АВТОМАТИЗАЦИЯ АНАЛИЗА УЯЗВИМОСТЕЙ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ ТЕХНОЛОГИИ TEXT MINING

Васильев В.И.[1], Вульфин А.М[2]., Кучкарова Н.В.[3]

Цель исследования: *разработка автоматизированной системы анализа уязвимостей программного обеспечения (ПО) промышленных информационно-управляющих систем (АСУ ТП) на основе технологии интеллектуального анализа текстов, написанных на естественном языке (Text Mining).*

Метод исследования: *сопоставление множества выявленных уязвимостей ПО и релевантных угроз безопасности информации путем оценки метрик семантической близости их текстовых описаний с использованием методов Text Mining.*

Полученные результаты: *предложена архитектура автоматизированной системы анализа уязвимостей ПО, позволяющая оценить уровень критичности уязвимостей и сопоставить их с наиболее подходящими по описанию (т.е. семантически близкими) угрозами из Банка данных угроз безопасности информации ФСТЭК России, обеспечивая при этом возможность ранжирования (приоритезации) рассматриваемых уязвимостей и угроз. Разработаны основные программные модули системы. Проведены вычислительные эксперименты с целью оценки эффективности ее применения. Показано, что применение разработанной системы позволяет повысить достоверность оценки степени критичности уязвимостей ПО, значительно сокращая затраты времени на поиск и сопоставление уязвимостей и угроз.*

Ключевые слова: *угрозы информационной безопасности, интеллектуальная фильтрация, векторное представление слов, лемматизация, семантическая близость.*

DOI: 10.21681/2311-3456-2020-04-22-31

1.Введение

Как показывает статистика последних лет, ситуация в мире в области кибербезопасности все более усложняется. В 2019г. было зафиксировано более 1,5 тыс. цифровых атак, что на 19% больше, чем в 2018г. Доля целенаправленных атак при этом выросла на 5% по сравнению с 2018г. и составила 60%[4]. Согласно данным[5], в 2019 г. было обнаружено более 22 тыс. новых уязвимостей. По крайней мере, треть из них оценивается экспертами как имеющие высокую и критическую степень риска. Важным шагом на пути решения данной проблемы является сбор, систематизация и обобщение информации из различных источников о ранее обнаруженных уязвимостях ПО (то, что сегодня входит в понятие Vulnerability Intelligence – «совокупность знаний об уязвимости» [1]).

Существует большое число официально признанных реестров и баз данных (БД) уязвимостей ПО (NDV, CVE, VulnDB, X-Force, Банк данных угроз безопасности информации ФСТЭК России и др.). Широкое применение получили различные системы классификации и оценки критичности уязвимостей (NIPC, SANC, nCircle, CVSS, WIVSS, и др.) [2]. В то же время, работа с указанными БД и системами предполагает «ручной» поиск и анализ уязвимостей с учетом особенностей ПО конкретной организации, что, как правило, требует больших затрат времени со стороны специалистов в области ИБ. Поэтому во многих исследованиях сегодня активно поднимается проблема автоматизации поиска и анализа уязвимостей ПО с использованием существующих БД и систем оценки уязвимостей [3,4].

Одним из перспективных путей решения данной проблемы является применение методов и технологий интеллектуального анализа текстов (Text Mining). Суть данного подхода применительно к затронутой выше проблеме заключается в том, что исходные текстовые описания уязвимостей, хранящиеся в БД, еще до их

4 Актуальные киберугрозы: итоги 2019 года [Positive Technologies Research]. [Электронный ресурс]. URL: https://www.ptsecurity.com/ru~ru/research/analytics/cybersecurity-threatscape-2018/ (дата обращения 01.08.2020).

5 Risk Based Security. Обзор уязвимостей: отчет за 2019 год. [Электронный ресурс]. URL: https://www.ict.moscow/research/obzor-uiazvimostei-otchet-za-2019-god/ (дата обращения 01.08.2020).

1 Васильев Владимир Иванович, доктор технических наук, профессор, профессор кафедры вычислительной техники и защиты информации Уфимского государственного авиационного технического университета, г. Уфа, Россия, e-mail:vasilyev@ugatu.ac.ru

2 Вульфин Алексей Михайлович, кандидат технических наук, доцент кафедры вычислительной техники и защиты информации Уфимского государственного авиационного технического университета, г. Уфа, Россия, vulfin.alexey@gmail.com

3 Кучкарова Наиля Вакилевна, магистр, старший преподаватель кафедры вычислительной техники и защиты информации Уфимского государственного авиационного технического университета, г. Уфа, Россия, nailya_kuchkarov@mail.ru

экспертной оценки (в значительной степени субъективной) содержат значительный объем существенной (скрытой) информации, которую нужно выявить из исходного текста и правильно интерпретировать, что и позволяют методы Text Mining. Об интересе к данному направлению и его перспективности говорит достаточно большое число появившихся в последние годы публикаций [5-14], касающихся тех или иных аспектов решения данной задачи с применением Text Mining.

В отличие от указанных работ, авторы данной статьи предлагают при анализе описания уязвимостей ПО использовать дополнительно информацию, полученную путем сопоставления описаний этих уязвимостей с описаниями спроецированных (связанных с ними) угроз, взятыми из БД угроз. Аналогичная идея, связанная с сопоставлением описаний уязвимостей и угроз, была высказана ранее в [15], где была предложена автоматизированная методика выявления скрытых взаимосвязей уязвимостей, зафиксированных с помощью сканера, и угроз ИБ на основе вычисления базовых метрик CVSS и их последующего сравнения, на основании чего составляется список пар «уязвимости – релевантные угрозы» для конкретной информационной системы (ИС). Ниже в 3-ей части статьи будет произведено сравнение результатов, полученных с помощью предложенного авторами подхода, и результатов, полученных с помощью методики [16].

