Интернет-журнал «Науковедение» ISSN 2223-5167 http://naukovedenie.ru/

Том 9, №3 (2017) http://naukovedenie.ru/vol9-3.php

URL статьи: http://naukovedenie.ru/PDF/52TVN317.pdf

Статья опубликована 22.06.2017

Ссылка для цитирования этой статьи:
Лапко А.Н., Рябоконь В.В., Куцакин М.А., Григорян Д.Р., Шиндряев А.В. Экспериментальное исследование метода идентификации массивов бинарных данных // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 9, №3 (2017) http://naukovedenie.ru/PDF/52TVN317.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.

УДК 004.67

Лапко Александр Николаевич
ФГКВОУ ВО «Академия Федеральной службы охраны Российской Федерации», Россия, Орёл[1]
Сотрудник
Кандидат технических наук
E-mail: lan46@mail.ru

Рябоконь Владимир Владимирович
ФГКВОУ ВО «Академия Федеральной службы охраны Российской Федерации», Россия, Орёл
Сотрудник
Кандидат технических наук
E-mail: mimicria@mail.ru
РИНЦ: http://elibrary.ru/author_profile.asp?id=860017

Куцакин Максим Алексеевич
ФГКВОУ ВО «Академия Федеральной службы охраны Российской Федерации», Россия, Орёл
Сотрудник
E-mail: max_kooks@mail.ru
РИНЦ: http://elibrary.ru/author_profile.asp?id=860545

Григорян Даниил Рубенович
ФГКВОУ ВО «Академия Федеральной службы охраны Российской Федерации», Россия, Орёл
Сотрудник
E-mail: daniil96grigoryan@yandex.ru

Шиндряев Алексей Владимирович
ФГКВОУ ВО «Академия Федеральной службы охраны Российской Федерации», Россия, Орёл
Сотрудник
E-mail: shialsha@mail.ru

Экспериментальное исследование метода идентификации массивов бинарных данных

Аннотация. В статье рассматривается задача идентификации нечетких дубликатов среди массивов бинарных данных в составе исходных текстов программного обеспечения. Предложен подход к снижению вычислительной сложности метода идентификации в условиях больших объемов исходных и эталонных данных на основе схемы независимых перестановок. Приведены результаты экспериментальной проверки предложенного метода идентификации, показавшие возможности его использования в технологическом процессе автоматизированного

[1] 302034, Россия, г. Орёл, ул. Приборостроительная, д. 35

контроля информационных объектов, а также его эффективность по отношению к существующим методам.

Вклад авторов. Рябоконь Владимир Владимирович - автор внес существенный вклад в организацию и проведение экспериментальной части. Собрал и проанализировал основные результаты, касающиеся времени идентификации массивов бинарных данных. Лапко Александр Николаевич - автор внес существенный вклад в разработку метода идентификации массивов бинарных данных, одобрил окончательную версию статьи перед ее подачей для публикации. Куцакин Максим Алексеевич - автор внес существенный вклад в исследовании, направленном на поиск и обоснование параметров представленного метода, осуществил написание статьи. - Григорян Даниил Рубенович - автор участвовал в поиске и анализе существующих подходов оценки сходства двух объектов. Шиндряев Алексей Владимирович - автор участвовал в поиске и анализе существующих подходов оценки сходства двух объектов.

Ключевые слова: идентификация; массивы бинарных данных; тематические исследования; большие данные; экспериментальное исследование

Введение

В современных условиях, обусловленных политико-экономическими тенденциями, направленными на реализацию стратегии импортозамещения в области инфокоммуникационных систем, в том числе и в интересах органов государственной власти и управления, использование программного обеспечения с открытым исходным кодом приобретает особую важность. Наличие открытого исходного кода в ключевых проектах системного и прикладного программного обеспечения обеспечивает возможность его использования в системах, обрабатывающих информацию различного уровня конфиденциальности, путем проведения соответствующих сертификационных испытаний с учётом требований современной нормативной базы в области информационной безопасности.

Между тем, одной из тенденций в разработке сложных проектов с открытым исходным кодом является широкое использование в их рамках проприетарных компонентов, например, драйверов устройств, оформленных в виде предварительно откомпилированных модулей - массивов бинарных данных. С точки зрения сертификационных испытаний программного обеспечения, идентификация таких информационных объектов как массивы, содержащие в своём составе бинарные данные, представляет несомненный интерес, поскольку содержимым данных массивов может являться исполняемый код с неизвестной и потенциально вредоносной функциональностью.

Прецедентный подход

В общем случае прецедентный подход базируется на методологии принятии решения по аналогии [2]. В [3] проведено обширное исследование использования методологии *case-based reasoning* (CBR) в различных предметных областях, в том числе в рамках автоматизации поддержки принятия решений в области надежности и безопасности сложных технических систем. Очевидно, что применение методологии, базирующейся на прецедентном подходе в процессе экспертного аудита информационных объектов программного обеспечения, реализуемого в ходе тематических исследований, может позволить сократить отводимые на него трудозатраты за счет повторного использования информационных объектов, аудит которых был проведен ранее.

