УДК 004.056.52

ДОСЛІДЖЕННЯ АЛГОРИТМІВ МОРФОЛОГІЧНОГО АНАЛІЗУ У DLP-СИСТЕМАХ ДЛЯ ЗАПОБІГАННЯ ВИТОКУ ІНФОРМАЦІЇ

DOI 10 36994 / 2788- 5518- 2022-02-04-16

Близнюк А.В. *Державний університет інтелектуальних технологій і зв'язку, Київ,Україна. blizniykartem@gmail.com*

Анотація: *У статті досліджується використання морфологічного методу аналізу текстової інформації в системах протидії витоку інформації. Дається визначення поняттю морфологічного аналізу в DLP-система та аналізується актуальність проблеми забезпечення інформаційної безпеки в сучасному світі з постійним розвитком цифровізації. Досліджуються проблеми та особливості морфологічного аналізу враховуючи особливості різних мов. У зв'язку з особливостями української та російської мови, які найчастіше використовуються в нашому інформаційному просторі, і наявністю великої кількості слів виняткiв в них, здійснення морфологічного аналізу може бути ускладнене. Для подолання цих труднощів існує можливість вибору методу морфологічного аналізу. У статті алгоритми морфологічного аналізу текстової інформації порівнюються за основними характеристиками, зокрема точність визначення слова при морфологічному аналізі, можливість алгоритму до самонавчання, а також часові затрати на впровадження і налаштування того чи іншого методу в робочу діяльність компанії. Описані переваги та недоліки різних підходів до реалізації, наведені структурні особливості алгоритмів стемінгу - який полягає у виокремленні основної частини слова без закінчення та суфіксу, а не тільки у виділенні кореня, їх приклади та програмні рішення, які доступні на ринку.*

Ключові слова: *DLP-система, морфологічний аналіз, запобігання витоку інформації, алгоритми стемінгу, інформаційна безпека.*

RESEARCH OF MORPHOLOGICAL ANALYSIS ALGORITHMS IN DLP-SYSTEMS TO PREVENT INFORMATION LEAKAGE

Artem Blyzniuk . *State University of Intellectual Technologies and Communication, Kyiv,Ukraine. blizniykartem@gmail.com*

Abstract: *The article investigates the use of morphological method of text information analysis in information leakage control systems. The concept of morphological analysis in DLP-systems is defined and the relevance of the problem of information security in the modern world with the constant development of digitalization is analyzed. The problems and features of morphological analysis are investigated taking into account the peculiarities of different languages. Due to the peculiarities of the Ukrainian and Russian languages, which are most often used in our information space, and the presence of a large number of exception words in them, the implementation of morphological analysis can be complicated. To overcome these difficulties, it is possible to choose a method of morphological analysis. In this article, the algorithms of morphological analysis of text information are compared by the main characteristics, in particular, the accuracy of word detection in morphological analysis, the ability of the algorithm to self-learn, as well as the time required to implement and configure a particular method in the company's work. The advantages and disadvantages of different approaches to implementation are described, the structural features of stemming algorithms - which consists in isolating the main part of the word without ending and suffix, and not only in isolating the root, their examples and software solutions that are available on the market are given.*

Key words: *DLP-system, morphological analysis, information leakage prevention, stemming algorithms,*
information security.

Вступ

Розвиток інформатизації та цифровізації сучасного постіндустріального суспільства, де переважає інноваційний сектор з високопродуктивною економічною системою, все частіше стикається з новою глобальною проблемою, а саме – проблемою інформаційної

безпеки. Левова частка інтересів підприємства визначається інформаційним середовищем в якому воно знаходиться. Тому коли середовище змінюється це також має вплив на інтереси підприємства, зокрема такі зміни можуть становити реальну загрозу майбутній діяльності підприємства.

Необхідно зазначити, що в довідниковій літературі відсутній єдиний погляд на поняття «інформаційна безпека підприємства», що є вкрай актуальним питанням на етапі розвитку і вдосконалення сфери ІТ, яке супроводжується введенням інформаційних систем у всі сфери діяльності людини, постійною взаємодією підприємств на теренах саме інформаційного простору.