2. Text mining и анализ уязвимостей

2.1 Предварительная подготовка и формализация текстовых данных на естественном языке

Существующие базы данных (БД) уязвимостей (NDV, CVE List, Банк угроз безопасности информации ФСТЭК России и др.) содержат огромное количество информации об уязвимостях ПО, поступающей из различных источников. Так, на конец 2019г. в базе данных NDV содержалось более 136 тысяч записей уязвимостей[6], В Банке данных угроз безопасности информации (БДУ) ФСТЭК России по состоянию на 01.08.2020г. хранилось 217 записей с описанием угроз и 28010 записей с описаниями уязвимостей ПО.

Для того чтобы перейти к использованию методов машинного обучения, необходимо прежде всего произвести предварительную обработку указанной информации (т.е. текстовых описаний уязвимостей, записанных на естественном языке) с помощью следующих операций [17]:

- нормализация (приведение текста к более простому виду удаление знаков пунктуации, аббревиатур, стоп-слов, не несущих смысловой нагрузки союзов, предлогов, междометий.);
- стеммизация (приведение слова к его корню, путем устранения суффиксов, приставок, окончаний);
- лемматизация (приведение слова к смысловой канонической форме – инфинитив, именительный падеж единственного числа и т.д.).

В результате удается сократить текст, убрав из него все несущественные для последующего изучения детали. Следующим шагом преобразования полученного «рафинированного» текста является переход от слов и предложений к их векторному представлению в многомерном семантическом пространстве признаков.

Широкую известность в качестве метода векторного представления слов (Word Embedding) получил разработанный в 2013г. группой исследователей под руководством Т. Миколова (корпорация Google) алгоритм Word2Vec [18]. Данный алгоритм обучается на прочтении большого количества документов (в нашем случае – текстовых описаний из БД уязвимостей) с последующим запоминанием того, какое слово возникает в схожих контекстах. По завершении процесса обучения на достаточном количестве данных Word2Vec генерирует вектор заданной длины для каждого слова в образованном таким образом словаре, в котором слова со схожим значением располагаются ближе друг к другу. Разновидности данного алгоритма – модель непрерывного «мешка слов» (Continuous Bag-Of-Words, CBOW), когда по текущему слову в предложении предсказываются слова из его контекста, и модель Skip-Gram, когда по окружению слова, т.е. по его контекстным словам, предсказывается центральное слово сегмента текста. В качестве расширения алгоритма Word2Vec предложен алгоритм Doc2Vec. Он формирует так называемый paragraph vector (вектор абзаца) – алгоритм обучения без учителя, который создает пространство признаков фиксированной длины из документов разной длины. Для оценки меры семантической близости слов (точек в рассматриваемом многомерном пространстве) при этом могут использоваться различные метрики расстояния (евклидова, косинусная метрика и др.) [19,20].

2.2 Структурно-функциональная организация автоматизированной системы анализа уязвимостей ПО

Автоматизированная система анализа уязвимостей ПО предназначена для автоматизации процесса обработки накапливаемых с помощью хостовых сканеров безопасности данных об обнаруженных уязвимостях. Ядром системы является механизм сопоставления текстовых описаний уязвимостей и связанных с ними угроз безопасности информации, что позволяет уточнить и актуализировать перечень рассматриваемых уязвимостей, и кроме того, осуществить приоритезацию указанных угроз с учетом дополнительной информации о наличии зависимостей между угрозами и уязвимостями ПО.

Архитектура системы включает в себя следующие основные подсистемы:

- подсистему локального хранения актуальной копии БДУ ФСТЭК (I);
- подсистему сопоставления угроз и уязвимостей на основе их текстового описания (II);
- подсистему оценки актуальных угроз и уязвимостей для корпоративной информационной системы (III).

Детализированная структурно-функциональная организация системы представлена на рисунке 1. Рассмотрим подробнее состав каждой из подсистем.