Необходимость идентификации информационных объектов программного обеспечения связана с тем, что в исходных текстах программного обеспечения могут присутствовать

массивы бинарных данных полностью или частично соответствующие массивам, проанализированным ранее экспертным путем. Подобная ситуация является прецедентом и позволяет эксперту-аналитику повторно использовать данные о неизвестном массиве бинарных данных на основе его аналогов в репозитории и не выполнять этап его анализа.

При этом точно совпадающие по содержимому массивы бинарных данных (так называемые «полные дубликаты») могут оперативно идентифицироваться с использованием контрольных сумм. При идентификации «нечетких дубликатов» [4] задача идентификации сводится к сравнению последовательностей бинарных данных, отличающихся по своим размерам и содержанию, и получению меры их подобия.

Очевидно, что выбор метода идентификации или их сочетания зависит от конкретных задач, решаемых системой сбора, обработки и представления информации об объектах.

Методы оценивания подобия объектов

Прецедентный подход, реализуемый в рамках систем, поддерживающих деятельность испытательных лабораторий, основан на необходимости оценивания подобия (близости) найденного информационного объекта ранее проанализированным эталонным объектам, которые хранятся в соответствующем репозитории. Эта необходимость обусловлена гипотезой о том, что информационные объекты с одинаковой или близкой функциональностью, или структурой данных являются подобными, а, значит, с точки зрения проведения их анализа может быть применен подход повторного использования.

При этом следует отметить, что процесс оценивания подобия, в общем случае, существенно зависит от предметной области, в рамках которой рассматриваются оцениваемые объекты и при их сложной структуре этот процесс может быть достаточно сложным, в том числе и в вычислительном отношении.

Основной проблемой, связанной с реализацией процесса контроля информационных объектов на основе прецедентов, является снижение эффективности их извлечения по мере роста репозитория объектов. При этом методологический аппарат оценивания подобия массивов бинарных данных должен удовлетворять условию низкой вычислительной сложности, что обусловлено большими объёмами исходных данных и ограничениями на временной ресурс проведения тематических исследований.

Наиболее очевидной мерой подобия между двумя объектами является расстояние между выборками представляющих их данных, а одним из путей анализа их подобия является определение наиболее подходящей функции расстояния (метрики) и вычисление матрицы расстояния между парами всех выборок. В [5] представлены основные типы метрик, которые используются в задачах оценивания близости: евклидова метрика, мера сходства Хэмминга, вероятностная мера сходства, мера сходства Роджерса-Танимото, манхэттенская метрика, расстояние Чебышева, метрики Махалонобиса, Журавлева, Брея-Кертиса, Чекановского, Жаккара и др.

Так, например, в предметной области анализа двоичных последовательностей одной из самых распространенных метрик является мера сходства (расстояние) Хэмминга, то есть количество различающихся позиций для их содержимого. Расстояние Хэмминга широко используется в различных задачах распознавания, таких как поиск близких дубликатов, классификация документов, исправление ошибок, обнаружение вирусов и т.д.

Другим известным подходом является использование методов, применяемых для анализа сигналов в системах передачи информации, в области анализа дискретных сообщений [6]. В подобных методах в качестве признакового пространства предлагается использовать,

например, взаимную корреляционную функцию с лагом, учитывающим разницу в размерах анализируемых массивов.

В [7] проведено исследование проблем сравнения и классификации дискретных данных, а также предложен подход с использованием предварительного их сжатия на основе спектрального импульсного преобразования и сравнения данных путём оценки евклидова расстояния между полученными в результате сжатия значениями спектра.

В качестве отдельной группы следует выделить методы, базирующиеся на использовании хэш-функций. Их особенностью является поиск компромисса между вычислительной сложностью применяемых хэш-функций и качеством получаемых результатов, поскольку уменьшение вычислительных затрат неизбежно связано со слабостью полученных сигнатур, которая приведёт к большой вероятности ошибки при сравнении [8].

Очевидно, что вычислительная сложность представленных методов определяется произведением сложности вычисления метрики $O(n^2)$ и сложности построения матрицы расстояния между парами всех выборок $O(K)$, где K - количество сравниваемых эталонных образцов из репозитория, n - размер сравниваемых массивов. При существенном увеличении объема репозитория вычисление метрики по принципу «каждый с каждым» делает невозможным использование подобных методов для идентификации массивов бинарных данных из-за ограничений на временной ресурс проведения тематических исследований.

