МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра теоретических основ компьютерной безопасности и криптографии

Разработка системы обнаружения и предотвращения утечек

АВТОРЕФЕРАТ
дипломной работы

студента 6 курса 631 группы
специальности 10.05.01 Компьютерная безопасность
факультета компьютерных наук и информационных технологий
Пронина Никиты Евгеньевича

Научный руководитель
старший преподаватель _____ А. А. Лобов
22.01.2024 г.

Заведующий кафедрой
д. ф.-м. н., доцент _____ М. Б. Абросимов
22.01.2024 г.

Саратов 2024

ВВЕДЕНИЕ

Каждый день компании по всему миру вынуждены защищать свою информацию от различного рода утечек. Их последствия могут быть самыми разными: начиная от репутационных потерь и заканчивая банкротством. Данные атаки могут быть исходить как извне, так и изнутри компании. Чтобы защититься от них, компании прибегают как разного рода средствам защиты, одними из которых являются DLP системы – системы, предотвращающие потери данных.

Целью данной работы является, изучение DLP систем, методов поиска утечки данных в текстовых сообщениях, а также построение программной модели имитирующей работу DLP системы, которая ориентирована на анализ электронной почты.

Дипломная работа состоит из введения, 3 разделов, заключения, списка использованных источников и 7 приложений. Общий объем работы – 71 страница, из них 40 страниц – основное содержание, включая 28 рисунков и 0 таблиц, список использованных источников из 20 наименований.

КРАТКОЕ СОДЕРЖАНИЕ

1 О DLP Системах

В данной главе приводится общая информация о DLP системах

1.1 Описание DLP системы

Наилучшим техническим вариантом для предотвращения утечки данных является применение систем класса DLP (Data Loss/Leakage Prevention). Они контролируют все наиболее вероятные каналы утечки позволяют идентифицировать информацию самыми современными способами, что обеспечивает наименьшее количество ложных срабатываний [1].

1.2 Классификация DLP систем

В зависимости от критериев классификации DLP системы делятся на несколько классов. По локализации (сетевой архитектуре) DLP-системы подразделяются на хостовые и шлюзовые.

По механизму определения степени конфиденциальности передаваемых данных выделяют два вида DLP-систем:

• Системы, которые устанавливают конфиденциальность на основе анализа маркеров документа,

• Системы, которые для этого проводят анализ содержимого документа.

С точки зрения *правомерности отслеживания действий* пользователя можно выделить легальные и нелегальные DLP-системы.

По способности блокирования информации, опознанной как конфиденциальная, выделяют системы с активным и пассивным контролем действий пользователя.

1.3 Принцип работы DLP системы

Современная система защиты от утечки информации, как правило, является распределённым программно-аппаратным комплексом, состоящим из большого числа модулей различного назначения. Часть модулей

функционирует на выделенных серверах, часть — на рабочих станциях сотрудников компании, часть — на рабочих местах сотрудников службы безопасности.

Выделенные сервера могут потребоваться для таких модулей как база данных и, иногда, для модулей анализа информации. Эти модули, по сути, являются ядром и без них не обходится ни одна DLP-система.

1.4 Примеры существующих DLP систем

В данной подглаве были рассмотрены следующие существующие DLP системы:

1. InfoWatch ;
2. SearchInform ;
3. Ростелеком-Солар;
4. Zecurion;
5. Falcongaze.

Вывод

В данной главе было проделано следующее:

1. Дано определение DLP системы;
2. Описана классификация DLP систем относительно четырех критериев;
3. Описан принцип работы DLP системы, расположение её модулей внутри корпоративной сети, а также назначение каждого из модулей;
4. Рассмотрены и оценены пять существующих DLP систем.

2 Методы анализа

В данной главе описаны основные методы анализа текста, используемые в DLP системах.

2.1 Лингвистический метод анализа

Лингвистический метод анализа – метод, который работает напрямую с содержанием файла и документа.

Данный метод включает в себя два вида анализа текста: морфологический и семантический.

1) Морфологический анализ предполагает дословный и описательный разбор конкретной части текста. Он призван расчленить исследование на более мелкие составляющие и определить суть, роль каждого элемента в нем.

На практике выделяют два вида морфологического анализа:

1) Дословный;

2) Попредложный.

2) Семантический анализ — этап в последовательности действий алгоритма автоматического понимания текстов, заключающийся в выделении семантических отношений, формировании семантического представления текстов.

В общем случае семантическое представление является графом, семантической сетью, отражающим бинарные отношения между двумя узлами – смысловыми единицами текста.

2.2 Статистический метод анализа

Статистический метод анализа – метод анализа, в котором исследуются качественные и количественные характеристики текста, такие как объем текста, частота встречаемых символов и слов, процентное соотношение частей речи и так далее. При этом смысловая нагрузка текста игнорируется. В данном методе можно выделить два подметода: графематический и контент анализ.

1) *Графематический анализ* – метод, создающий базу для

последующего морфологического и синтаксического анализа, на основе выделения слов, цифровых комплексов, формул и т.д. анализ направлен на разбивку текста на слова, разделители и т.д., сборку слов, написанных в разрядку, выделение устойчивых оборотов, фамилии, имени, отчества, даты и т.п., выделение электронных адресов и имен файлов, выделение предложений из входного текста абзацев, заголовков, примечаний [9].