6 National Vulnerability Database. [Электронный ресурс] URL: https://nvd.nist.gov/ (дата обращения: 01.08.2020).

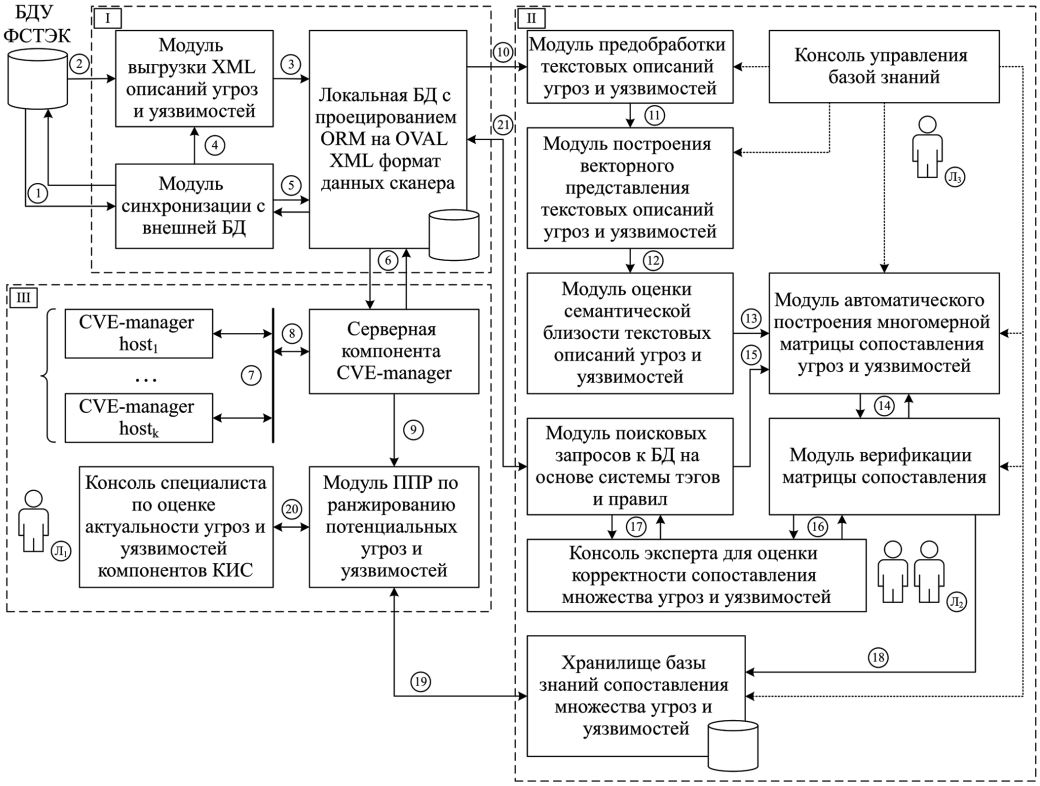

Рис.1. Структурно-функциональная организация подсистемы отбора и анализа актуальных угроз и уязвимостей на основе оценки семантической близости их текстовых описаний

Подсистема локального хранения актуальной копии БДУ ФСТЭК (I) предназначена для построения СУБД с объектно-ориентированным проецированием (ORM) хранимых сущностей, характеризующих угрозы и уязвимости в формате открытого языка описания и оценки уязвимостей (OVAL) [21], на сериализуемые файлы с выбранной XML-схемой. Модуль синхронизации с внешней БД сопоставляет (5) временные метки изменений данных внешнего хранилища БДУ ФСТЭК (1) и метки в локальном хранилище. По результатам сопоставления принимается решение о запуске (4) механизма синхронизации. Модуль выгрузки XML-описаний угроз и уязвимостей из внешней базы подключается (2) к серверу БДУ и выполняет импорт данных (3) в локальную СУБД в требуемом формате.

Подсистема сопоставления угроз и уязвимостей на основе их текстового описания (II) предназначена для построения базы знаний, описывающей отображение множества уязвимостей на множество угроз.

БДУ, помимо формальных метрик, содержит текстовое описание уязвимости и угрозы, характеризующее особенности их проявления и возможности эксплуатации злоумышленником. Модуль предобработки текстовых описаний угроз и уязвимостей извлекает (10) данные из локального хранилища и выполняет цепочку подготовительных преобразований текстовых описаний (фильтрацию и нормализацию) сущностей для передачи (11) в модуль построения их формализованных векторных представлений. Модуль оценки семантической близости текстовых описаний использует (12) формализованные векторы признаков каждой сущности для попарной оценки сходства на основе косинус-метрики. Далее, модуль автоматизированного построения многомерной матрицы сопоставления угроз и уязвимостей на основе оценок семантической близости формирует (13) матрицу отображения множества уязвимостей на множество угроз вида

$$\begin{pmatrix} & V_1 & V_2 & \cdots & V_m \\ T_1 & d(T_1,V_1) & d(T_1,V_2) & \cdots & d(T_1,V_m) \\ T_2 & d(T_2,V_1) & d(T_2,V_2) & \cdots & d(T_2,V_m) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ T_n & d(T_n,V_1) & d(T_n,V_2) & \cdots & d(T_n,V_m) \end{pmatrix} \quad (1)$$

Эксперты (Л2) с помощью консоли доступа выполняют оценку (16) корректности сопоставления множества угроз и уязвимостей и выполняют корректировку в случае необходимости. В процессе верификации (14) матрицы сопоставления эксперты опираются (17) на имеющийся механизм поисковых запросов к локальной БД на основе системы тегов и правил фильтрации, предусмотренных БДУ ФСТЭК (21, 15). Верифицированные сопоставления угроз и уязвимостей помещаются в хранилище базы знаний для последующего использования экспертами в ходе аудита ИБ корпоративной ИС. Специ-

алист по знания (Л3) управляет работой модулей предобработки и векторизации текстовых описаний, а также следит за метриками качества базы знаний.

Подсистема оценки актуальных угроз и уязвимостей для корпоративной информационной системы (III) с помощью клиент-серверного сканера (CVE-manager) обеспечивается сбор (7, 8) данных об уязвимостях программного обеспечения рабочих станций и серверов КИС. Применяется связка ПО CVE-manager и ScanOVAL для ОС Linux и Windows, управляемое серверной компонентой, и взаимодействующее (6) с локальной БД. Результаты поиска уязвимостей с помощью сканеров безопасности представляются в виде XML документов с разметкой на языке OVAL. Применение графических интерфейсов работы с найденными уязвимостями ScanOVAL и WEB-интерфейс БДУ ФСТЭК позволяют выполнить фильтрацию найденных уязвимостей по 15 параметрам. Однако, ввиду значительного количества выявляемых уязвимостей на отдельных хостах (более 200 уязвимостей для системы с систематическим обновлением минимального набора прикладного ПО), ручная фильтрация даже наиболее критических по оценкам уязвимостей может занять длительное время. Существующие решения [15] позволяют упростить поиск и сопоставление актуальных угроз и уязвимостей для конкретных версий ПО, но дальнейшая автоматизация процедуры подбора актуальных угроз и уязвимостей на основе данных интеллектуальной фильтрации и оценки семантической близости их текстовых описаний позволит масштабировать решение для крупных ИС. С помощью консоли специалист по ИБ (Л1) выполняет оценку (20) актуальных угроз и уязвимостей для отдельных узлов КИС, руководствуясь рекомендациями модуля поддержки принятия решений по ранжированию и сопоставлению потенциальных угроз и уязвимостей, полученных (9) в результате сканирования ПО ИС, и механизмами интеллектуальной фильтрации (19) на основе извлекаемых из базы знаний (рис. 2).