Масштаби дистанційної та віддаленої роботи дають змогу як кіберзлочинцям, так і інсайдерам набагато більше можливостей для крадіжки інформації. Великі ризики можуть бути пов'язані з використанням так званих "тіньових ІТ" (Shadow IT), тобто інформаційних сервісів, що реалізовані на зовнішніх, не корпоративних ресурсах, власники яких не несуть ніякої відповідальності за інформацію, що обробляється.

Використання великих обсягів конфіденційної та "чутливої" інформації, ставить перед відділом забезпечення інформаційної безпеки чи кібербезпеки комплексну задачу для забезпечення належного функціонування компанії, запобігання витоків інформації, розмежування доступів відповідно до посадових обов'язків. Все це значною мірою впливає на впровадження та апробацію такої системи, як DLP.

Виконання досліджень

Обійтись тільки адміністративними мірами для запобігання ІТ-інцидентів, пов’язаних із витоком інформації, на сьогоднішній день вже не вдається. Справа тут не тільки в збільшенні обсягів та різноманіття цінної інформації, а в принциповій відкритості інформаційних систем фінансових установ.

На рис. 1 зображено графік з порівнянням частки витоку інформації за різними типами даних. Лідируючим типом стабільно є персональні дані, близько восьмидесяти відсотків пов'язані з викраденням або ненавмисним розкриттям персональних даних.

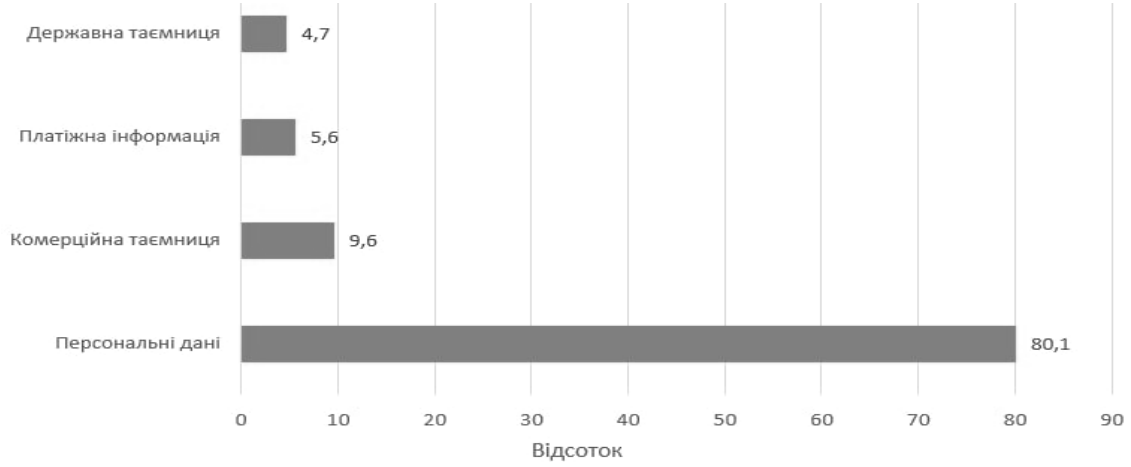

Рис. 1. Графік порівняння витоку інформації за різними типами даних

Основним каналом витоку інформації залишається мережа та електронна пошта, що також вказує на необхідність захисту конфіденційної та персональної інформації, зокрема з використанням DLP-систем.

Якщо порівнювати сфери, де найчастіше трапляються витоки інформації, то перше місце стабільно за останні роки займає індустрія високих технологій, як у західних країнах так і у країнах колишнього СНД.

Через досить високу інформатизацію та діджиталізацію систем у країнах Європи та США на другому місці посідає медичний сектор. Персональні дані пацієнтів, інформація про страхові поліси є конфіденційною, але часто не достатньо захищеною.

Також слід зауважити, що значна частина, а саме 51,2%, порушень та випадків витоку інформації була скоєна персоналом та керівниками компаній. Тому в умовах пандемії та віддаленої роботи працівників компанії та підприємств питання захисту інформації стає ще більше актуальним.

