В практическом плане разработанный MMScraper позволит составить экспериментальную базу для исследования задач интеллектуальной обработки текста.

Нам видится важным продолжить исследования результатов краулинга и реализации дополнительных возможностей MMScraper, улучшающих результаты работы на очень крупных сайтах всемирной паутины. Осуществление таких возможностей предусмотрено расширяемой архитектурой краулера.

Список использованных источников:

1. A.H.F. Laender, B. A. Ribeiro-Neto, Juliana S.Teixeria. A brief survey of web data extraction tools // ACM SIGMOD Record 31(2), pp 84-93. 2002
2. Baeza-Yates R., Castillo C. Crawling the Infinite Web: Five Levels are Enough // Lecture Notes in Computer Science. Algorithms and Models for the Web-Graph, Third International Workshop. 2004. Vol. 3243. P. 156–167.

УНИКАЛЬНАЯ ИДЕНТИФИКАЦИЯ ЦИФРОВЫХ СИСТЕМ НА ОСНОВЕ ДОЗУ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Пучков А. В.

Иванюк А. А. – д-р. техн. наук, доцент

В современных встраиваемых системах практически обязательным компонентом является оперативное запоминающее устройство (ОЗУ), часто динамическое (ДОЗУ). Помимо своего основного назначения ДОЗУ может быть использовано в качестве криптографического примитива, обеспечивающего уникальную неклонируемую идентификацию цифровой системы.

Криптографические примитивы, в основе которых лежит структурная сложность физических, в частном случае электронных, систем, являются предметом изучения физической криптографии, одним из ключевых понятий которой является физически неклонируемая функция (ФНФ). В контексте цифровых систем ФНФ в общем случае можно рассматривать как устройство, задача которого состоит получении значений ответов на подаваемые входные воздействия – запросы, при этом пары запрос-ответ являются уникальными, непредсказуемыми и неклонируемыми на других экземплярах интегральных схем, выпущенных в рамках конкретного технологического процесса [1]. Функционирование таких устройств основывается на том факте, что в процессе производства цифровых систем принципиально невозможно управлять значениями отдельных их параметров, которые принимают уникальные для конкретного экземпляра цифровой системы значения. Основной задачей ФНФ является извлечение таких параметров, например,задержек распространения сигналов или различий результирующих частот идентичных тактовых генераторов. Уникальность пар запрос-ответ ФНФ позволяет использовать их как криптографический примитив в методах неклонируемой идентификации и аутентификации цифровых систем.

Статические и динамические оперативные запоминающие устройства характеризуются тем, что при включении питания часть запоминающих ячеек находится в состоянии 0, в то время как другая часть – в состоянии 1, причем распределение таких ячеек в общем случае случайно и статистически уникально, что позволяет рассматривать ОЗУ как ФНФ, где запросом является включение питания, а ответом – содержимое массива запоминающих элементов после включения питания [2]. Практическое применение такой ФНФ вызывает затруднение в той связи, что ее использование может быть произведено лишь всякий раз при включении питания ОЗУ, что в подавляющем большинстве случаев означает включение цифровой системы.В случае ДОЗУ альтернативным подходом может являться отключение регенерации части запоминающих элементов, что может быть неоднократно выполнено в процессе функционирования цифровой системы [2]. Стоит отметить, что ДОЗУ получает все более широкое распространение, в том числе во встраиваемых системах из-за своей сравнительно низкой стоимости.

Для экспериментального исследования данного метода использовались интегральные схемы ДОЗУMicron M45W8MW16, которыми были оснащены имеющиеся в наличии 10 плат быстрого прототипирования на основе FPGA Digilent Nexys 4. Данные интегральные схемы ДОЗУ имеют объем 128 Мбит и обладают интерфейсом, схожим с ОЗУ, т.е. обеспечение регенерации запоминающих элементов достигается использованием дополнительной логики, которой можно удобно управлять при помощи доступных пользователю регистров.

Был разработан аппаратно-программный комплекс, позволяющий осуществлять управление данными ОЗУ с рабочей станции, устанавливая различные режимы работы, а также производить чтение региона ДОЗУ для последующего анализа. Задачей экспериментов было оценить различные методы извлечения уникальности интегральных схем ДОЗУ.

