Министерство образования и науки Российской Федерации
Московский физико-технический институт
(национальный исследовательский университет)

Физтех-школа радиотехники и компьютерных технологий
Кафедра системного программирования ИСП РАН

Выпускная квалификационная работа бакалавра

Автоматическая генерация сигнатур сетевых протоколов и приложений

Автор:
Студент Б01-009а группы
Дурнов Алексей Николаевич

Научный руководитель:
канд. физ.-мат. наук
Гетьман Александр Игоревич

Москва 2024

Аннотация

Автоматическая генерация сигнатур сетевых протоколов и приложений

Дурнов Алексей Николаевич

Краткое описание задачи и основных результатов, мотивирующее прочитать весь текст

Содержание

1 Введение

В этой части надо описать предметную область, задачу из которой вы будете решать, объяснить её актуальность (почему надо что-то делать сейчас?). Здесь же стоит ввести определения понятий, которые вам понадобятся в постановке задачи.

Интернет-провайдеры и сетевые администраторы хотят идентифицировать тип сетевого трафика, который проходит через их сеть, для того, чтобы предоставлять своим клиентам лучший сервис, предлагая им высокое качество обслуживание (QoS), а также планировать свою инфраструктуру и управлять ею. Сетевой трафик можно классифицировать как в соответствии с использующимся протоколом, так и в соответствии с использующимся приложением. Вторая классификация более трудоёмкая, но позволяет решать более широкий спектр задач: формирование трафика в сети, улучшение качества обслуживание, а также предоставление более детального биллинга.

В области классификации сетевого трафика было проведено множество исследований, что привело к разработке многих методов. Самый наивный метод классификации сетевого трафика это идентификация по номеру порта. Однако современные приложения используют динамическое распределение портов и туннелирования трафика, например, по протоколу HTTP, данный метод даёт очень плохие и неточные результаты. Чтобы преодолеть эти ограничения идентификации были введены более совершенные методы.

Первый подход основан на сопоставлении сигнатур полезной нагрузки. Сигнатура полезной нагрузки - это часть данных полезной нагрузки, которая является статичной и различимой для приложений и может быть описана, как последовательность строк или шестнадцатиричных чисел. Второй подход основан на алгоритмах машинного обучения. Для этого подхода используются такие признаки потоков и пакетов, как задержки между пакетами, размеры пакетов и другие, а полезная нагрузка пакетов не анализируется, поэтому он менее точен, чем сигнатурный подход. Однако он может применяться для зашифрованного сетевого трафика, так как в приближении полезная нагрузка зашифрованного трафика представляет собой белый шум, поэтому сигнатурный подход не возможен. Также существуют и гибридные методы, которые используют эти два подхода вместе.

Методы, которые анализируют полезную нагрузку пакетов, являются очень эффективными и точными для идентификации трафика. Их часто называют глубокой проверкой пакетов (DPI). DPI является очень трудоемким и ресурсоемким процессом. Система DPI должна искать сигнатуры в полезной нагрузке пакетов, чтобы точно классифицировать поток. Такой подход не нов, системы обнаружения вторжения (IDS), основанные на DPI, с помощью сигнатур находят интернет-червей и другой трафик, угрожающий безопасности. Далее рассматриваемые методы будут сосредаточены на идентификации приложений среди безобидного трафика.

По началу сигнатуры извлекались вручную. Постоянное появление новых приложений и их частые обновления подчёркивают необходимость автоматической генерации сигнатур, так как ручная операция извлечения сигнатур занимает много времени, а также может быть разница в качестве сигнатур в зависимости от оператора извлечения. Методы автоматической генерации сигнатур должны быть основаны не на семантическом анализе протоколов, так как хотя они и повышают точность сигнатур, но не могут быть применены к анализу высокоскоростного трафика в режиме реального времени.

Поэтому данная работа посвещена исследованию различных методов автоматической генерации сигнатур полезной нагрузки для классификации сетевого трафика по

протоколам и приложением.

2 Постановка задачи

Здесь надо максимально формально описать суть задачи, которую потребуется решить, так, чтобы можно было потом понять, в какой степени полученное в результате работы решение ей соответствует. Текст главы должен быть написан в стиле технического задания, т.е. содержать как описание задачи, так и некоторый набор требований к решению

3 Обзор существующих решений

Здесь надо рассмотреть все существующие решения поставленной задачи, но не просто пересказать, в чем там дело, а оценить степень их соответствия тем ограничениям, которые были сформулированы в постановке задачи.

4 Исследование и построение решения задачи

Здесь надо декомпозировать большую задачу из постановки на подзадачи и продолжать этот процесс, пока подзадачи не станут достаточно простыми, чтобы их можно было бы решить напрямую (например, поставив какой-то эксперимент или доказав теорему) или найти готовое решение.

5 Описание практической части

Если в рамках работы писался какой-то код, здесь должно быть его описание: выбранный язык и библиотеки и мотивы выбора, архитектура, схема функционирования, теоретическая сложность алгоритма, характеристики функционирования (скорость/память).

6 Заключение

Здесь надо перечислить все результаты, полученные в ходе работы. Из текста должно быть понятно, в какой мере решена поставленная задача.