Основою виявлення витоків інформації в сучасних DLP-системах є морфологічний аналіз текстів. Головною перевагою цієї технології є універсальність алгоритмів аналізу, які дозволяють проводити оцінку як повідомлень в різних месенджерах, так і тексту електронних документів. Також перевагами використання цього методу є можливість працювати з вмістом, здатність до навчання лінгвістичного алгоритму, масштабованість і простота налаштування. До недоліків можна віднести залежність від мови, яка використовується, і необхідності застосування імовірнісного підходу.

Морфологічний аналіз являє собою процес визначення граматичного значення словоформи і виділення її основи, або, інакше кажучи, виділення ключових слів у текстовому потоці. Кожен алгоритм морфологічного аналізу складається з двох основних складових - декларативної і процедурної. При цьому декларативна складова являє собою таблиці структурованих даних, необхідних для аналізу, а процедурна компонента містить безпосередньо алгоритми аналізу і допоміжні процедури.

У зв'язку з особливостями української та російської мови і наявністю великої кількості слів винятків в них, здійснення морфологічного аналізу може бути ускладнене. Для подолання цих труднощів існує можливість вибору методу морфологічного аналізу, серед яких найбільш відомими є три основних.

Першим методом є складання морфологічного словника для конкретного підприємства вручну, враховуючи всі ключові слова, які здатні будь-яким чином вказати на витік інформації. Вищеописаний спосіб слід використовувати, коли в наявності відносно невелика множина ключових слів, аналізуючи корінь цих слів. Якщо інформаційна система підприємства (організації) складна, містить велику кількість різноманітних ресурсів, то використання даного методу буде ускладнене, особливо на початкових етапах.

Особливістю другого методу є використання алгоритму стемінгу (англ. stemming), який полягає у виокремленні основної частини слова без закінчення та суфіксу, а не тільки у виділенні кореня. Трапляється, що результати стемінгу можуть бути схожі на визначення кореня слова, проте алгоритм цього методу використовує інші правила та принципи. Тому слово після обробки алгоритмом стемінгу (стематизації) може відрізнятися від морфологічного кореня слова.

Стемінг досить часто використовується в лінгвістичній морфології та для інформаційного пошуку. Багато пошукових систем застосовують цей алгоритм для встановлення синонімічних зв’язків, якщо в них збігаються форми після стемінгу. Для української та російської мови найбільш популярними алгоритмами стемінгу є:

- алгоритм стемінгу Портера;
- Stemka;
- Mystem;
- визначення слова по його суфіксу і афіксу.

Алгоритм стемінгу Портера, інакше званий "Snowball", був розроблений в 1979 році спочатку для англійської мови (тепер фактично стандартний алгоритм стематизації для англійської мови), згодом адаптований під аналіз на основі слов'янських мов. При використанні даного алгоритму стемінг відбувається через ряд правил на основі безлічі існуючих суфіксів і закінчень, які відсікаються. При цьому цей алгоритм не використовує безпосередньо базу основ слів. Для будь-якої мови спочатку необхідно визначити який клас закінчень, має слово, що перевіряється, за допомогою спеціальних правил. Для української мови закінчення матимуть такий вигляд:

- якщо слово іменник воно матиме одне із закінчень а|ев|ов|е|ями|ами|еи|и|ей|ой|ий|й|иям|ям|ием|ем|ам|ом|о|у|ах|иях|ях|ы|ь|ию|ью|ю|ия|ья|я|і|ові|ї|ею|єю|ою|є|еві|ем|єм|ів|їв|\'ю;
- якщо слово відноситься до дієслів, то матиме одне із закінчень сь|ся|ив|ать|ять|у|ю|ав|али|учи|ячи|вши|ши|е|ме|ати|яти|є;
- для дієприкметників закінчення відповідно ий|ого|ому|им|ім|а|ій|у|ою| ій|і|их|йми|их;

- для прикметників закінчення ими|ій|ий|а|е|ова|ове|ів|є|їй|єє|еє|я|ім|ем| им|ім|их|іх|ою|йми|іми|у|ю|ого|ому|ої;

- для дієприслівників закінчення ив|ивши|ившись.

У тому числі перевіряються словотвірні частини, а також частини слова після першої голосної літери, або кінець слова, що не містить голосних.

