Использование методов анализа контента в DLP системах
Королев В. В.

Королев Виталий Владимирович / Korolev Vitalij Vladimirovich – аспирант,
кафедра информатики и кибернетики,
Байкальский государственный университет, г. Иркутск

Аннотация: *в статье говорится об основных методах контекстного контроля в системах защиты информации. Рассмотрены преимущества и недостатки данных методов.*
Ключевые слова: *контент, контекст, категоризация, сигнатуры.*

Введение

На сегодняшний день наиболее эффективным и очень популярным подходом к защите от утечек информации с компьютеров является механизм контекстного контроля – анализа формальных признаков документа, а также запрет или разрешение передачи данных для конкретных пользователей в зависимости от форматов данных, типов интерфейсов и устройств, сетевых протоколов, направления передачи, дня недели и времени суток и т.д.

Однако во многих случаях требуется более глубокий уровень контроля – например, проверка содержимого передаваемых данных на наличие персональной или конфиденциальной информации в условиях, когда порты ввода-вывода не должны блокироваться, чтобы не нарушать производственные процессы. В таких ситуациях дополнительно к контекстному контролю необходимо применение технологий контентного анализа и фильтрации, позволяющих выявить и предотвратить передачу неавторизованных данных, не препятствуя при этом информационному обмену в рамках служебных обязанностей сотрудников.

Контент и контекст

Применяя методы контентной фильтрации необходимо четко различать определения контента и контекста. Под термином контент следует понимать любую значимую информацию для организации (основное содержимое документа).

Контекст же включает в себя формальные признаки документа, такие, как:

- источник информации,
- размер,
- информация об отправителе,
- информация о получателе,
- метаданные,
- время,
- формат и пр.

Любое DLP решение включает в себя в первую очередь контекстный анализ информации. В таком случае система проверяет контекст, в котором передается информация, т.е. извлекает метки файла, смотрит его размер, анализирует поведение пользователя и т.д. Но зачастую анализа контекста недостаточно, требуется анализ основного содержания информации. Это означает, например, что при проверке на секретность стандартных офисных документов в формате .docx система сначала переведет их в текстовый формат, а затем, используя заранее подготовленные данные, вынесет по этому тексту вердикт. Процесс проверки является очень сложным, трудоемким, повышающим ценность используемой DLP системы.

Методы контент-анализа

Первым шагом в контентном анализе является открытие документа. Затем система преобразует файл в специализированный формат, удобный для подачи его на вход алгоритма. Если это простой текстовый документ – это легко, но когда необходимо проанализировать двоичный файл это становится намного сложнее. В таком случае DLP системы решают это с помощью технологии File cracking [1].

Затем наступает основной этап - анализ контента передающегося файла. Тут появляются различные варианты, отражающие суть происходящего анализа. На данный момент существует несколько популярных технологий фильтрации, каждая из которых имеет как преимущества, так и определенные недостатки. Например, самый распространенный метод цифровых отпечатков обеспечивает сравнительно высокую точность, однако требует предварительного сбора отпечатков. А эффективность морфологического анализа прямо пропорциональна размеру и качеству словарной базы, собранной на начальном этапе реализации проекта.

Рассмотрим шесть основных методов анализа контента, используемых в DLP системах, их сильные и слабые стороны.

1. Сигнатуры

Самый простой метод контроля — поиск в потоке данных некоторой последовательности символов. Иногда запрещенную последовательность символов называют «стоп-выражением», но в более общем случае она может быть представлена не словом, а произвольным набором символов, например, определенной меткой. Если система настроена только на одно слово, то результат ее работы — определение 100%-го совпадения, т.е. метод можно отнести к детерминистским. Однако чаще поиск определенной последовательности символов все же применяют при анализе текста. В подавляющем большинстве случаев сигнатурные системы настроены на поиск нескольких слов и частоту встречаемости терминов.

Сильные стороны: простота пополнения словаря запрещенных терминов и очевидность принципа работы, это самый верный способ, если необходимо найти соответствие слова или выражения на 100%.

Слабые стороны: большинство производителей DLP-систем работают для Западных рынков, а английский язык очень «сигнатурен» — формы слов чаще всего образуются с помощью предлогов без изменения самого слова. В русском языке все гораздо сложнее, так как у нас есть приставки, окончания, суффиксы. Для примера можно взять слово «ключ», которое может означать как «ключ шифрования», «ключ от квартиры», «родник», «ключ или PIN-код от кредитной карты», так и множество других значений. В русском языке из корня «ключ» можно образовать несколько десятков различных слов. Это означает, что если на Западе специалисту по защите информации от инсайдеров достаточно ввести одно слово, в России специалисту придется вводить пару десятков слов и затем еще изменять их в шести различных кодировках. Реальное применение этого метода требует наличие лингвиста или команды лингвистов как на этапе внедрения, так и в процессе эксплуатации и обновления базы.

2. Регулярные выражения.

Этот метод применяется в большинстве DLP-продуктов. Он проверяет контент по определенным правилам. Регулярные выражения позволяют находить совпадения по форме данных, в нем нельзя точно указать точное значение данных, в отличие от «сигнатур». Такой метод детектирования эффективен для поиска:

- ИНН,
- КПП,
- номеров счетов,
- номеров кредитных карт,
- номеров телефонов,
- номеров паспортов,
- клиентских номеров.

Большинство DLP-систем укрепляют используемые базовые регулярные выражения своими собственными правилами дополнительного анализа (например, инициалы в непосредственной близости от адреса и номера кредитной карты).

Также использование регулярных выражений позволяет DLP-системе обеспечивать соответствие требованиям все более популярного стандарта PCI DSS, разработанного международными платежными системами Visa и MasterCard для финансовых организаций.

