ТЕХНОЛОГІЇ ІДЕНТИФІКАЦІЇ І АНАЛІЗУ КОНФІДЕНЦІЙНИХ ДАНИХ В DLP-СИСТЕМАХ

Вацлавик Олег, Явин Христина

Львівський державний університет безпеки життєдіяльності, Львів, Україна

Summary. The prevalence of information security (Gartner claims that about a third of companies already use DLP) removes only one part of the problem-random leakage-do not affect malicious behavior. The question here is faster in the perception of DLP-systems as a software capable of independently, without the efforts of the security services of information, to deal with leaks, which is fundamentally wrong.

Keywords: DLP systems, digital imprints, artificial Intelligence, linguistic method.

DLP-системах зазвичай використовуються три методи ідентифікації: імовірнісний, детерміністський і комбінований. Системи, засновані на першому методі, здебільшого використовують лінгвістичний аналіз контента і «цифрові відбитки» даних. Такі системи прості в реалізації, але недостатньо ефективні і характеризуються високим рівнем помилкових спрацьовувань. Системи, що використовують детермінований підхід (мітки файлів), дуже надійні, але їм не вистачає гнучкості. Комбінований підхід поєднує обидва методи з аудитом середовища зберігання і обробки даних, що дає можливість досягти оптимального вирішення проблеми захисту конфіденційності інформації.

У системах DLP застосовуються складні механізми аналізу: порівняння по шаблонах з використанням словників і регулярних виразів, лінгвістичний і контекстний аналіз, цифрові відбитки. Словники і шаблони зручно застосовувати в конкретних областях, наприклад, для контролю номерів кредитних карт і інших персональних даних.

У лінгвістичному і контекстному аналізі використовуються морфологія і статистичні моделі, враховується контекст, характер відправника і одержувача інформації. Цей метод хороший для динамічних даних. Цифрові відбитки (аналогічні сигнатурам в антивірусних продуктах) підходять для контролю статичних даних, наприклад, для захисту інтелектуальної власності.

Через DLP-систему проходять всі інформаційні потоки підприємства, і система повинна визначати, чи відноситься інформація, що передається до тієї, що захищається. Для цього використовують наступні технології:

− Сигнатури – пошук в потоці даних "заборонених" слів, послідовності символів ("стоп-слів");

− Лінгвістичні методи – працюють із словоформами, аналізують весь текст (наприклад, визначення частоти зустрічальності термінів);

− Цифрові відбитки – хеш-функції зразків конфіденційних документів;

− Регулярні вирази – дозволяють знаходити збіги за формою даних (а не за самими даними), типу номерів кредитних карток;

− Мітки – установка на файли, що містять конфіденційну інформацію, спеціальних «міток»;

− Штучний інтелект – самонавчальний алгоритм аналізу даних "Vector Machine Learning".

Методом аналізу є пошук в потоці даних деякої послідовності символів («стоп-слів»). У переважній більшості випадків сигнатурні системи налаштовані на пошук декількох слів і частоту зустрічальності термінів.

Метод аналізу масок є розширенням функціонала пошуку сигнатур і є пошуком такого змісту, який неможливо точно вказати в базі "стоп-слів", але можна вказати його елемент або структуру. До такої інформації слід віднести будь-які коди, які характеризують персону або підприємство: ІНН, номери рахунків документів і так далі. Шукати їх за допомогою сигнатур неможливо.

Лінгвістичний метод аналізу тексту несе на собі характеристику всього класу методів аналізу вмісту. З погляду класифікації хеш-аналіз, аналіз сигнатур, аналіз масок – є "контентною фільтрацією", тобто фільтрацією трафіку на основі аналізу вмісту.

Технологія лінгвістичного аналізу дозволяє автоматично визначати тематику і ступінь конфіденційності аналізованого фрагмента інформації на підставі термінів, що зустрічаються в ньому, і їх поєднань. Лінгвістичний аналіз виконується на основі заздалегідь створеної бази контентної фільтрації (БКФ).

База контентної фільтрації – це база даних, яка представляє собою виділений на основі імовірнісних і математичних методів ієрархічно організований список (дерево) категорій з довільною кількістю вкладених рівнів, і що містить слова і вирази, наявність яких в документі дозволяє визначити тематику і ступінь конфіденційності інформації.

БКФ не тільки описує категорії інформації, яка циркулює в компанії, але і враховує різні атрибути її конфіденційності, в т.ч. специфіку діяльності компанії, її вимоги до безпеки.

Результатами проведення лінгвістичного аналізу тексту автоматично привласнюються ті або інші категорії, відповідні його тематиці і змісту. У аналізованій інформації можуть зустрітися терміни (слова і словосполучення) з різних категорій, тому вона може бути віднесена до однієї або декількох категорій БКФ.

База контентної фільтрації і точність детектування конфіденційної інформації

Надійність і точність ідентифікації конфіденційних даних в корпоративних інформаційних потоках за допомогою технології лінгвістичного аналізу залежать від бази контентної фільтрації, на основі якої здійснюється аналіз.

Тому важливо створити базу, яка забезпечить надійні результати фільтрації інформації за категоріями. Основним методом лінгвістичного аналізу за допомогою БКФ є пошук в аналізованому фрагменті інформації слів і словосполучень, що описують конфіденційні дані і структурованих за категоріями.

Література

1. В.Мирошниченко, Д.Ходоров, Е.Щеглова DLP-решения: Как помешать торговле корпоративными секретами [Електронний ресурс] // Інвестгазета №14 10.05.2011 – Режим доступу: https://investgazeta.delo.ua/praktika/dlpreshenija-kak-pomeshat-tor-273582/
2. DLP-решения - информационная безопасность – Режим доступу: http://securitysoftline.ru
3. 3. А. Прозоров ALL ABOUT DLP [Електронний ресурс] – Режим доступу: http ://bis-expert.ru/blog/2560/51911
4. Загальні положення щодо захисту інформації в комп'ютерних системах від несанкціонованого доступу: НД ТЗІ 1.1–002–99.–К.. ДСТСЗІ СБ України, 1999. – 16 с.