Після зчитування слова всі літери конвертуються в нижній регістр, після чого перевіряється наявність апострофа та літери "ґ" у слові та видаляються при наявності. Після цього аналізуються закінчення слова у відповідності до частин мови та видаляються при їх наявності. Якщо слово в закінченні має голосну літеру – вона також видаляється. В кінці аналізуються суфікси найвищого ступеню порівняння прикметників, подвоєння, чи є у слові м'який знак та видаляються при їх наявності, а у подвоєнні – одна із літер.

В результаті виконання алгоритму створюється необхідна для впізнання частина слова. Основною перевагою алгоритму стемінгу Портера є відсутність словників основ, що істотно підвищує швидкодію. Часові витрати на роботу алгоритму наведені на рисунку 2.2. Також слід зазначити, що програмна реалізація даного підходу класифікатора не потребує великих обсягів пам'яті на дисковому просторі.

До негативних властивостей даного алгоритму можна віднести можливість втрати частини інформації з аналізованого слова. Крім того, вразливістю даного алгоритму є можливість помилки з боку оператора, що задає правила перевірки.

"Stemka" - російсько-український ідентифікатор морфології, який був створений за допомогою спеціально розробленого морфологічного модуля. Алгоритм заснований на ймовірнісній моделі. Складається масив даних, який включає в себе пари - дві останні букви основи та суфікс, таким чином, виходять моделі різних слів. Після цього визначається ймовірність появи моделей в тексті, і при ймовірності 1/10000 модель відсікається. Результат являє собою таблицю переходів кінцевого автомату, за якими сканується слово.

Для розглянутих алгоритмів "Snowball" і "Stemka" в процесі налагодження було сформульовано правило, яке передбачає наявність хоча б однієї голосної букви в основі.

Mystem був розроблений в 1998 Іллею Сегаловіч. Модель будується в вигляді лісу інвертованих префіксних дерев суфіксів і інвертованого префіксного дерева для основ, для цього використовується словник з перерахуванням всіх граматичних форм слова.

Алгоритм Mystem, функціонування якого полягає в наступному. Чергове аналізоване слово піддається поділу на основу і суфікс. Такий поділ проводиться програмою на основі вже наявного дерева суфіксів. Далі відбувається зіставлення вихідної основи з уже наявними в словнику (блок 2) для знаходження відповідностей. Якщо відповідність знайдено, алгоритм закінчує роботу, і результатом є гіпотеза для словникового слова. Якщо ж відповідність знайти не вдалося, то алгоритм продовжує роботу з пошуку потрібної гіпотези. Для цього відбувається генерація гіпотетичної моделі слова, що базується на основі слова, суфікс і найближчій основі з наявного словника (блок 3).

Після цього отримана модель знову звіряється зі словником. При позитивному результаті алгоритм закінчує роботу. Якщо результат знову негативний і збіг не знайдено, то алгоритм продовжує виробляти вищевказані дії, поступово зменшуючи основу на одну букву (блок 4) до тих пір, поки основа не буде знайдена, або поки кількість букв не скоротиться до двох.

У цьому випадку відбувається ранжування всіх отриманих в ході роботи алгоритму гіпотез (блок 5) по продуктивності і відсікання менш продуктивних. В результаті на виході алгоритму виходить набір гіпотез для неіснуючого в наявному словнику слова. Перевагами даного алгоритму є простота реалізації і словників, а також можливість визначення форм слова, яких не було в словнику. До недоліків цього алгоритму можна віднести те, що хоча алгоритм і працює з українською мовою, проте більш орієнтований на російську.

Четвертим методом є визначення слова по його суфіксу і афіксу, та приведення слова до його початкової форми. Даний спосіб є найбільш раціональним завдяки особливостям слов'янських мов, однак для підвищення якості роботи алгоритму необхідно додати якомога більше слів-винятків. Таким чином, використання одного з перерахованих вище

алгоритмів морфологічного аналізу дозволяє розпізнати конфіденційну інформацію в потоці інформації, що перехоплюється.

Існує кілька критеріїв, що впливають на вибір алгоритму морфологічного аналізу для DLP-систем. До них відносять:

- точність визначення слова при морфологічному аналізі;
- здатність до навчання;
- часові витрати на налаштування системи.