Сильные стороны: Правила быстро обрабатываются и легко конфигурируются. Регулярные выражения позволяют определить специфичный для каждой организации тип контента. Большинство продуктов поставляются с начальными наборами правил. Эта технология хорошо известна и легко используется в различных продуктах.

Слабые стороны: данный способ склонен к частым ложным срабатываниям. Предлагает очень слабую защиту для неструктурированного контента. С помощью регулярных выражений можно найти конфиденциальную информацию только определенной формы.

3. Database Fingerprinting

Иногда переводится как: «Точное совпадение данных». Этот метод использует либо дамп базы данных или базу данных в реальном времени (с помощью ODBC связи) и только ищет точные совпадения. Например, создается политика, чтобы проанализировать данные по номерам кредитных карт из клиентской базы, тем самым игнорируя аналогичные данные, используемые внутри. Более продвинутые инструменты позволяют искать комбинации данных.

Сильные стороны: Очень низкое количество ошибочных результатов (близкие к 0). Позволяет защитить конфиденциальные данные клиентов, игнорируя аналогичные личные данные сотрудников.

Слабые стороны: Текущие соединения могут повлиять на производительность базы данных. Большие базы данных влияют на работу устройства.

4. Partial Document Matching

Эта метод проверяет полное или частичное совпадение с защищенным контентом. Таким образом возможно настроить политику безопасности так, чтобы защитить конфиденциальный документ, и DLP решение будет искать или полный текст документа, или его часть [2].

Большинство решений основывается на методе, известном как циклическое хеширование, где Вы берете хеш части содержания, смещаете предопределенное число символов, затем берете другой хеш и продолжаете идти, пока документ полностью не загружен как ряд наложения значений хэш-функции. Исходящее содержание выполнено через тот же метод хеша и значения хэш-функции, сравненные для соответствий. Много продуктов используют циклическое хеширование в качестве основы, затем добавляют более усовершенствованный лингвистический анализ.

Сильные стороны: Возможность защитить неструктурированные данные. Низкое количество ложных срабатываний (некоторые поставщики скажут нулевые ложные положительные стороны, но любое общее предложение/текст в защищенном документе может инициировать предупреждения). Не полагается на полное соответствие больших документов; могут находиться нарушения политики на даже частичном соответствии.

Слабые стороны: ограничения производительности на суммарный объем содержания, которое может быть защищено. Общие фразы/формулировка в защищенном документе могут инициировать ложные срабатывания. Необходимо точно знать, какие документы необходимо защитить. Легко избежать с помощью шифрования.

5. Статистический анализ

В данном методе используется обучение машины, байесовский анализ и другие статистические методы, для того чтобы проанализировать контент и найти нарушения в его содержимом, которое напоминает защищенное содержание. Эта категория включает широкий диапазон статистических методов, которые варьируются значительно по реализации и эффективности. Некоторые методы аналогичны методам, используемым в борьбе со спамом.

Сильные места: Может работать с большим количеством контента. Может использовать такие политики безопасности, как «предупреждение» о какой либо исходящей информации, напоминающей защищенный контент.

Слабые места: Частые ложные срабатывания. Требует длительной подготовки и настройки — чем больше, тем лучше.

6. Концептуальный/лексический анализ

Этот метод использует комбинацию словарей, правил и других исследований, чтобы защитить контент, который напоминает «идею». Например, можно настроить политику безопасности на предупреждения о том, что исходящая информация по содержанию напоминает защищенный контент, а также на подсчет количества ключевых фраз, подсчета слов, похожих на нарушения.

Сильные места: Не все корпоративные политики или содержание могут быть описаны, используя готовые примеры, сигнатуры; Концептуальный анализ может определять нарушения политики безопасности, когда другие методы не могут даже думать о контроле.

Слабые места: В большинстве случаев данные функции создаются и настраиваются поставщиком DLP и не могут меняться, переопределяться пользователем.

7. Категоризация

Данный метод использует предварительно созданные категории с правилами и словарями для общих типов уязвимых данных.

Сильные места: Легкая настройка. Экономит время генерации политики. Для многих организаций категории могут встретить большой процент своих потребностей защиты данных.

Слабые места: Подходит только для легко категоризированных правил и содержания.

Рассмотренные методы формируют основу контент анализа для большинства продуктов DLP на рынке. Не все продукты включают все методы, и могут содержаться различия между реализациями. Большинство продуктов может также объединить методы в цепочку, создавая сложные комбинации.

Недостатки методов контентной фильтрации.

Независимо от конкретной используемой технологии суть контентной фильтрации остается неизменной. Каждый раз, проверяя исходящий документ, система пытается угадать, является ли он конфиденциальным. Как следствие, основной недостаток контентной фильтрации очевиден — он заключается в сравнительно невысокой точности всех современных алгоритмов.

Существуют две серьезные проблемы DLP-решений, построенных на технологиях контентной фильтрации. Первая проблема - невысокая точность фильтрации не позволяет обнаружить все конфиденциальные документы, покидающие корпоративную сеть. А также возможно высокое число ложных срабатываний системы, когда вполне легальные документы признаются строго секретными. И эти ложные срабатывания вполне способны вызвать настоящую панику у штатного офицера безопасности».

Кроме того, подавляющее большинство механизмов фильтрации является ресурсоемким и потому, как правило, реализуется на специальном сервере. Такой подход автоматически сопровождается проблемами, связанными с копированием информации на различные мобильные носители (прежде всего

флэшки). Теоретически для фильтрации такого трафика можно использовать локальные агенты, передающие информацию на сервер, однако на практике этот метод малоэффективен и труднореализуем.

Литература

1. Understanding and Selecting a Data Loss Prevention Solution, Securosis, L.L.C., 2011.
2. Data Leakage Detection, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. Vol. 23. №. 1, January 2011.