Метод идентификации с использованием независимых перестановок

В противоположность рассмотренным методам A. Broder предложил метод, основанный на представлении документа в виде последовательности перекрывающихся подстрок определенной длины [9], также известный как метод «шинглов». Метод шинглов базируется на гипотезе о том, что схожие документы имеют существенное количество одинаковых шинглов, то есть множества их шинглов существенно пересекаются.

При большом количестве шинглов подсчет мощности пересечения множеств нецелесообразен. С целью уменьшения вычислительной сложности метода расчет мощности пересечения множеств осуществляется не для полной таблицы шинглов, а некоторой её выборки, получаемой с помощью случайных перестановок строк таблицы (перемешиваний) [10]. Независимые перестановки осуществляются с использованием набора взаимно однозначных и независимых хэш-функций $h_i(S)$, применяемых к элементам двух множеств. При этом для каждой хэш-функции из набора выбирается только минимальное значение сигнатуры $h_i^{\min}(S)$, соответствующее определённому шинглу. При использовании независимых перестановок массив бинарных данных представляется в виде вектора, содержащего конечный набор минимальных значений сигнатур хэш-функций:

$$A' = \left[h_1^{\min}, h_2^{\min}, ..., h_n^{\min}\right], \quad (1)$$

где

$$h_i^{\min} = \min_j [h_{ij}]. \quad (2)$$

Модификация метода шинглов для идентификации массивов бинарных данных заключается в переходе от сравнения отдельных слов текста к сравнению отдельных блоков массива бинарных данных [11]. Схематично получение меры близости массивов бинарных данных на основе их разделения на шинглы - блоки байтов, представлено на рис. 1.

Для массивов A и B минимальные значения сигнатур хэш-функций совпадают тогда и только тогда, когда элементы, генерирующие эти минимальные значения, находятся в обоих массивах, вероятность их совпадения определяется выражением 3:

$$P(h_i^{\min}(A) = h_i^{\min}(B)) = \frac{|A \cap B|}{|A \cup B|} = J(A,B), \quad (3)$$

то есть равна коэффициенту сходства Жаккара [12].

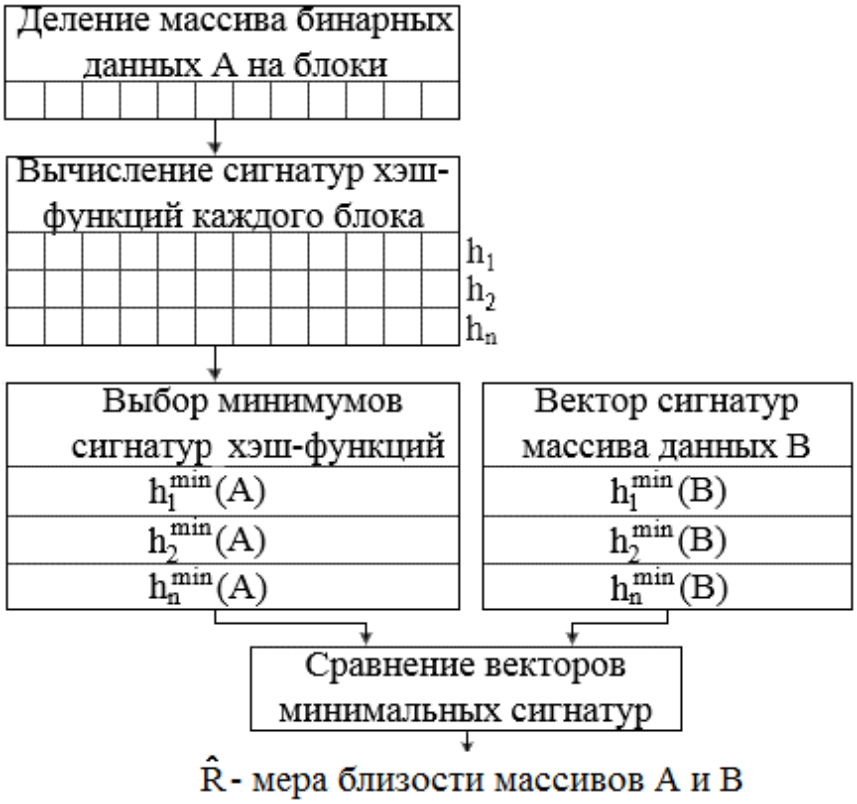

Рисунок 1. *Получение меры близости массивов бинарных данных (составлено авторами)*

При этом многократное применение различных хэш-функций количеством *n* для перестановок аналогично схеме повторных независимых испытаний Бернулли, в которой количество успешных наступлений события подчиняется биномиальному распределению.