Згідно з дослідженнями точність визначення повинна бути не нижче 95 97%. Здатність до навчання дозволяє зімітувати на етапі налаштування системи небезпечні ситуації, тим самим дозволивши системі самостійно адаптуватися під необхідні параметри, що істотно скорочує час введення системи в експлуатацію.

Для співробітника, що не має навичок аналітика, створення словника для морфологічного аналізу може зайняти досить багато часу, що істотно позначиться на швидкості введення системи в експлуатацію.

Порівняння алгоритмів морфологічного аналізу, що використовуються в DLP-системах, наведені в табл. 1.

З огляду на дані, що представлені в таблиці, найбільш оптимальним алгоритмом морфологічного аналізу для використання в DLP-системах є алгоритм Mystem. Незважаючи на тривалість налаштування, алгоритм має найвищу з перерахованих алгоритмів ймовірність правильного визначення слова і має важливу здатність до навчання.

Таблиця1.

Порівняння алгоритмів морфологічного аналізу

Назва алгоритму	Точність визначення слова при морфологічному аналізі	Можливість до навчання	Часові затрати на налаштування
Складання морфологічного словника	80–85%	-	1–2 дні
Алгоритм стемінгу Портера;	85–90%	-	3–5 годин
Визначення по суфіксу і афіксу	90–96%	-	1–2 дні
Mystem	92–97%	+	2–3 дні
Stemka	87–95%	+	1–1,5 дні

Висновки

DLP продукт виконує функціонал аналіз та моніторингу вихідного трафіку, аналізу інформаційних потоків за декількома технологіями, глибокого аналізу змісту інформації, автоматичний захист конфіденційних даних у кінцевих інформаційних ресурсах, на рівні шлюзів, що передають дані, і в системах, що зберігають дані статично.

Було проведено дослідження методи, які використовує DLP-система для морфологічного аналізу текстових даних в інформаційних системах, проаналізовані різні алгоритмічні підходи для розв'язання труднощів, що пов'язані з особливостями мов.

Визначено, що найбільш оптимальним алгоритмом морфологічного аналізу для використання в DLP-системах є алгоритм Mystem. Незважаючи на тривалість налаштування, алгоритм має найвищу з перерахованих алгоритмів ймовірність правильного визначення слова і має важливу здатність до навчання.

Література

1. Метод стемінгу україномовних текстів для класифікації документів на базі алгоритму Портера. URL: https://clck.ru/UzvxN.
2. Використання адаптованої DLP-системи для блокування витоків інформації. URL: https://cyberleninka.ru/article/n/ispolzovanie-adaptirovannoy-dlp-sistemy-dlya-blokirovaniya-utechek-informatsii/viewer.
3. DLP у документах. URL: https://mirobase.com/spravka/64/.
4. Технології захисту конфіденційної інформації від внутрішніх загроз. URL: http://dspace.nbuv.gov.ua/bitstream/handle/123456789/50925/8_s_78-88.pdf?sequence=1.
5. Morphological box of the DLP, 2022. URL: https://www.researchgate.net/figure/Morphological-box-of-the-DLP_fig1_330727964.

References

1. Metod stemingu ukraïnomovnih tekstiv dlya klasifikacií dokumentiv na bazi algoritmu Portera. [A method of stemming Ukrainian-language texts for document classification based on Porter's algorithm]. URL: https://clck.ru/UzvxN.
2. Vikoristannya adaptovanoï DLP-sistemi dlya blokuvannya vitokiv informacií. [Using an adapted DLP system to block information leaks]. URL: https://cyberleninka.ru/article/n/ispolzovanie-adaptirovannoy-dlp-sistemy-dlya-blokirovaniya-utechek-informatsii/viewer.
3. DLP u dokumentah. [DLP in documents]. URL: https://mirobase.com/spravka/64/.
4. Tehnologiï zahistu konfidencijnoï informacií vid vnutrishnih zagroz. [Ttechnologies for protecting confidential information from internal threats]. URL: http://dspace.nbuv.gov.ua/bitstream/handle/123456789/50925/8_s_78-88.pdf?sequence=1.
5. Morphological box of the DLP. URL: https://www.researchgate.net/figure/Morphological-box-of-the-DLP_fig1_330727964.