2.3 Архитектура конвейера по обработке данных текстовых описаний угроз и уязвимостей БДУ

Функциональная схема конвейера подготовки текстовых данных и оценки семантической близости текстовых описаний угроз и уязвимостей представлена на рисунке 3.

Ключевыми этапами обработки являются:

Загрузка данных из локальной БД (1) – необходима для преобразования текстовых полей каждой записи в единое текстовое описание для последующей обработки.

Нормализация (2) текстовых описаний угроз и уязвимостей- включает шаги по символьной фильтрации, токенизации и фильтрации с использованием общего и специализированного (формируемого экспертами) «стоп-словарей». Заключительным шагом является лемматизация с применением инструментов pymystem3.

Экспертная структурно-семантическая разметка (3) текста на основе системы doccano – позволяет выделить семантические особенности текстовых описаний (ключевые слова, ключевые словосочетания, отношения между сущностями) и уточнить состав специализированного стоп-словаря.

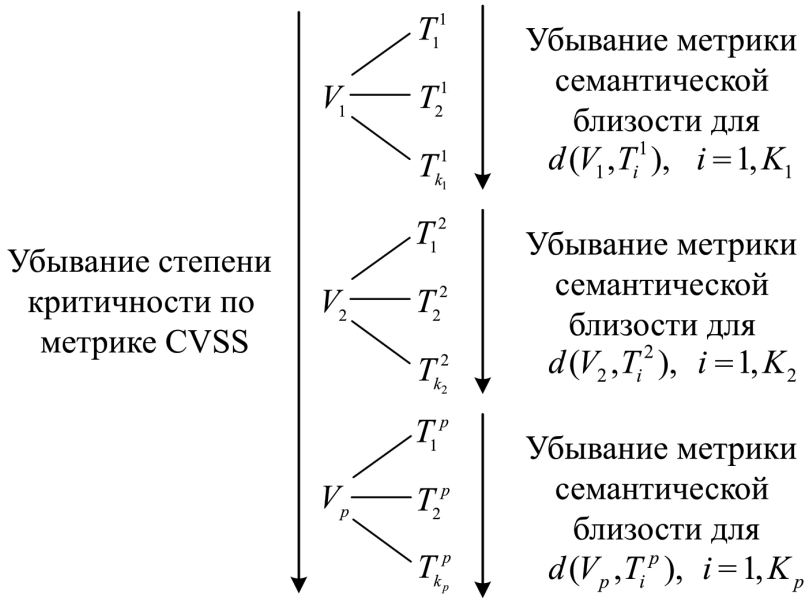

Рис.2. Список актуальных уязвимостей, ранжированных по степени критичности, и сопоставленные с ними угрозы (в порядке убывания метрики семантической близости)

