А. А. Мачуев, А. П. Дураковский

АВТОМАТИЗАЦИЯ АНАЛИЗА И КЛАССИФИКАЦИИ ФАЙЛОВОГО МАТЕРИАЛА

Тот ли файл перед нами, можно определить, зная его структуру, а по известной структуре можно произвести классификацию. Гораздо большего внимания подобный подход заслуживает, если речь идет не об общепринятых типах файлов, с которыми мы сталкиваемся в повседневной жизни, а когда перед нами данные совершенно неизвестного происхождения и мы не можем судить о каких-либо параметрах этих данных без детального «ручного» анализа.

Анализ данных, как было сказано выше, «вручную» представляет собой весьма трудоемкий процесс. Для автоматизации и ускорения этого процесса разработано специальное программное обеспечение (СПО). Разработка и тестовые испытания СПО производились на примере файлов с общеизвестными расширениями (см. рисунок 1), особенности структуры которых известны заранее [1]. Дальнейший этап использования СПО — анализ неизвестных данных и, возможно, иные прикладные задачи.

Рисунок представляет собой структуру файлов с расширением .EXE. Под структурой файла понимается набор значений байт файла. Для удобства восприятия человеком структура файла отображается в виде графика. На графике по оси абсцисс отложен номер байта — смещение данного байта относительно начала файла, по оси ординат — значение байта. Совершенно иначе будут выглядеть графики для архивных файлов, файлов с изображениями или звуком. Файлы с различными расширениями характеризуются каждый своей особой, характерной только для данного расширения, структурой.

Далее в работе решается задача по методам классификации различных структур известных расширений. По графическому представлению можно определить совпадение нескольких структур. Компьютерную информацию, представляющую структуру файла в виде массива байт, можно описать в виде ряда статистических величин: математического ожидания, дисперсии, энтропии. На рисунке 1 также приведено графическое представление структуры файла с рассчитанным значением математического ожидания (на графике — жирная кривая).

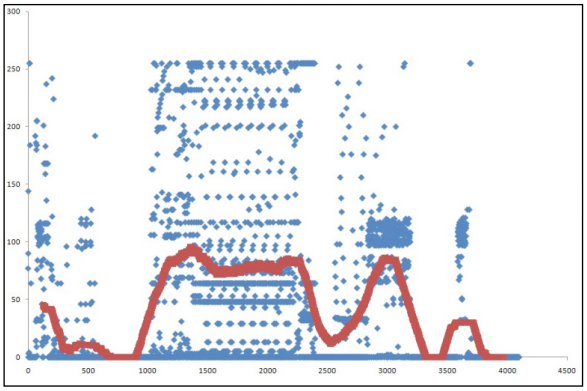

Рис. 1. Структура .EXE файла со значением математического ожидания

Далее по полученным статистическим величинам происходит выделение характерных зон в файле. Характерная зона представляет собой набор точек со схожими свойствами, например по значению байт.

Исходя из свойств зон им назначаются условные типы: 1, 2, 3… таким образом, что зональная картина для примера выше принимает вид: 1 2 3 2 3 2. В данном случае тип зоне присваивается по значению точек.

Полученные последовательности сравниваются при помощи алгоритмов глобального выравнивания (Needleman-Wunch) и локального выравнивания (Smith-Waterman) [2].

Данные алгоритмы дают численную оценку соответствия сравниваемых файлов. В результате если на вход СПО подать файлы с известной структурой и файлы неизвестной природы, то на основании предложенных известных файлов СПО определит тип анализируемых неизвестных файлов.

СПИСОК ЛИТЕРАТУРЫ:

1. Форматы файлов. URL: http://www.filetypes.ru/docx.
2. Pairwise Sequence Alignment. URL: http://web.cecs.pdx.edu/~ps/CapStone03/SimilarityDiscussion.html.

А. Н. Мироненко, С. В. Белим

МНОГОУРОВНЕВАЯ СИСТЕМА ФИЛЬТРАЦИИ СПАМА В ПОТОКЕ ЭЛЕКТРОННОЙ ПОЧТЫ

Система имеет два уровня фильтрации: анализ формальных признаков сообщения с запросом повторной отправки и контентный, основанный на двухслойной сети формальных нейронов. Оба метода описаны ниже.

Первый уровень фильтрации. Заголовок сообщения содержит 12 ключевых полей [1]. Наиболее важными и представляющими интерес являются следующие:

1. From — адрес отправителя;

2. Message-ID — уникальный идентификатор сообщения. Состоит из адреса узла-отправителя и номера (уникального в пределах узла);

3. In-Reply-To — указывает на Message-ID, для которого это письмо является ответом (с помощью этого почтовые клиенты могут легко выстраивать цепочку переписки — каждый новый ответ содержит Message-ID для предыдущего сообщения);

4. Subject — тема письма.

Для данного метода распознавания спама важны поля From, Message-ID и Subject, Date. Алгоритм работы приведен ниже.

1. По POP3 [2] сообщение сохраняется на ПК. Выделяется заголовок, в нем ключевые поля From: (*АдресОтправителя*), Subject: (*ТемаСообщения*), Message-ID (*IDСообщения*), Date: (Дата попадания сообщения в карантин).

2. Проводится предварительная фильтрация сообщения по спискам.

Если (*АдресОтправителя* входит в *БелыйСписок*),

тогда (сообщение не спам, и оно перемещается во входящие).

Иначе

Если (*АдресОтправителя* входит в *ЧерныйСписок*),

тогда (сообщение спам, и оно удаляется).

Иначе (сообщение помещается в карантин) и (Шаг 3).

3. Создается таблица сообщений, попавших в карантин, следующего вида: