмові від розуміння токена як лінгвістичної морфеми (наприклад, у алгоритмах BPE чи SentencePiece) на користь суб'єктивно значущої сутності (ім'я, дата, місце). Для структурування OSINT-даних запропоновано семикласову таксономію (T1–T7) [3], що включає ідентифікатори, часові, реляційні, просторові, тематичні, числові та контекстуальні дані. Описавши об'єкт оцінки ефективності, варто перейти до компоненти дослідження, що стосується вже саме оцінки ефективності.

Аналітичний огляд досліджень у сфері оцінки ефективності словників паролів демонструє поступовий перехід від класичних методів до складних моделей, що враховують як структурні характеристики паролів, так і контекстні дані користувачів. Розвиток цієї тематики відбувся у працях, де були запропоновані моделі TarGuess, які інтегрують відкриті дані (OSINT), включно з персональними атрибутами та інформацією із соціальних мереж, для побудови цільових словників. Ефективність таких словників значно зростає завдяки використанню реальних цифрових слідів користувачів, що підтверджує важливість ймовірнісного та статистичного аналізу у вимірюванні їх стійкості. Водночас ці роботи піднімають питання етики та приватності, адже використання персональних даних у словниках може створювати додаткові ризики [4]. Трендом останніх років стало застосування нейромережевих моделей: запропоновано PassGAN, який використовує генеративні змагальні мережі для створення словників. Цей підхід дозволяє моделювати розподіл паролів без явних правил, що робить словники більш адаптивними та здатними відображати реальні патерни користувачів. Метричний аналіз у цьому випадку включає оцінку ентропії та варіативності згенерованих словників, що дозволяє порівнювати їх ефективність із класичними методами [5].

Підсумувавши дослідження інших авторів щодо оцінки ефективності словників паролів загалом та виокремивши необхідне для нашого дослідження, що стосується цільових словників, варто зауважити, що останнє десятиліття ознаменувалося переходом від простої частотної статистики до складних моделей представлення знань, де базові показники: CR (Coverage Rate) та DSR

(Dictionary Size Ratio) залишаються основними для оцінки компактності та влучності. Але варто згадати і використання математичної близькості за метрики Дамерау-Левенштейна [6], котра стала стандартом для моделювання транспозицій (перестановок сусідніх знаків), що відображає механічні помилки користувачів. Щодо порівняння моделей, звернемо увагу, що класичні ймовірнісні граматики (PCFG) розбивають пароль на сегменти [L][D][S], але ігнорують семантичний зв'язок. Нейромережеві підходи, такі як PassGAN та PASSLLM, демонструють здатність вивчати логіку витоків, проте в цільових атаках вони у 10 000 разів повільніші за комбінаторні методи та часто страждають від перенавчання (overfitting) [7]. Ідеї Representational Learning (PLR), в своєю чергою, використовують латентний простір паролів для генерації навколо опорних точок (pivots), проте вимагають значних обчислювальних ресурсів.

Як висновок, варто звернути увагу, що в межах даного представлення – окреслено бачення команди щодо оцінки ефективності сформованих словників, при формуванні яких буде використано запропоновану методику СРП з роботи [1]. Але, вже можна виокремити те, що перехід до інтегрованої моделі оцінки, що поєднує OSINT-профілювання, семантичну токенізацію та метод перманентної декомпозиції, дозволяє перетворити аудит парольної політики з хаотичного підбору на науково обґрунтовану процедуру. Запропонований підхід не лише підвищує точність генерації при цільовому використанні, а й забезпечує обчислювальну стійкість систем захисту в умовах зростання цифрового сліду користувачів.

1. С. В. Бабич, «Інформаційна технологія складання розкладу занять згідно перманентної декомпозиції», дисертація кандидата технічних наук, Хмельницький національний університет, Хмельницький, Україна, 2023.
2. M. Bazzell, Open Source Intelligence Techniques: Resources for Searching and Analyzing Online Information, 9th ed. IntelTechniques, 2022.
3. B. Rader, “Targeted password cracking with OSINT data,” Purdue School of Engineering & Technology, IUPUI, 2023.
4. M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek, “Password cracking using probabilistic context-free grammars,” in Proc. IEEE Symp. Security and Privacy (SP), Oakland, CA, USA, 2009, pp. 391–405.
5. Y. Xie, J. Wang, and J. Yan, “TarGuess II: A statistical framework for targeted password guessing,” in Proc. ACM Conf. Computer and Communications Security (CCS), London, UK, 2020, pp. 1–14.
6. F. J. Damerau, “A technique for computer detection and correction of spelling errors,” Communications of the ACM, vol. 7, no. 3, pp. 171–176, 1964.
7. B. Hitaj, P. Gasti, G. Ateniese, and F. Perez-Cruz, “PassGAN: A deep learning approach for password guessing,” in Proc. Int. Conf. Applied Cryptography and Network Security (ACNS), Bogota, Colombia, 2019, pp. 217–237.