DOI:10.21681/2311-3456-2020-04-22-31

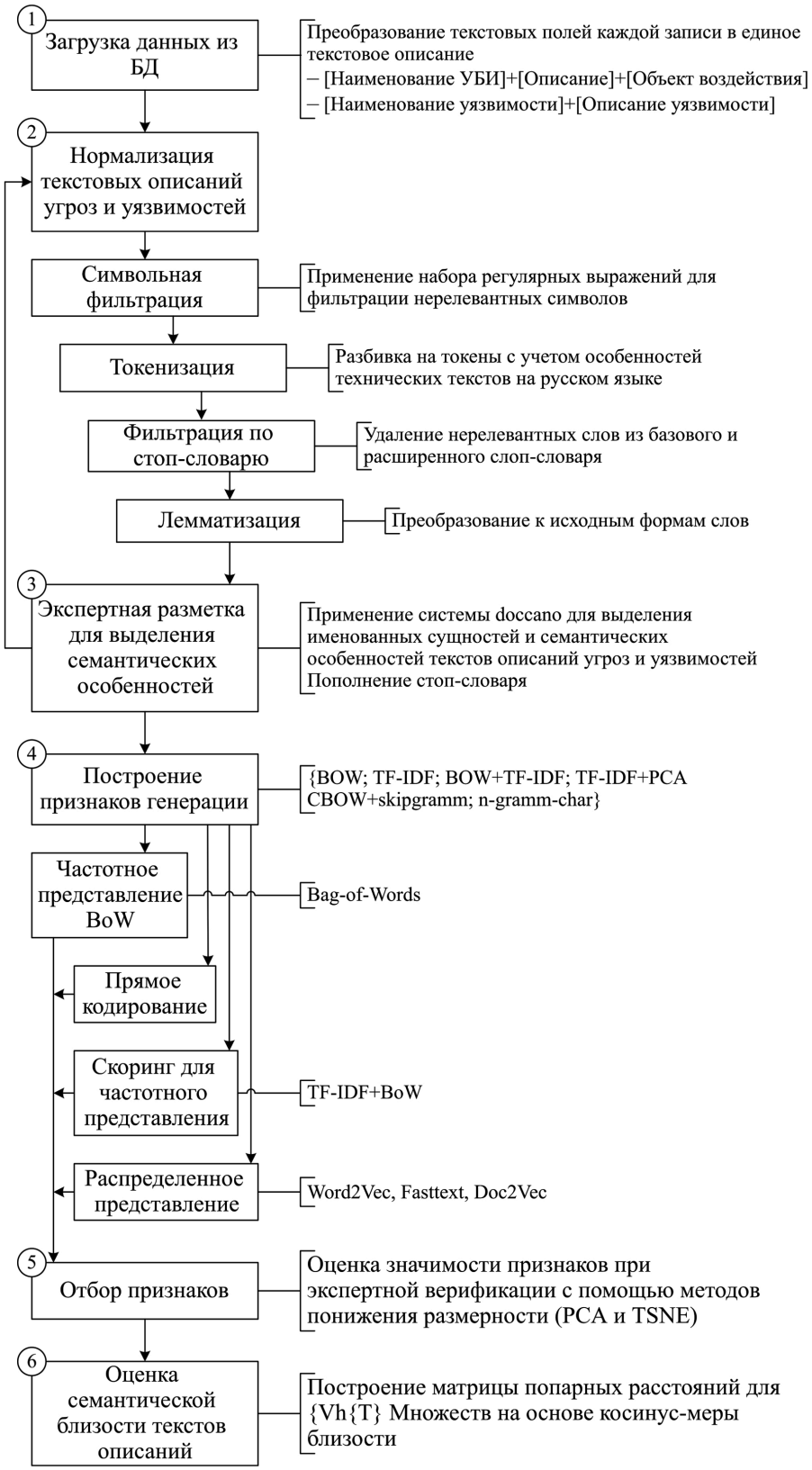

Рис.3. Функциональная схема конвейера подготовки текстовых данных и оценки семантической близости текстовых описаний угроз и уязвимостей

Ключевым является этап (4) построения формализованного вектора признаков текстовых описаний. Применяемые схемы частотного представления (Bag of Word, BoW), прямого кодирования, скоринга для частотного представления (BoW + TF-IDF) и распределенное представление (с помощью моделей Word2Vec, Doc2Vec и FastText) позволяют сформировать гетерогенный вектор признаков, однозначно определяющий текстовое описание сущности в признаковом пространстве. На следующем этапе (5) выполняется отбор наиболее значимых признаков с помощью экспертной оценки структуры двухмерной визуализации стохастического вложения соседей с t-распределением (TSNE) редуцированного пространства признаков с помощью метода главных компонент (PCA) или сингулярного разложения (SVD) (рисунок 4). Визуальная оценка при разной заданной избирательности модели векторизации позволяет оценить, насколько компактными и отделимыми являются формируемые группы текстовых описаний угроз и уязвимостей. Основная задача данного этапа – оценить, имеется ли структура в данном признаковом пространстве у данных, т.е. присутствует ли группировка семантически близких сущностей в отделимые кластеры.

Заключительным является этап (6) оценки семантической близости текстовых описаний и формирование матрицы попарных расстояний на основе косинус-метрики.

3. Результаты эксперимента по построению отношения «уязвимости – релевантные угрозы»

Доступная база уязвимостей содержит 27085 записей, база угроз – 217 записей. Из текстовых описаний объемом 740634 слова был сформирован словарь 12884 уникальных слов. После процедуры предобработки и нормализации построена модель Doc2Vec с помощью фреймворка Gensim. Размер формализованного вектора признаков выбран равным 100, количество эпох обучения модели равно 40. Параметры модели подбирались экспериментально.

В качестве иллюстрации работы системы рассмотрим выявленную хостовым сканером безопасности уязвимость BDU:2015-00285 «Уязвимость программного обеспечения Flash Player, позволяющая удаленному злоумышленнику нарушить конфиденциальность, целостность и доступность защищаемой информации». Данной уязвимости эксперт в ручном режиме поставил в соответствие угрозу УБИ.192 (таблица 1).

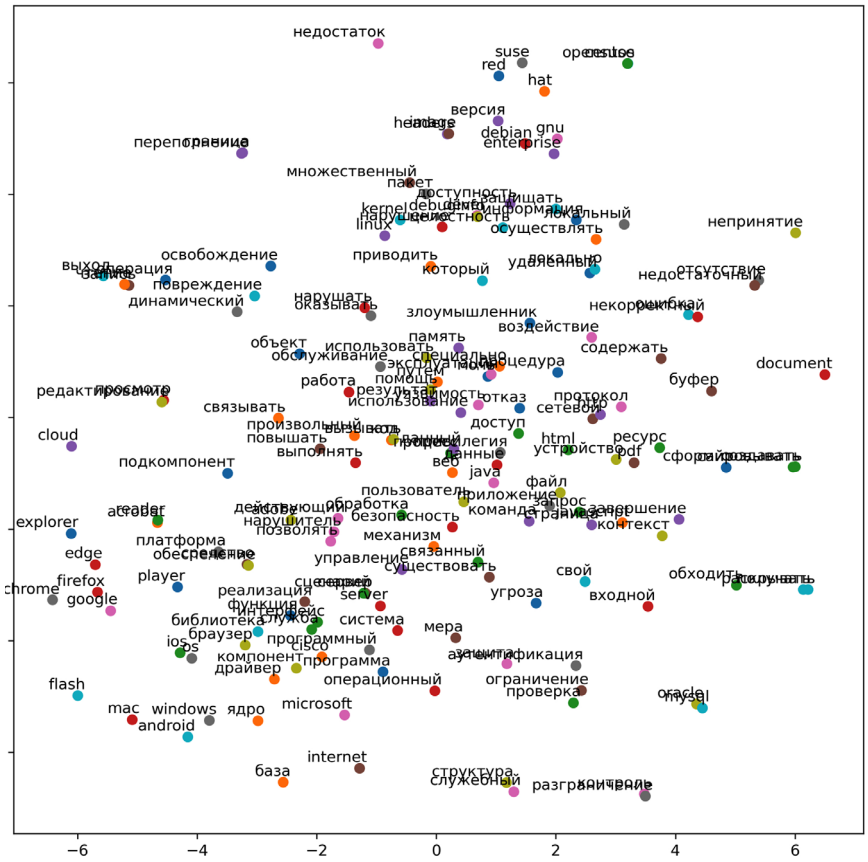

Рис.4. Двухмерная визуализация стохастического вложения соседей с t-распределением (TSNE) редуцированного пространства признаков распределенного векторного представления Doc2Vec