В данном случае близости $\hat{R}$ представляет собой частоту наступления события совпадения минимальных значений для *n* хэш-функций, математическое ожидание полученной меры близости МБД описывается выражением 4, а её среднеквадратическое отклонение - выражением 5.

$$M_R = J(A,B), \quad (4)$$

$$\sigma_R = \sqrt{\frac{J(A,B)\cdot(1-J(A,B))}{n}}. \quad (5)$$

Выбор параметров метода

Метод независимых перестановок, применяемый для получения меры близости массивов бинарных данных, основан на использовании набора независимых хэш-функций и получения минимальных хэш-значений.

При этом получаемая мера близости подчиняется биномиальному распределению в случае, когда служащие для перестановок хэш-функции помимо независимости обладают свойством дискретного равномерного распределения результатов по всем возможным значениям. По результатам анализа алгоритмов некриптографических хэш-функций в качестве базовой для метода независимых перестановок выбрана функция, основанная на линейном конгруэнтном методе:

$$h_i(s_j) = (seed[i] \cdot h_i(s_{j-1}) + s_j) \bmod m, \quad (6)$$

где: s_j - байт данных, $seed[i]$ - коэффициент хэш-функции, m - значение модуля.

Данный выбор обусловлен низкой вычислительной сложностью, а также возможностью получения набора хэш-функций для независимых перестановок с помощью различных значений коэффициента функции $seed[i]$.

Кроме того, в работе [13] доказана гипотеза о равномерном распределении значений хэш-функции с использованием критерия согласия Пирсона. Результаты, представленные на рис. 2, показывают, что при достаточно большом количестве статистических испытаний хеш-функция вида (6) обладает высокой равномерностью хеширования.

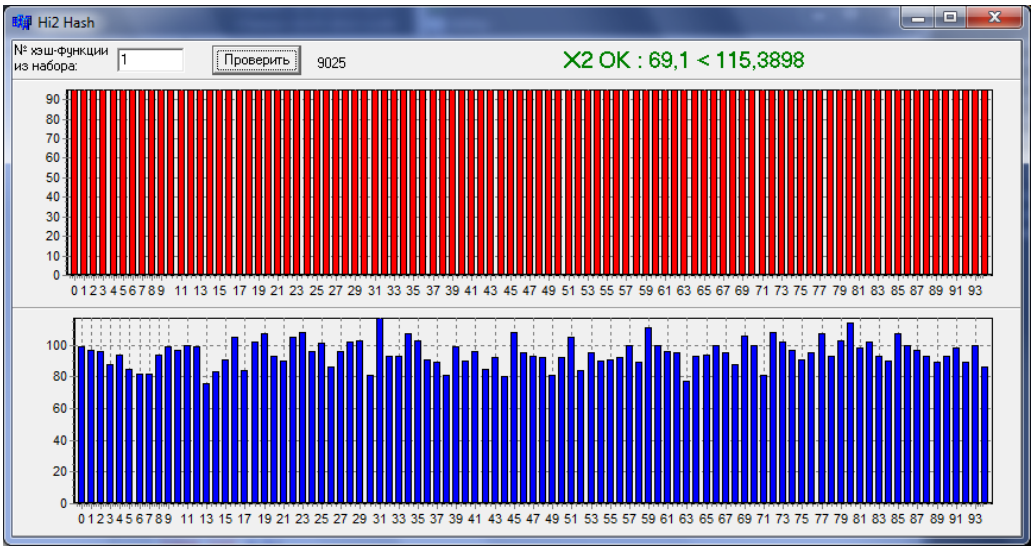

Рисунок 2. *Пример программы расчета по критерию χ^2 (составлено авторами)*

Для программной реализации алгоритма идентификации осуществлен выбор подходящие значения для размера блоков (W_b), на которые будет разбиваться массив бинарных данных, и количество хэш-функций n, используемых для независимых перестановок.

При малых значениях размера блока ($W_b < 10$) наблюдаются существенные отклонения от аналитически рассчитанных значений, а для минимального размера блока $W_b = 1$ (байт) математическое ожидание меры близости массивов близко к единице [14]. Это обусловлено высокой вероятностью совпадения коротких блоков байт массивов бинарных данных и малым перекрытием блоков.

Для задачи идентификации массивов бинарных данных использован размер блока W_b = 16 (байт), дальнейшее увеличение размера блока не оказывает влияния на точность получаемой меры близости, но приведёт к увеличению вычислительной сложности алгоритма.

На выбор количества хэш-функций влияет необходимая точность получаемой меры близости. С ростом количества испытаний, то есть с увеличением количества хэш-функций n, дисперсия меры близости стремится к нулю, а частота появления события в испытаниях - к истинной вероятности наступления события. Таким образом, точность идентификации возрастает пропорционально $\sqrt{n}$.

Для удобства расчётов при идентификации массивов бинарных данных с приемлемой точностью достаточно задать n = 100. Дальнейшее увеличение количества хэш-функций оказывает всё меньшее влияние на точность идентификации при существенном увеличении вычислительной сложности метода.