2) *Контент-анализ* – это метод сбора данных и анализа содержания текста. Слово «контент» (содержание) имеет отношение к словам, рисункам, символам, понятиям, темам или же иным сообщениям, которые могут быть объектом коммуникации. Слово «текст» означает нечто написанное, видимое или произнесенное, которое выступает как пространство коммуникации.

Контент-анализ позволяет исследователю выявлять содержание в источнике коммуникации. Он позволяет поэкспериментировать с содержанием и рассмотреть его с использованием методов, отличных от обычного прочтения книги или просмотра телевизионной программы. С помощью контент-анализа исследователь может сравнить содержание множества текстов и анализировать его с помощью количественной методики (например, диаграмм, таблиц).

Недостаток статистического метода в том, что алгоритм не способен самостоятельно обучаться, формировать категории и типизировать. Как следствие – зависимость от компетенций специалиста и вероятность задания хеша такого размера, при котором анализ будет давать избыточное количество ложных срабатываний.

Вывод

В данной главе были подробно описаны два метода анализа текста – лингвистический и статистический. Лингвистический метод анализа текста оценивает содержание документа, определяя роль каждого элемента текста, его связь с другими элементами текста и выделяя смысловую составляющую всего текста. Данный метод даёт точную оценку при поиске утечки данных, однако ресурсозатратен и сложен в реализации.

Статистический метод анализ оценивает содержание документа относительно качественных и количественных характеристик текста, игнорируя смысловую нагрузку. Данный метод прост в реализации и не требует большого количества ресурсов, однако чувствителен к настройке параметров, а также компетенций специалиста, в связи с чем может выдавать некорректный результат.

3 Программная реализация

В данной главе будет рассмотрена программа, которая имитирует работу DLP системы. Данная программа написана на языке Python. При написании программы использовались следующие библиотеки:

1. NLTK – для обработки и анализа текста;
2. SQLite – для создание и ведения базы данных;
3. Response – для создания и обработки http запросов;
4. Docx – для обработки .docx файлов.

3.1 Описание программы

Данная программа моделирует процесс общения пользователей. Каждый пользователь имеет свой уровень доступа, который варьируется от 1 до 10. Каждый документ в системе также имеет уровень доступа, аналогичный пользовательскому. После отправки сообщения, программа анализирует его содержание и сравнивает с архивом известных документов, а также с таблицы запрещенных комбинаций слов. В случае нахождения запрещенных комбинаций слов или высокого совпадения с документом, на который у одного из пользователей не хватает прав, система блокирует отправку сообщения. Также каждое сообщение проверяется лингвистическим методом, и, в случае высокого совпадения с запрещенной информацией, администратор получает уведомление с предупреждением о возможной утечке.

Вся информация, которая необходима для функционирования программы, хранится в базе данных. Она состоит из 8 таблиц: «Пользователи», «Файлы», «Журнал событий», «Точность», «Сообщения», «Файлы из сообщений», «Список доступа к файлам» и «Список запрещенных комбинаций слов».

3.2 Работа программы

Для оценки корректности работы алгоритмов программы были созданы следующие условия:

1. Добавлены в базу данных пользователей user5 и user8 с уровнями

доступа 5 и 8 соответственно;

2. Добавлены в базу данных два документа с уровнем доступа 6, текстыдокументов перечислены в приложениях Е и Ж соответственно;

3. Разрешены пользователю user5 доступ к документу из приложения Е;

4. Добавлены фразу «bad person» в список запрещенных фраз;

Проведены следующие эксперименты:

1. Отправка в теле сообщения текста документа, к которому у одного из пользователей нет доступа;

2. Отправка во вложенном файле текста документа, к которому у одного из пользователей нет доступа;

3. Отправка запрещенной фразы в теле сообщения;

4. Отправка запрещенной фразы во вложенном файле;

5. Отправка в теле сообщения текста, похожего по смыслу на текст документа, к которому у одного из пользователей нет доступа;

6. Отправка во вложенном файле текста, похожего по смыслу на текст документа, к которому у одного из пользователей нет доступа;

7. Отправка текста содержащего информацию в теле сообщения или во вложенном файле из документа, к которому пользователю user5 выдали доступ;

Вывод

В данной главе была описана программная реализация DLP системы, настроенной на анализ утечки данных в текстовых сообщениях. Принцип работы программы, а также методы анализа текста, используемые при реализации, основаны на информации, приведенной в первой и второй главе. Был проведены эксперименты с целью определения наличия или отсутствия утечки информации в сообщении. Результаты экспериментов подтвердили корректность работы алгоритмов.

ЗАКЛЮЧЕНИЕ

В данной работе были рассмотрены различные классы DLP систем, а также изучены методы работы, при помощи которых DLP системы обнаруживаюти предотвращают утечку конфиденциальных данных.

Также были рассмотрены различные методы анализа текста, которые используются при поиске утечки конфиденциальных данных в текстовых сообщениях.

Была разработан алгоритм, который анализирует текст сообщения, а также содержимое вложенного файла формата .docx двумя методами, на основе результатов которых делается вывод о наличии либо отсутствии утечки конфиденциальной информации:

1. Морфологический анализ текста
2. Семантический анализ текста

Для демонстрации алгоритма была разработана программа с графическим интерфейсом. В ходе демонстрации программы алгоритм подтвердил свою работоспособность.

Таким образом, все поставленные задачи выполнены, цели работы достигнуты.