УДК

ПРЕДЛОЖЕНИЕ ПО МОДИФИКАЦИИ АЛГОРИТМА СЕМАНТИЧЕСКОГО АНАЛИЗА В ИНФОРМАЦИОННО-АНАЛИТИЧЕСКИХ СИСТЕМАХ БЕЗОПАСНОСТИ

Кисарина П.А. [1], Полянский Д.А.[1]

[1] ИБ ИИТиР ВлГУ, Владимир

Данная статья дает представление о некоторых основных недостатках внедренных бизнесом DLP-систем в международной практике. Проанализированы основные структурные недостатки данных систем, выявлены факторы, коррелирующие с ними. Сформирована экспериментальная установка, позволяющая оценить влияние изменения данных факторов на ошибки 1 и 2 рода в работе систем. Приведены результаты проведенных исследований с применением алгоритмов, включающих влияние выявленных факторов, в бизнес-системах различной направленности.
Ключевые слова: DLP-система, утечки информации, потоки данных, семантические алгоритмы, защита информации.

PROPOSAL FOR THE MODIFICATION OF THE SEMANTIC ANALYSIS ALGORITHM IN INFORMATION AND ANALYTICAL SECURITY SYSTEMS

Kisarina P.A.[1], Polyansky D.A.[1]

[1]IITiR VlSU, Vladimir

This article gives an idea of some of the major shortcomings of DLP systems implemented by business in international practice. The main structural deficiencies of these systems are analyzed, the factors correlating with them are revealed. An experimental setup has been formed, which makes it possible to evaluate the effect of changes in these factors on errors 1 and 2 of the order of the systems. The results of the studies carried out using algorithms, including the influence of identified factors, in business systems of various kinds are presented.
Keywords: DLP-system, information leaks, data streams, semantic algorithms, information protection.

Введение

Формализованный подход к ведению бизнеса позволяет четко структурировать организационные действия в виде документации, часть из которой предполагает отображение потенциально возможных действий организации. На основании массивов корпоративных данных можно сформировать совокупности показателей, которые затем, преобразовав в модели оценки эффективности различного рода, использовать сравнимо с инсайдерской аудиторской информацией.

Для противодействия подобным угрозам информационной безопасности во многих организациях внедряются DLP-системы, позволяющие анализировать массивы данных корпоративного общения. Одним из разделов анализа в таких системах является семантико-синтаксический анализ данных.

При анализе ранее проведенных исследований таких систем [1], [5] прослеживается тенденция – внедрение систем может использоваться как средство защиты информации и предотвращения ее утечек, однако работа таких систем несовершенна, прежде всего, потому, что работа автоматизированного анализатора формализована и не может включать в себя анализ данных в контексте, в результате чего при работе такой системы наблюдаются ошибки первого и второго рода (а именно ошибки первого рода - пропуск утечки конфиденциальной информации, а также высокая вероятность ложных срабатываний – ошибок 2 рода).

Исходя из вышеуказанных исследований становится очевидно, что работа семантического анализатора в рамках DLP-системы нуждается в добавлении в нее некоторой категоризации и индексации данных с их коррекцией на контекст, которое позволило бы уменьшить ошибки ложных срабатываний системы, а также количество ошибок первого рода.

Цель исследования

В рамках данного исследования предлагается экспериментальное исследование алгоритма, усовершенствование которого заключается:

- в возможности коррекции индексации данных в зависимости от их расположения по тексту;
- в категоризации исследуемых массивов данных по структурированным словарям со специфично подобранной сигнатурной базой

Таким образом, сформирована цель исследования - повышение качества проведения анализа служебной переписки организации при оценке информационной безопасности.

Формулировка научной гипотезы звучит следующим образом: предположим, что заданная частота встречаемости лексем из подготовленных организацией сигнатурных баз будет более качественно свидетельствовать об определенном уровне её информационной защищенности при зависимости от категоризации сигнатурных баз и коррекции на их расположение по тексту.

Формулировка задачи исследования

Стандартная схема обработки текстов, используемых промышленными системами, состоит из этапов [6], [5]:

1. Токенизация;
2. Орфографическая коррекция;
3. Стемминг;

4. Задачи семантического анализа в соответствии с ожидаемыми результатами

Задача, решение которой предлагается получить в работе, определяется как оптимизационная, а именно: необходимость максимизировать количество найденных фрагментов, информативно для исследователя пересекающих периметр заданных сигнатурно ограничений при минимизации ошибок ложных срабатываний.