Оценивание эффективности метода

Экспериментальное исследование разработанного метода идентификации массивов бинарных данных проводилось с целью проверки возможности его использования в технологическом процессе автоматизированного контроля информационных объектов, а также установление его эффективности по отношению к существующим методам.

Эксперимент проводился в несколько этапов. На первом этапе оценивалась применимость разработанного метода идентификации в технологическом процессе автоматизированного контроля информационных объектов. При этом необходимо оценить точность идентификации массивов бинарных данных в зависимости от размера массивов.

Для построения такой зависимости фиксировались количество хэш-функций n = 100, количество совпадающих байт $S_b = 0.5 \cdot A_b$, размер блока $W_b = 16$, и генерировались случайным образом два тестовых массива размером $A_b = B_b = 64 \ldots 3\,072\,000$ (байт). Мера близости массивов вычислялась для 100 различных вариантов сгенерированных массивов данных каждого размера, результаты измерения представлены на рис. 3.

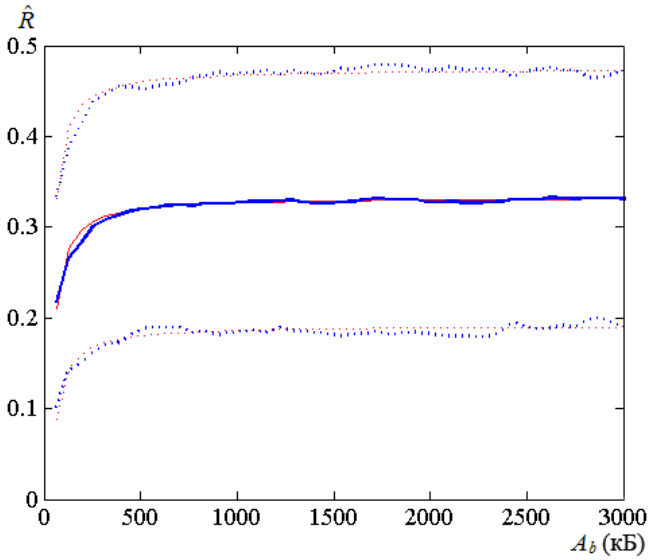

Рисунок 3. *Результаты измерений меры близости при изменяющемся размере массивов бинарных данных (составлено авторами)*

При этом аналитически рассчитанные мера близости и её среднеквадратическое отклонение совпадают со своими статистическими значениями на всём диапазоне измеренных значений.

Предполагается, что при дальнейшем увеличении размера массивов точность идентификации будет снижаться, однако на практике не удалось подтвердить это предположение вследствие ограничений на временной ресурс при проведении эксперимента.

Для проверки применимости разработанного способа идентификации был произведен анализ массивов бинарных данных в исходных текстах различных версий ядер ОС Linux, результаты представлены в таблице 1.

Таблица 1

Анализ количества и размеров массивов бинарных данных в выборке различных версий ядер ОС Linux (составлено авторами)

Номер версии	Количество МБД	Средний размер (байт)
Linux 3.7.1	5809	1465
Linux 3.8.1	5823	1463
Linux 3.9.1	5884	1463
Linux 3.10.1	5971	1461
Linux 3.11.1	6186	1458
Linux 3.12.1	6263	1465
Linux 3.13.1	6305	1456

Из табл. 1 видно, что количество массивов бинарных данных увеличивается с появлением каждой новой версии ядра, однако средний размер найденных массивов бинарных данных не превышает 1500 байт, а максимальный размер массива среди проанализированных составляет 252 641 байт.

Таким образом, разработанный метод идентификации может использоваться в технологическом процессе автоматизированного контроля информационных объектов.

На втором этапе для обеспечения возможности установления эффективности способа разработана процедура побайтного сравнения массивов бинарных данных, осуществляющая поиск максимального количества совпадающих байт при всех возможных вариантах смещений массивов относительно друг друга.

При этом в качестве экспериментальных входных данных для разработанного способа идентификации использовались массивы бинарных данных размером 1500 байт, сгенерированные случайным образом [15].

Для процедуры побайтного сравнения использовались сгенерированные случайным образом массивы бинарных данных размерами 1000 и 2000 байт, поскольку такое сравнение подразумевает итерации, сопровождающиеся побайтным сдвигом массивов относительно друг друга.

Выбор размеров входных данных для процедуры побайтного сравнения обусловлен большим количеством нечётких дубликатов массивов бинарных данных в исходных текстах различных версий ядра ОС Linux, полученных в результате разделения массива на два меньшего размера или обратного слияния. Таким образом, при одинаковом среднем размере массива в 1500 байт процедура побайтного сравнения оперирует массивами с двукратной разницей в размерах, имитируя типовую задачу идентификации массивов бинарных данных в исходных текстах.

