УДК 004

ВЫБОР ОПТИМАЛЬНОГО АЛГОРИТМА ДЕТЕКТИРОВАНИЯ ИНСАЙДЕРСКИХ АТАК ДЛЯ КОМПАНИЙ АНАЛИЗА СУДЕБНОЙ ПРАКТИКИ

Конкина Ольга Владимировна

студент
Самарский государственный аэрокосмический университет им. С.П. Королёва, Самара

author@apriori-journal.ru

Аннотация. Описан выбор оптимального алгоритма детектирования инсайдерской угрозы, путём сравнения нескольких алгоритмов на основе тестовых данных.

Ключевые слова: инсайдер; алгоритм; цифровой отпечаток; контекстный анализ; судебная практика.

OPTIMAL CHOICE DETECTION ALDORITHM INSIDER ATTACKS ANALYSIS FOR COMPANIES COURT

Konkina Olga Vladimirovna

student
Samara State Aerospace University, Samara

Abstract. Describes a selection of the optimal detection algorithm insider threats by comparing several algorithms based on test data.

Key words: insider; algorithm; digital fingerprint; contextual analysis; litigation.

С развитием информационных технологий в современном мире появляется все больше устройств, предназначенных для хранения и передачи информации. С каждым днём отследить передвижение важных для бизнеса компаний данных в таких условиях становится крайне сложно.

Владельцы компаний, от маленьких стартапов, до огромных холдингов, вынуждены прибегать к использованию различных программных средств для борьбы с инсайдерскими атаками. Известно, что идеального решения по защите от внутренних угроз не существует – все зависит от требований, заложенных в политиках ИБ. И основной задачей на сегодня остаётся способ точного определения является ли передаваемая информация конфиденциальной, несёт ли угрозу для интеллектуальной собственности компании её передача?

Для ответа на этот вопрос существует множество способов и алгоритмов определения, принадлежит ли передаваемый сотрудником файл к категории конфиденциальных.

Поскольку любая технология сама по себе практически бесполезна, ее необходимо рассматривать только в привязке к определенной задаче. Рассмотрим выбор оптимального алгоритма и конфигурации системы на примере компании, обрабатывающей большие объёмы документов от судов Российской Федерации. Конфиденциальными данными являются не все виды документов т.к. тексты большинства из них публикуются в широком доступе (определения, решения и постановления арбитражных дел), а вот тексты заявлений или каких-либо приложений, подаваемых участниками дела не подлежат огласки т.к. могут содержать информацию с персональными данными.

Первоначально нужно определить какие способы передачи данных нужно контролировать. Детально изучив специфику работы компании, было установлено, что основным способом передачи данных является электронная почта и различного рода чаты. Поэтому существует острая необходимость контролировать интернет трафик компании. Наилучшим

решением для такого рода задачи является программа – снифер. Суть DLP системы заключается в том, что система прослушивает канал на идущую по нему информацию и на основании каких-то своих внутренних алгоритмов выносит определенное суждение на то, можно ли признать эту информацию конфиденциальной. А после, на основе предварительно настроенных политик и полученного суждения, принимается решение о том, что делать дальше с этой информацией. В частности, система может разрешить ее передачу, заблокировать ее или сообщить об этом факте сотруднику безопасности. Таким образом, ключевым параметром любой DLP-системы является алгоритм фильтрации трафика, который позволяет вынести суждение о конфиденциальности тех или иных данных.

Итак, на следующем шаге необходимо выбрать наилучший алгоритм обработки файлов для определения принадлежности файла к списку конфиденциальных. Были рассмотрены следующие варианты – поиск регулярных выражений, цифровые отпечатки (digital fingerprints), лингвистический/морфологический анализ.

Каждый из алгоритмов реализован и протестирован на наборе тестовых данных, результаты тестирования и анализа работы алгоритмов представлены в таблице ниже.

При поступлении информации из судов в компании анализа судебной практики, документы проходят обработку. На основе поступившей информации заполняется база атрибутов с выделением основных данных, которые и требуется защищать от утечек. Но отдельный атрибут практически не несёт никакой ценности, а вот сочетание нескольких значений уже является конфиденциальными. Например, отдельно взятый ИНН (индивидуальный номер налогоплательщика) является обычной последовательностью чисел, а вот в сочетании с адресом, ОГРН или ФИО можно сделать вывод о принадлежности значений к тому или иному физическому или юридическому лицу. После анализа работы алго

Таблица 1

Результаты тестирования и анализа работы алгоритмов

Название алгоритма	Краткое описание	Преимущества	Недостатки	Целевые типы данных
Поиск регулярных выражений	Поиск производится по заранее определенным кускам текста – образцам или шаблонам	Простая, понятная и легко настраиваемая технология	Большое количество ложных срабатываний. Технология бесполезна для неструктурированной информации	Структурированная конфиденциальная информация
Лингвистический / морфологический анализ	Поиск и анализ информации производится на основе заранее заданного словаря	Простой алгоритм. Низкие требования к производительности при наличии словарной базы	Большое количество ложных срабатываний. Требуется большая подготовительная работа для создания словарной базы	Полностью не структурированный контент
Цифровые отпечатки (digital fingerprints)	Поиск по заранее снятым «отпечаткам» (хеш-функциям), которые сравниваются с текущим трафиком	Практически полное отсутствие ложных срабатываний	При больших объёмах возникает проблема с нагрузкой на систему. Алгоритм работает только на точных совпадениях	Информация в базах данных или хранилищ данных

ритмов был сделан вывод, что наиболее подходящим алгоритмом для работы с такими данными являются цифровые отпечатки, но алгоритм необходимо доработать и реализовать получение составных цифровых отпечатков. Суть алгоритма в том, что отпечатки снимаются не с отдельных полей или частей текста, а именно с сочетаний полей различного типа.

Таким образом, для отслеживания возможных инсайдерских утечек в компаниях, занимающихся контролем судебной практики, наиболее подходящим оказался алгоритм электронных отпечатков. Алгоритм имеет высокую степень детектирования, прост в использовании и удобен во внедрении.