В ходе первого из экспериментов верхняя половина адресного пространства ОЗУ была заполнена константными значениями '1', затем была отключена регенерация запоминающих элементов этого региона. Далее производилось многократное чтение содержимое указанной области с целью последующего анализа изменений, которые происходят в ней с течением времени. Выбор значения '1' обусловлен физическими процессами, проходящими в запоминающей ячейке, а именно '1' соответствует присутствию заряда, который в

отсутствии регенерации будет утекать из ячейки с течением времени. Экспериментальные данные показали, что изменение ячеек быстро насыщается, т.е. образуется некоторое распределение единичных и нулевых бит, слабо изменяющееся во времени (рисунок 1).

Альтернативный сценарий состоит в выключении регенерации ячеек без последующего их чтения, а затем включения регенерации с чтением результирующего содержимого ДОЗУ. Однако на рассматриваемых интегральных схемах ДОЗУ это приводит к формированию детерминированного распределения значений ячеек, что не позволяет построить алгоритм идентификации по этому сценарию (рисунок 2).

Наконец, самый простой подход, состоящий в выключении питания схемы ДОЗУ либо во введении ее в режим пониженного энергопотребления (что достигается, главным образом, отключением регенерации), если таковой поддерживается конкретной интегральной схемой, имеет очевидный недостаток, связанный с тем, что его невозможно применить во время нормального функционирования устройства, т.к. данные из памяти теряются.

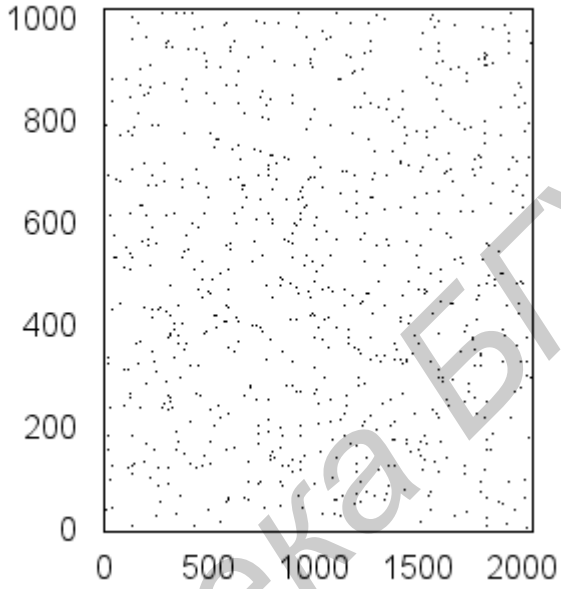

Рис. 1 – Фрагмент карты состояний запоминающих ячеек после многократного чтения без регенерации

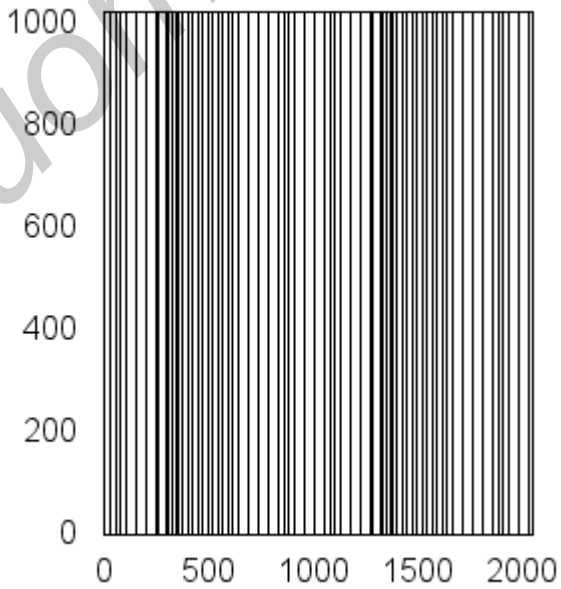

Рис. 2 – Фрагмент карты состояний запоминающих ячеек после включения регенерации без чтения

Тем не менее, распределение, экспериментально полученное с помощью чтения данных ОЗУ, выведенного из режима пониженного энергопотребления, схоже с тем, что было получено в предыдущем сценарии. Данные же, полученные после реального выключения питания, имеют характер случайных последовательностей, различающихся на разных интегральных схемах ДОЗУ.

В ходе первичных экспериментов было показано, что при помощи описанных сценариев можно осуществлять как извлечение уникальных характеристик каждой схемы ДОЗУ, так и производить генерацию истинно случайных последовательностей.