Тестовым стендом для проведения эксперимента являлся сервер проведения тематических исследований под управлением аппаратного гипервизора ESXi с виртуальной машиной ОС Windows 7 64-bit.

Время выполнения отдельно замерялось для трёх операций:

- побайтного сравнения двух массивов со смещениями и поиском максимального количества совпадающих байт;
- вычисления идентификатора в соответствии с разработанным алгоритмом;
- сравнения идентификаторов в соответствии с разработанным алгоритмом.

Результаты эксперимента представлены в таблице 2.

Таблица 2

Результаты замеров времени (составлено авторами)

Операция	Количество повторов	Время выполнения (сек.)	Среднее время операции (сек.)
Побайтное сравнение	10 000	40	0,004
	20 000	80	
	30 000	121	
Вычисление идентификатора	10	11	1,13
	20	23	
	30	34	
Сравнение идентификаторов	1 000 000	4	0,000004
	2 000 000	7	
	3 000 000	12	

По результатам эксперимента построен график зависимости времени идентификации $T_и$ (ч.) от количества массивов бинарных данных N_M и количества эталонных образцов в базе данных (репозитории) N_B для процедуры побайтного сравнения и разработанного способа идентификации (рис. 4).

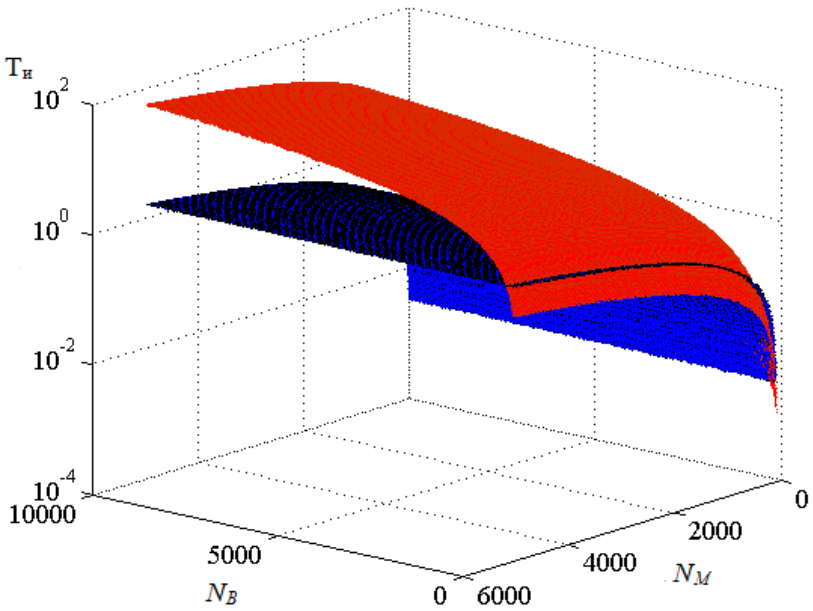

Рисунок 4. *Зависимость времени идентификации от количества массивов бинарных данных и эталонных образцов в репозитории для побайтного сравнения и разработанного метода (составлено авторами)*

Из графика видно, что время идентификации массивов бинарных данных для разработанного способа практически не зависит от количества эталонных образцов в репозитории, а на процедуру побайтного сравнения оказывают влияние оба параметра.

При количестве эталонных образцов в репозитории $N_B > 283$ время идентификации для разработанного метода не превысит соответствующего времени побайтного сравнения. Применительно к процессу автоматизированного контроля информационных объектов предполагаемый рабочий объём репозитория составляет 100 000 эталонных образцов, и будет только увеличиваться с проведением очередных тематических исследований. Для фиксированного размера репозитория N_B = 100 000 эталонных образцов построен график зависимости времени идентификации $T_и$ (ч.) от количества массивов бинарных данных N_M для процедуры побайтного сравнения и разработанного метода идентификации (рис. 5).