Таблица 1

Экспертное сопоставление угроз и уязвимостей из БДУ ФСТЭК

Угроза	Уязвимость	Воздействие/уровень опасности
УБИ.192 Угроза использования уязвимых версий программного обеспечения.	BDU:2015-00285 Уязвимость программного обеспечения Flash Player, позволяющая удаленному злоумышленнику нарушить конфиденциальность, целостность и доступность защищаемой информации	Критический уровень опасности (базовая оценка CVSS 2.0 составляет 10)

Используя текстовое описание уязвимости, с помощью разработанного модуля автоматизированной системы осуществим выбор семантически близких по описанию угроз из БДУ ФСТЭК. На рисунке 5 показаны результаты подбора 10 релевантных угроз, отсортированных в порядке убывания метрики семантической близости.

Как видно из рисунка, угроза УБИ.192 попадает в данный перечень, что совпадает с результатом предварительного экспертного оценивания. Аналогичным образом, для выбранных в процессе экспертного анализа и сбора данных сканерами уязвимостей (поиск установленных версий ПО с имеющимися уязвимостями по БДУ) производится подбор соответствующих угроз. Финальная стадия анализа позволяет упростить работу эксперта, значительно сократив время на поиск и сопоставление уязвимостей и угроз.

Применяемые для префильтрации средства [15,16] позволяют упростить поиск и сопоставление актуальных угроз и уязвимостей для конкретных версий ПО и сократить количество просматриваемых экспертом угроз для отдельной уязвимости с 200 до 4.

Сравнение процедуры анализа уязвимостей WEB-браузера Firefox с [16] приведены в таблице 2.

Рис. 5. Релевантные угрозы, отсортированные в порядке убывания нормированной метрики семантической близости (score) к данной уязвимости BDU:2015-00285

Таблица 2

Сравнение процедуры анализа уязвимостей

Параметр	Поиск по тегам	Система [16]	Автоматизированная система на основе Text Mining
Ввод информации	Вручную, графический WEB-интерфейс БДУ	Формирование запроса оператором в графическом интерфейсе	Автоматизированная обработка результатов работы сканеров уязвимостей
Количество найденных уязвимостей	41	41	48
Количество сопоставленных угроз	2 (ручное сопоставление)	8 (задается на основе сформированной матрицы)	10 (задается пороговыми и количественными метриками, определяющими чувствительность фильтра на основе сформированной матрицы)
Затраченное время	Более 11 минут	20 с	< 5 с

Согласно оценке [16], время, затрачиваемое на сопоставление угроз и уязвимости «вручную» для полного списка, при этом составляет более 2 часов, применение же предлагаемых решений позволяет сократить время анализа до 20 секунд. Предлагаемая система для сопоставления на основе анализа текстовых описаний позволяет выполнить ранжирование оставшихся угроз по степени их семантической близости к конкретной уязвимости, тем самым дополнительно снижая когнитивную нагрузку на эксперта и уменьшая время анализа.

Заключение

Рассмотрена архитектура системы анализа критичных уязвимостей ПО с использованием технологии Text Mining, основанная на алгоритмах векторного представления слов и оценки семантической близости текстовых описаний уязвимостей, выявленных с помощью сканеров безопасности, и описаний релевантных угроз из Банка данных угроз безопасности информации ФСТЭК России. Программная реализация клиент-серверного прототипа данной системы и интеграция с модулями существующих решений позволяют:

- автоматизировать процесс сопоставления и ранжирования угроз ИБ для каждой выявленной уязвимости на рабочих станциях и серверах в составе корпоративной информационной системы;
- в несколько раз сократить время ручного анализа экспертом результатов работы сканеров за счет интеллектуальной фильтрации и ранжирования списка угроз;
- снизить когнитивную нагрузку на эксперта и повысить достоверность оценки степени критичности уязвимостей ПО за счет использования дополнительной информации о фактически существующих зависимостях между выявленными уязвимостями и потенциальными угрозами;
- масштабировать решение для крупных ИС за счет интеграции с существующими БД уязвимостей и формализации знаний экспертов о прецедентах сопоставления угроз и уязвимостей в пополняемой базе.

Исследование выполнено при финансовой поддержке Минобрнауки России (грант ИБ) в рамках научного проекта № 1/2020.

Рецензент: *Цирлов Валентин Леонидович, кандидат технических наук, доцент кафедры ИУ-8 «Информационная безопасность» МГТУ им. Н.Э. Баумана, г. Москва, Россия. E-mail: v.tsirlov@bmstu.r*

Литература

1. Smyth V. Vulnerability Intelligence // ITNOW, Dec. 2016. P.26-27.
2. Федорченко А.В., Чечулин А.А., Котенко И.В. Исследование открытых баз уязвимостей и оценка возможностей их применения в системах анализа защищенности компьютерных сетей // Информационно-управляющие системы. 2014. №5. С.72-79.
3. Tao Wen, Yuquing Zhang, Gang Yang. A Novel Automatic Severity Vulnerability Assessment Framework // Journal of Communications, Vol. 10. №5. May 2015. pp. 320-329.
4. Detection and Remediation Method for Softwere Security / Jessoo Jurn, Taeeun Kim, Hwankuk Kim, An Automated Vulnerability // Sustainability, May 2018. №10. 1657. DOI: 10.3390/ su10051652012.
5. Spanos G., Angeis L., Toloudis D. Assessment of Vulnerability Severity using Text Mining // Proceedings of the 21st Pan-Hellenic Conference, Sept.2017, Larissa, Greece. pp. 1-6.