Подход к решению

Для целей коррекции на контекст слова из сигнатурных баз (или далее по тексту – словарей) были «взвешены» исходя из следующего предположения (формула 1):

$$\text{Вес слова} = \frac{\text{Кол — во контекстов, где употребление данного слова относилось к искомому результату}}{\text{Кол — во контекстов, где в ходе анализа нашлось слово}} * 10 \quad (1)$$

После формирования сигнатурных баз, было проведено их стеммирование или приведение слов к базовой словоформе, позволяющее снизить количество ошибок второго рода (пропусков информации).

После первоначального формирования словаря сформированной группой и весов, соответствующих этим «стоп-словам», прохождение по объекту исследования начиналось заново, результаты соотносились с экспертной оценкой, полученной в результате влияния на оценку человеческого фактора.

Для каждого стоп-слова определялся процент «несоответствия», когда слово в автоматизированной экспериментальной системе свидетельствовало о присутствии в контексте тревожных признаков, однако экспертно это не подтверждалось. Процент определялся относительно количества всех контекстов, взятых в эксперимент. Далее слово, выдававшее наибольший процент погрешности, убиралось из словаря.

Элементы алгоритма анализатора можно следующим образом:

1) Рабочий массив данных подается на анализ, вводится параметр окрестности = a

2) В тексте ищется первое слово, совпадающее со словом из словаря, после чего анализируется фрагмент текста с шинглом, определяющимся согласно следующей формуле 2:

$$\text{Шингл} = \{\text{Искомое слово} - a \text{ слов} | \text{Искомое слово} + a \text{ слов}\}, \quad (2)$$

причем во всех случаях, когда Искомое слово-a слов < 0, шингл определяется как (формула 3):

$$\text{Шингл} = \{0|\text{Искомое слово} + a \text{ слов}\}, \quad (3)$$

3) Выделенному в п.2 куску текста присваивается определенный вес в соответствии с весом, указанным в словаре, причем, если во фрагменте найдено 2 и более слов, их вес суммируется, подозрительный фрагмент с данными о «весе подозрения» попадает на вывод

4) В рамках найденного «слепка» или шингла суммарный вес корректируется на удаленность слов между собой исходя из линейной зависимости, т.е. по формуле 4:

$$\text{Вес шингла} = \sum_{\text{все слова из шингла}} \text{вес слова в шингле} - \sum_{\text{все слова из шингла}} \text{кол} - \text{во слов между индексированными словами} \quad (4)$$

5) Далее следует переход к следующему фрагменту аналогично п.2, т.е. ищется следующее слово, совпавшее со словарем, кусок текста фрагментируется и проверяется на вес (п.3)

6) В результате анализа всего текста фрагменты с весами ранжируются по убыванию веса и идут на вывод программы, первыми выводятся фрагменты с наибольшим весом, причем, возможно задать ограничение (параметр = b), ограничивающий вывод фрагментов со слабым весом (не выводятся фрагменты с весом<=b). При выставлении пустым поле параметра, можно увидеть все совпадения по тексту.

7) Для снижения ошибок 2 рода также в алгоритм работы была включена возможность специфицирования сигнатурных баз добавлением новых индикативных слов с экспертно выставленным весом.

Ход эксперимента

Рассмотрению подвергались служебные переписки в организациях, сферой деятельности которых являлись:

• информационные технологии,

• юридические услуги,

• банковская деятельность.

В качестве метода измерения факторов по всем факторам выбран прямой метод измерения, т.к. искомую величину устанавливают непосредственно из опыта.

К средствам измерений в представленной работе относится система, представляющая собой ПК с установленным на нем разработанным в рамках исследования ПО.

Эксперимент был проведен в двух фазах – активной, когда были сформированы сигнатурные базы, и пассивной – когда сравнивались результаты работы программы до и после усовершенствования.

Абстрактный пример работы разработанной программы

Исходный фрагмент:

X: пришли пожалуйста с среды банка результат …

Y: 1 - результат

Y: Добрый день, что можно оттуда почистить с диска (C)?

Z: все можно чистить

Y: Коллеги, а пароль ... кто-нибудь менял? т.к. текущий прописанный в параметрах маршрута не подходит (не могу залогиниться в лм с ним)

*Z: test123** можно использовать*

Y: спасибо

Z: Коллеги, нужна помощь, - закинуть пару файлов в облако

Z: Спасибо, сделали уже

Y: Коллеги, сейчас будет экстренно перезапущена среда (оповещение из банка)

Z: test1518

Z: пароль от теста

Y: коллеги, можно со среды что-то удалить? места нет на диске

Z: наверное можно удаляй

Выборка слов для формирования словарей в анализируемом массиве данных:

X: пришли пожалуйста с среды банка результат …

Y: 1 - результат

Y: Добрый день, что можно оттуда почистить с диска (C)?

Z: все можно чистить

Y: Коллеги, а пароль ... кто-нибудь менял? т.к. текущий прописанный в параметрах маршрута не подходит (не могу залогиниться в лм с ним)

*Z: test123** можно использовать*

Y: спасибо

Z: Коллеги, нужна помощь, - закинуть пару файлов в облако

Z: Спасибо, сделали уже

Y: Коллеги, сейчас будет экстренно перезапущена среда (оповещение из банка)

Z: test1518

Z: пароль от теста

Y: коллеги, можно со среды что-то удалить? места нет на диске

Z: наверное можно удаляй

Вывод результатов:

Фрагмент с окрестным контекстом	Вес
закинуть пару файлов в облако Z: Спасибо, сделали	8
Y: Коллеги, а пароль ... кто-нибудь менял?	7
Z: test1518 Z: пароль от теста Y: коллеги	7
X: пришли пожалуйста с среды банка результат … Y: 1 - результат	5
среды что-то удалить? места нет на диске Z: наверное можно удаляй	4
можно оттуда почистить с диска (C)? Z: все можно чистить	4

Обсуждение результатов

Из результатов видно, что предлагаемые методы повышения эффективности работы DLP-систем в части их семантического анализа источников служебных данных ощутимо влияют на ошибки 2 рода, в тоже время незначительно корректируя анализ в вопросе совершения ошибок 1 рода.

Вероятно, повышение качества анализа в результате снижения количества ошибок 2 рода достигается, прежде всего, гибкостью сигнатурных баз (а именно возможностью добавления специфических для каждого вида организаций сигнатурных слов в словари, т.е. адаптивностью исследуемых текстов к специфике работы организации наряду с базовой экспертной категоризацией), а также корректированием весов шинглов на контекст, т.е. удаленность слов-индикаторов друг от друга.

Гипотеза «предположим, что заданная частота встречаемости лексем из подготовленных организацией сигнатурных баз будет более качественно свидетельствовать об определенном уровне её информационной защищенности при ее зависимости от категоризации и контекста» - подтверждена относительно ошибок 2 рода и не подтверждена относительно ошибок 1 рода.

Список литературы

1. Литвиненко Е. Последние тенденции эволюции DLP-систем // "Information Security/ Информационная безопасность" #2, 2018
URL: http://itsec.ru/articles2/pronsol/poslednie-tendencii-evolucii-dlpsistem
2. Агурьянов И. Обходим DLP-системы //SecurityLab – 2013
URL: https://www.securitylab.ru/blog/personal/aguryanov/30720.php
3. Умысков А. В., Тимофеев А. С. Рекомендации по внедрению систем предотвращения утечек конфиденциальной информации (DLP-систем) в

информационные системы предприятий // Молодой ученый. — 2016. — №13. — С. 231-233. — URL https://moluch.ru/archive/117/32050/»

4. Васильев В. DLP-системы: человеческий фактор под контролем? // PC Week №22 (921) 13 декабря 2016, URL: https://www.itweek.ru/security/article/detail.php?ID=190703

5. Шихов Е. «Обзор DLP-систем на мировом и российском рынке //Anti-Malware – 2014. URL: https://www.anti-malware.ru/analytics/Technology_Analysis/DLP_market_overview_2014»

6. Сердюк В., Ванерке В. DLP на страже информации // itsec –2017 URL:https://www.dialognauka.ru/upload/Information_Security_3_2017_Serdiouk_Vanerke_DLP_info.pdf