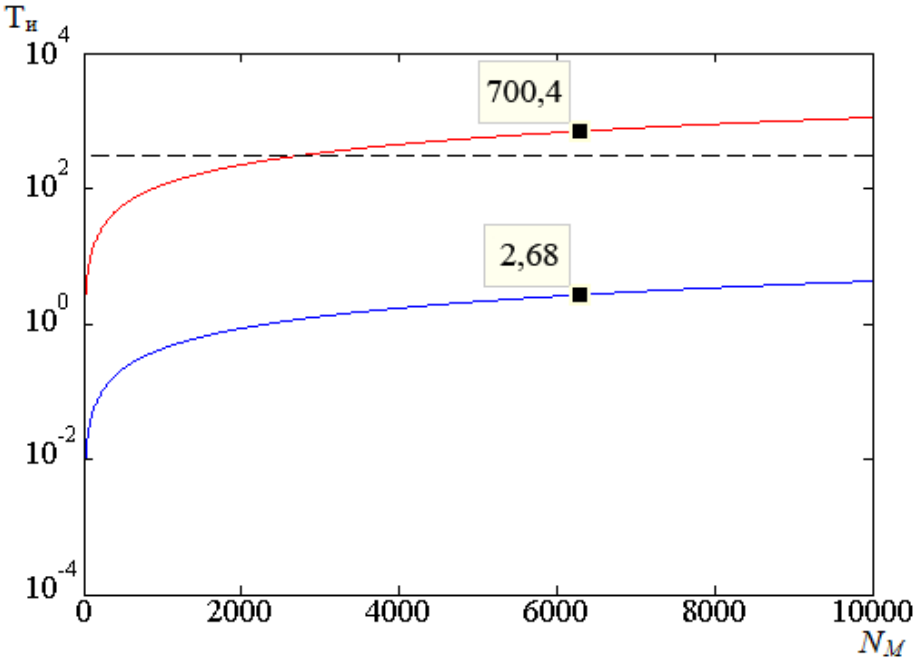

Рисунок 5. *Зависимость времени идентификации от количества массивов бинарных данных для побайтного сравнения и разработанного метода (составлено авторами)*

Заключение

При проведении тематических исследований исходных текстов ядра ОС, содержащих примерно 6000 массивов бинарных данных, и репозитории, содержащем 100 000 эталонных образцов, время идентификации при использовании процедуры побайтного сравнения составит более 700 часов при допуске в $T_{доп}$ = 300 часов на данный вид работ.

В аналогичных условиях время идентификации при использовании разработанного метода составляет менее 3 часов. Таким образом, для усреднённых типовых условий функционирования достигнуто существенное уменьшение вычислительной сложности метода идентификации массивов бинарных данных.

ЛИТЕРАТУРА

1. Рябоконь В.В. Алгоритмы поиска и идентификации массивов бинарных данных в исходных текстах программного обеспечения [Текст] / Диссертация на соискание ученой степени кандидата технических наук. - Орел, 2016. - 130 с.
2. Варшавский П.Р., Алехин Р.В. Метод поиска решений в интеллектуальных системах поддержки принятия решений на основе прецедентов // Information Models and Analyses. Vol.2. 2013. Number 4. С. 385-392.
3. Berman A.F., Nikolaychuk O.A., Yurin A.Yu. Automated Planning with the Aid of Case-based Reasoning and Group Decision-making Methods // Computer Communication & Collaboration. vol. 2. 2014. Issue 1. pp. 7-15.
4. Фролов А.С. Разработка алгоритма нечеткого поиска на основе хэширования // Молодой ученый. 2016. №13. С. 357-360.
5. Шрейдер Ю.А. Что такое расстояние? // Популярные лекции по математике. М.: Физматгиз. 1963. Выпуск 38. 76 С.
6. Султанов Р.О., Еланцев М.О., Кощеев Н.М., Животов В.В. Поиск и классификация структурных элементов методом взаимной корреляции на примере распознавания автомобильного номера // Приволжский научный вестник. 2016. №5 (57). С. 71-74.
7. Тверетин А.А. Обработка информации на основе спектрального импульсного преобразования для сравнения и классификации дискретных данных, циркулирующих в промышленном предприятии // Автореферат диссертации на соискание ученой степени кандидата технических наук. Самара. 2010. С. 23.
8. Tridgell A. Efficient Algorithms for Sorting and Synchronization. URL: https://www.samba.org/~tridge/phd_thesis.pdf (дата обращения: 26.06.2016).
9. Broder A. On the resemblance and containment of documents. URL: http://gatekeeper.dec.com/ftp/pub/dec/SRC/publications/broder/positano-final-wpnums.pdf (дата обращения: 26.06.2016).
10. Broder A., Charikar M., Frieze A., Mitzenmacher M. Min-Wise Independent Permutations. URL: http://www.cs.princeton.edu/courses/archive/spring04/cos598B/bib /BroderCFM-minwise.pdf (дата обращения: 26.06.2016).
11. Рябоконь В.В. Подходы к идентификации массивов бинарных данных // Телекоммуникации. Выпуск №2. 2016. С. 26-32.
12. Розенберг Г.С. Поль Жаккар и сходство экологических объектов // Самарская Лука: Проблемы региональной и глобальной экологии. №1. 2012. С. 190-202.
13. Лебеденко Е.В., Рябоконь В.В. Проверка гипотезы о равномерном распределении значений хэш-функции // Вопросы кибербезопасности. 2016. №2(15). С. 36-40.
14. Лебеденко Е.В., Рябоконь В.В., Игнатов Ю.Н. Выбор управляемых параметров алгоритма идентификации массивов бинарных данных // Интернет-журнал «Науковедение». Том 8. Выпуск №3. 2016. URL: http://naukovedenie.ru/PDF/108TVN316.pdf (дата обращения: 02.02.2017).
15. Шубин Д.Н., Шинаков Ю.С. Объектно-ориентированный подход к разработке математических моделей семейств псевдослучайных последовательностей // T-Comm: Телекоммуникации и транспорт. 2015. Том 9. №7. С. 21-24.