6. Learning to Predict Severity of Software Vulnerability Description / Han Z., Li X., Xing Z., Liu H., Feng Z. // Proceedings of the 2017 International Conference on Software Maintenance and Evolution (ICSME), Shanghai, China, Nov. 2017. pp. 125-136.
7. Lee Y., Shin S. Toward Semantic Assessment of Vulnerability Severity: A Text Mining Approach // Proceedings of ACM CIKM Workshop (EYRE' 18), 2018. [Электронный ресурс]. URL: https://www.CEUR-WS.org/Vol1-2482/papers.pdf (дата обращения 01.08.2020).
8. О проблеме выявления экстремистской направленности в текстах// Вестник Новосибирского государственного университета / Ананьева М.И., Кобозева М.В., Соловьев Ф.Н., Поляков И.В., Чеповский А.М.// Серия: Информационные технологии. 2016.Т.14.С.5-13.
9. Сравнительный анализ специальных корусов текстов для задач безопасности / Лаврентьев А.М., Рябова Д.М., Тихомирова Е.А., Фокина А.И., Чеповский А.М., Шерстинова Т.Ю. // Вопросы кибербезопасности. 2020. №3(37). С.54-60.
10. Mittal S. et al. Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities // 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE. 2016. pp. 860-867.
11. Benjamin V. et al. Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops // 2015 IEEE international conference on intelligence and security informatics (ISI). – IEEE. 2015. С. 85-90.
12. de Boer M. H. T. et al. Text Mining in Cybersecurity: Exploring Threats and Opportunities // Multimodal Technologies and Interaction. 2019. Т. 3. №. 3. pp. 62.
13. Nunes E. et al. Darknet and deep net mining for proactive cybersecurity threat intelligence // 2016 IEEE Conference on Intelligence and Security Informatics (ISI). IEEE. 2016. pp. 7-12.
14. Epishkina A., Zapechnikov S. A syllabus on data mining and machine learning with applications to cybersecurity // 2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC). IEEE/ 2016. pp. 194-199.
15. Селифанов В.В., Юракова Я.В., Карманов И.Н. Методика автоматизированного выявления взаимосвязей уязвимостей и угроз безопасности информации в информационных системах // Интерэкспо Гео-Сибирь, 2018. – С.271-276.
16. Применение методов автоматизации при определении актуальных угроз безопасности информации в информационных система с применением банка данных угроз ФСТЭК России / Селифанов В. В., Звягинцева П.А., Юракова Я.В., Слонкина И.С. // Интерэкспо Гео-Сибирь. 2017. Т. 8. С.202-209.
17. Петренко С. А., Петренко А. С. Моделирование систем обработки больших данных кибербезопасности // Информационные системы и технологии в моделировании и управлении. 2016. С. 279-284
18. Mikolov T., Chen K., Corrado G. Dean J. Efficient Estimation of Word Representation in Vector Space // Proceedings of Workshop at ICLR, 2013. [Электронный ресурс]. URL: https://www.arXiv.1301.3781 (дата обращения 01.08.2020).
19. Бондарчук Д.В. Векторная модель представления знаний на основе семантической близости термов // Вестник ЮрГУ.Серия: Вычислительная математика и информатика, 2017. Т.6. С.73-83.
20. Ali A., Alfaycz F., Alquhayz H. Semantic Similarity Measures Between Words: A Brief Survey // Sci.Int. (Lahore), №30 (6). 2018. pp. 907-914.
21. Gupta S., Gupta B. B. Detection, avoidance, and attack pattern mechanisms in modern web application vulnerabilities: present and future challenges //International Journal of Cloud Applications and Computing (IJCAC). 2017. Vol. 7. №. 3. pp. 1-43.

AUTOMATION OF SOFTWARE VULNERABILITIES ANALYSIS ON THE BASIS OF TEXT MINING TECHNOLOGY

Vasilyev V.I.[7], Vulfin A.M.[8], Kuchkarova N.V.[9]

Purpose: *the development of automated system of software vulnerabilities analysis for information-control systems on the basis of intelligent analysis of texts written on the natural language (Text Mining).* ***Methods:*** *the idea of the used investigation method is based on matching the set of extracted software vulnerabilities and relevant information security threats by means of evaluating the semantic similarity metrics of their textual description with use of Text Mining methods.* ***Practical relevance:*** *the architecture of the automated system of software vulnerabilities analysis is developed, the application of which allows us to evaluate the level of vulnerabilities criticality and match it with the most suitable by discretion (i.e. semantically similar) threats from the Bank of information security threats of FSTEC Russia while ensuring vulnerabilities and threats. The main software modules of the system have been developed.*

7 Vladimir Vasilyev, Dr.Sc.(Eng.), Professor, Professor of Department of Computer Engineering and Information Security, Ufa State Aviation Technical University, Ufa, Russia, E-mail: vasilyev@ugatu.ac.ru

8 Alexey Vulfin, Ph.D., Associate Professor of Department of Computer Engineering and Information Security, Ufa State Aviation Technical University, Ufa, Russia, E-mail: vulfin.alexey@gmail.com

9 Nailya Kuchkarova, M. Sc., Senior Lecturer of Department of Computer Engineering and Information Security, Ufa State Aviation Technical University, Ufa, Russia, E-mail: nailya_kuchkarov@mail.ru

Computational experiments were carried out to assess the effectiveness of its application. The results of comparative analysis show that application of the given system allows us to increase the credibility of evaluating the criticality degree of vulnerabilities, considerably decreasing the time for a search and matching vulnerabilities and threats.