# Аналіз витоків паролів на наявність патернів та можливості їх субслівної токенізації

УДК 004.056

Сергій Бабич[1], Петро Голуб[2], Богдан Слив'як[3]

*Національний водного господарства та природокристування,*
*[1]s.v.babych@nuwm.edu.ua, [2]p.p.holub@nuwm.edu.ua, [3]slyviak_ak23@nuwm.edu.ua*

Паролі використовуються повсюдно як засіб автентифікації користувачів, відповідно складність паролів є одним із основних факторів захищеності інформаційних систем. Кореляції ж у рамках однієї із останніх збірок витоків паролів, а саме RockYou2024 демонструють стабільну наявність популярних патернів, що стосуються персональної інформації чи цифрового сліду людини [1]. Урядові та відомчі мережі, включно з індустріальними системами управління (SCADA), також залишаються вразливими до використання передбачуваних паролів, що становить загрозу критичні інфраструктурі [2, 3].

Метою роботи є дослідження тенденцій та кореляцій у масштабних витоках паролів, а також використання отриманих результатів для подальшого дослідження граматичних особливостей парольних фраз та їх субслівної токенізації, задля верифікації складності паролів базуючись на цифровому сліді людини. Відповідні результати можуть бути надзвичайно важливими для аудиту парольних політик всередині організації або створення релевантних цільових словників паролів на основі розвідки з відкритих джерел (OSINT) [4].

Новизна даного дослідження полягає у розширенні класичних підходів до побудови словників паролів та інтеграції українського соціокультурного і лінгвістичного контексту. Сучасні дослідження підтверджують, що традиційні моделі генерації паролів ігнорують вплив зовнішніх семантичних факторів (соціальних трендів, популярної лексики), що суттєво знижує їхню адаптивність [5]. Запропонований підхід субслівної токенізації враховуватиме унікальні регіональні аспекти, такі як транслітерацію кириличних слів, крос-розкладне введення та використання специфічних абревіатур.

Очікується, що це дозволить підвищити точність та швидкість підбору паролів. Даний фактор є важливим, оскільки грунтуючись на аналізі витоків паролів [2], варто відзначити, що за останні 15 років середня довжина паролів зросла з 8 до 10 символів, а їхня середня інформаційна ентропія — із 40 до 50 біт. Оцінка ентропії $H$ здійснюється за класичною формулою Шеннона [3]:

$$E = -\sum_{i=1}^{n} P_i log_2 P_i \quad (1)$$

де $P_i$ — імовірність появи $i$-го символу чи токена з доступного простору.

Загалом існує сукупність факторів, що дозволяє прослідкувати ускладнення паролів, проте залишається конструктивна слабкість через наявність передбачуваної інформації.

Для розв'язання поставленої задачі пропонується використовувати алгоритмічну модель генерації та скорингу токенів. Вона опрацьовує неструктурований масив вхідних OSINT-даних (імена, дати, локації, специфічний сленг) та виконує їх багаторівневу обробку. Процес включає вилучення базових сутностей, їх розбиття на субслівні одиниці (склади,

біграми), застосування алгоритмів транслітерації та додаткових мутацій. З метою формування оптимального словника, кожен токен отримує скорингову оцінку від 0 до 100 балів (табл. 1).

Таблиця 1

Критерії скорингової оцінки субслівної токенізації

| Критерій | Опис параметрів токена | Макс. бал |
|---|---|---|
| Пріоритет джерела | Прямі збіги (імена, дати), ініціали | 40 |
| Локалізація | Транслітерація, крос-розкладне введення (kbswap), N-грами | 30 |
| Частотність | Мультипликатор ваги за частоту появи токена у вхідних OSINT-даних | 20 |
| Довжина | Оптимальний розмір токена для формування паролю (найвища вага 3-6 символів) | 10 |

Отже, ускладнення паролів не усуває людського фактору, користувачі продовжують спиратися на власний цифровий слід та локальний контекст. Виокремлення граматичних особливостей формування парольних фраз і поєднання їх із субслівною токенізацією та OSINT-розвідкою дозволить створювати високорелевантні цільові словники паролів. Запропонована скорингова модель генерації токенів може бути ефективним інструментом для тестування на проникнення та комплексного аудиту безпеки інформаційних систем.

1. Слатвінська В., Бевза В., Вплив збою CrowdStrike на мега-витік паролів: чи є зв’язок? Ч. 1. Вісник Хмельницького національного університету. Технічні науки. – 2024. – № 4 (339). – С. 332-338. DOI: 10.31891/2307-5732-2024-339-4-52.
2. Rodrigues G.A.P. et al., From RockYou to RockYou2024: Analyzing Password Patterns Across Generations, Their Use in Industrial Systems and Vulnerability to Password Guessing Attacks. Journal of Internet Services and Applications. – 2025. – Vol. 16, No. 1. DOI: 10.5753/jisa.2025.5041.
3. Собина В.О., Тарадуда Д.В., Демент М.О., Захист інформації відомчої інформаційно-телекомунікаційної мережі за допомогою парольної системи. Проблеми надзвичайних ситуацій. – Харків: НУЦЗ України, 2021. – № 2 (34).
4. Різак М., Котик О., Використання OSINT для захисту персональних даних. The 14th International Scientific Conference «ITSec». – Тернопіль, 2025. – С. 165-167.
5. Yang X. et al., KAPG: Adaptive Password Guessing via Knowledge-Augmented Generation. – 2025.