Lapko Alexander Nikolaevich
The academy of federal security guard service of the Russian Federation, Russia, Orel
E-mail: lan46@mail.ru

Ryabokon' Vladimir Vladimirovich
The academy of federal security guard service of the Russian Federation, Russia, Orel
E-mail: mimicria@mail.ru

Kutsakin Maxim Alekseevich
The academy of federal security guard service of the Russian Federation, Russia, Orel
E-mail: max_kooks@mail.ru

Grigoryan Daniil Rubenovich
The academy of federal security guard service of the Russian Federation, Russia, Orel
E-mail: daniil96grigoryan@yandex.ru

Shindryaev Alexey Vladimirovich
The academy of federal security guard service of the Russian Federation, Russia, Orel
E-mail: shialsha@mail.ru

Experimental research of binary data arrays identification method

Abstract. The article deals with the problem of identifying fuzzy duplicates among binary data arrays in the source code of the software. An approach is proposed to reduce the computational complexity of the identification method in conditions of large volumes of initial and reference data on the basis of min-wise independent permutations scheme. The results of experimental verification of the proposed identification method are presented, showing the possibilities of its use in the technological process of information objects automated control, as well as its effectiveness in relation to existing methods.

Keywords: identification; binary data arrays; certification tests; big data; experimental research

REFERENCES

1. Ryabokon V.V. An algorithms of binary data arrays research and identification in software source code [Text] / Dissertation for the degree of candidate of technical sciences. - Orel, 2016. - 130 p.
2. Varshavsky P.R., Alekhin R.V. A method for finding solutions in intelligent decision support systems based on use cases // Information Models and Analyzes. Vol.2. 2013. Number 4. P. 385-392.
3. Berman A.F., Nikolaychuk O.A., Yurin A.Yu. Automated Planning with the Aid of Case-based Reasoning and Group Decision-making Methods // Computer Communication & Collaboration. Vol. 2. 2014. Issue 1. pp. 7-15.
4. Frolov A.S. Development of the algorithm of fuzzy search based on hashing // Young scientist. 2016. № 13. Pp. 357-360.
5. Shreider Yu.A. What is the distance? // Popular lectures on mathematics. Moscow: Fizmatgiz. 1963. Issue 38. 76 C.

6. Sultanov R.O., Elantsev M.O., Koshcheev N.M., Zhivotov V.V. Search and classification of structural elements by the method of mutual correlation on the example of recognition of a car number // Privolzhsky scientific herald. 2016. №5 (57). Pp. 71-74.

7. Tveretin A.A. Information processing on the basis of spectral impulse transformation for comparison and classification of discrete data circulating in an industrial enterprise // The dissertation author's abstract on competition of a scientific degree of a Cand. Tech. Sci. Samara. 2010. P. 23.

8. Tridgell A. Efficient Algorithms for Sorting and Synchronization. URL: https://www.samba.org/~tridge/phd_thesis.pdf (date of circulation: June 26, 2016).

9. Broder A. On the resemblance and containment of documents. URL: http://gatekeeper.dec.com/ftp/pub/dec/SRC/publications/broder/positano-final-wpnums.pdf (reference date: June 26, 2016).

10. Broder A., Charikar M., Frieze A., Mitzenmacher M. Min-Wise Independent Permutations. URL: http://www.cs.princeton.edu/courses/archive/spring04/cos598B/bib/BroderCFM-minwise.pdf (date of circulation: June 26, 2016).

11. Ryabokon V.V. Approaches to the identification of arrays of binary data // Telecommunications. Issue number 2. 2016. P. 26-32.

12. Rosenberg G.S. Paul Jacquard and similarity of ecological objects // Samara Luke: Problems of regional and global ecology. №1. 2012. P. 190-202.

13. Lebedenko E.V., Ryabokon V.V. Testing the hypothesis of a uniform distribution of hash values // Cybersecurity issues. 2016. № 2 (15). Pp. 36-40.

14. Lebedenko E.V., Ryabokon V.V., Ignatov Yu.N. Choice of controllable parameters of the algorithm for identification of arrays of binary data // Internet journal "Naukovedenie". Volume 8. Issue №3. 2016. URL: http://naukovedenie.ru/PDF/108TVN316.pdf (reference date: 02/02/2017).

15. Shubin D.N., Shinakov Yu.S. Object-oriented approach to the development of mathematical models of families of pseudorandom sequences // T-Comm: Telecommunications and Transport. 2015. Volume 9. №7. Pp. 21-24.