Список использованных источников:

1. Lao, Y. Reliable PUF-Based Local Authentication With Self-Correction / Y. Lao, B. Yuan, C. H. Kim, K. K. Parhi // IEEE transactions on Computer-Aided Design of Integrated Circuits and Systems. –2017. – Vol. 36, № 2. –P. 201-213.
2. Tehranipoor, F. DRAM based Intrinsic Physical Unclonable Functions for System Level Security / F. Tehranipoor, N. Karimian, K. Xiao, J. Chandy // Proc. of the 25th edition on Great Lakes Symposium on VLSI. – Pittsburgh, Pennsylvania, USA, 2015. – P. 15-20.

КЛАССИФИКАЦИЯ ТОНАЛЬНОСТИ ТЕКСТОВЫХ ДОКУМЕНТОВ С ПОМОЩЬЮ МЕТОДА ОПОРНЫХ ВЕКТОРОВ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Романов А.А.
Пилецкий И. И. – канд. физ.-мат. наук, доцент

С развитием сети Интернет, объём информации, создаваемой человечеством, значительно увеличился. Информация накапливается в различных источниках, таких как социальные сети, форумы, площадки для отзывов, блоги и новостные сайты, причём преимущественно она хранится в виде текстовых данных. Большой объём и слабая структурированность таких данных определяют необходимость создания систем автоматической обработки. Одной из актуальных задач обработки, является определение тональности текстовых документов. Определение тональности текстов востребовано, например, при анализе эффективности рекламных кампаний, сборе откликов о проведённых мероприятиях, определении репутации брендов, построении системы поддержки пользователей.

Текст на естественном языке может нести в себе не только информацию, но и её эмоциональную оценку. Эмоциональная оценка, выраженная в текстовом документе, называется тональностью или сентиментом (англ. sentiment – чувство, настроение). В понятии машинного обучения, задача определения эмоциональной оценки текста сводится к задаче классификации. В формальном виде задача классификации определяется следующим образом [1]. Пусть существует описание документа $d \in X$, где X – векторное пространство документов, и конечное множество классов $C = \{c_1, c_2, \dots, c_j\}$. Из множества документов с заранее известными классами $D = \{(d, c)$, где $(d, c) \in X \times C\}$, используя обучающий алгоритм, необходимо получить классифицирующую функцию γ, которая отображает документы в классы $\gamma: X \rightarrow C$. В решаемой задаче определения тональности множество C состоит из двух элементов: положительной и отрицательной эмоциональной оценки.

Задачу классификации на два класса успешно решают с использованием различных методов машинного обучения. Для применения алгоритмов машинного обучения текстовый документ необходимо представить в виде математического вектора. В качестве векторной модели используется «мешок термов» [1]. Текст в данной модели рассматривается как неупорядоченное множество термов. Термом может являться любое символьное выражение текста, например, слова, словосочетания, знаки пунктуации. Каждому терму сопоставляется некий вес. Вектор же формируется при упорядочивании всех уникальных термов в пространстве. Размерность вектора определяется числом уникальных термов во всей коллекции и является постоянной для всех документов.

Перед взвешиванием документов проводится предварительная очистка коллекции: приведение всех символов текстовых документов к нижнему регистру и удаление пунктуационных знаков. Далее из коллекции извлекаются уникальные термы, для каждого из которых рассчитываются статистические данные, необходимые для построения весовых схем. При необходимости, для повышения результатов классификации настраивается система фильтров, используя извлеченную статистику. Для взвешивания опробованы различные весовые схемы: бинарная, TF, TF-IDF, TF-RF и др. [2-5].

Для решения задачи был выбран один из алгоритмов машинного обучения с учителем – метод опорных векторов (supportvectormachine, SVM). Выбор данного алгоритма основан на его высокой точности в решении задач бинарной классификации коллекций текстовых документов различных тематик [6]. Основная идея метода SVM – поиск разделяющей гиперплоскости, максимально удалённой от ближайших к ней точек обоих классов в пространстве признаков [7]. В качестве ядра SVM взято линейное ядро, как наиболее эффективное при больших размерностях векторов и большом количестве объектов для обучения [8].

В качестве оценок качества результатов обучения и работы алгоритма выбраны четыре общепринятые характеристики: accuracy, precision, recall и f-measure. Для несмещенной оценки вероятности ошибки и избегания проблемы переобучения используется кросс-валидация по 5 блокам.

В процессе исследования тестируются различные коллекции документов на английском и русском языках. Коллекции содержат сотни тысяч документов и сопоставимое количество уникальных термов. Для эффективной обработки такого объёма информации был выбран фреймворк ApacheSpark. Библиотека MLlib данного фреймворка поддерживает реализацию метода опорных векторов с линейным ядром [9]. Используя данную библиотеку, на всех тестируемых коллекциях удалось достичь показателей качества более 80%.