Keywords: *information security threats, intelligent filtering, vector word representation, lemmatization, semantic proximity.*

References

1. Smyth V. Vulnerability Intelligence // ITNOW, Dec. 2016. P.26-27.
2. Fedorchenko A.V., CHechulin A.A., Kotenko I.V. Issledovanie otkrytyh baz uyazvimostej i ocenka vozmozhnostej ih primeneniya v sistemah analiza zashchishchennosti komp'yuternyh setej // Informacionno-upravlyayushchie sistemyyu. 2014. №5. S.72-79.
3. Tao Wen, Yuquing Zhang, Gang Yang. A Novel Automatic Severity Vulnerability Assessment Framework // Journal of Communications, Vol. 10. №5. May 2015. pp. 320-329.
4. Detection and Remediation Method for Softwere Security / Jessoo Jurn, Taeeun Kim, Hwankuk Kim, An Automated Vulnerability // Sustainability, May 2018. №10. 1657; doi: 10?3390/ su10051652012.
5. Spanos G., Angeis L., Toloudis D. Assessment of Vulnerability Severity using Text Mining // Proceedings of the 21st Pan-Hellenic Conference, Sept.2017, Larissa, Greece. pp. 1-6.
6. Learning to Predict Severity of Software Vulnerability Description / Han Z., Li X., Xing Z., Liu H., Feng Z. // Proceedings of the 2017 International Conference on Software Maintenance and Evolution (ICSME), Shanghai, China, Nov. 2017. pp. 125-136.
7. Lee Y., Shin S. Toward Semantic Assessment of Vulnerability Severity: A Text Mining Approach // Proceedings of ACM CIKM Workshop (EYRE' 18), 2018. [Электронный ресурс]. URL: https://www.CEUR-WS.org/Vol1-2482/papers.pdf (дата обращения 01.08.2020).
8. O probleme vyyavleniya ekstremistskoj napravlennosti v tekstah// Vestnik Novosibirskogo gosudarstvennogo universiteta / Anan'eva M.I., Kobozeva M.V., Solov'ev F.N., Polyakov I.V., CHepovskij A.M.// Seriya: Informacionnye tekhnologii. 2016.T.14.S.5-13.
9. Sravnitel'nyj analiz special'nyh korusov tekstov dlya zadach bezopasnosti / Lavrent'ev A.M., Ryabova D.M., Tihomirova E.A., Fokina A.I., CHepovskij A.M., SHerstinova T.YU. // Voprosy kiberbezopasnosti. 2020. №3(37). S.54-60.
10. Mittal S. et al. Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities //2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE. 2016. pp. 860-867.
11. Benjamin V. et al. Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops //2015 IEEE international conference on intelligence and security informatics (ISI). – IEEE. 2015. C. 85-90.
12. de Boer M. H. T. et al. Text Mining in Cybersecurity: Exploring Threats and Opportunities // Multimodal Technologies and Interaction. 2019. T. 3. №. 3. pp. 62.
13. Nunes E. et al. Darknet and deepnet mining for proactive cybersecurity threat intelligence //2016 IEEE Conference on Intelligence and Security Informatics (ISI). IEEE. 2016. pp. 7-12.
14. Epishkina A., Zapechnikov S. A syllabus on data mining and machine learning with applications to cybersecurity //2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC). IEEE/ 2016. pp. 194-199.
15. Selifanov V.V., Yurakova Ya.V., Karmanov I.N. Metodika avtomatizirovannogo vy`yavleniya vzaimosvyazej uyazvimostej i ugroz bezopasnosti informacii v informacionny`x sistemax //Intere`kspo Geo-Sibir`, 2018. pp.271-276.
16. Primenenie metodov avtomatizacii pri opredelenii aktual'nyh ugroz bezopasnosti informacii v informacionnyh sistema s primeneniem banka dannyh ugroz FSTEK Rossii / Selifanov V. V., Zvyaginceva P.A., YUrakova YA.V., Slonkina I.S. //Interekspo Geo-Sibir'. 2017. T. 8. C.202-209.
17. Petrenko S. A., Petrenko A. S. Modelirovanie sistem obrabotki bol'shih dannyh kiberbezopasnosti //Informacionnye sistemy i tekhnologii v modelirovanii i upravlenii. 2016. S. 279-284
18. Mikolov T., Chen K., Corrado G. Dean J. Efficient Estimation of Word Representation in Vector Space // Proceedings of Workshop at ICLR, 2013. [Электронный ресурс]. URL: https://www.arXiv.1301.3781 (дата обращения 01.08.2020).
19. Bondarchuk D.V. Vektornaya model' predstavleniya znanij na osnove semanticheskoj blizosti termov // Vestnik YUrGU.Seriya: Vychislitel'naya matematika i informatika, 2017. T.6. S.73-83.
20. Ali A., Alfaycz F., Alquhayz H. Semantic Similarity Measures Between Words: A Brief Survey // Sci.Int. (Lahore), №30 (6). 2018. pp. 907-914.
21. Gupta S., Gupta B. B. Detection, avoidance, and attack pattern mechanisms in modern web application vulnerabilities: present and future challenges //International Journal of Cloud Applications and Computing (IJCAC). 2017. Vol. 7. №. 3. pp. 1-43.