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

Том 8, №3 (2016) http://naukovedenie.ru/index.php?p=vol8-3

URL статьи: http://naukovedenie.ru/PDF/108TVN316.pdf

Статья опубликована 29.06.2016.

Ссылка для цитирования этой статьи:
Лебеденко Е.В., Рябоконь В.В., Игнатов Ю.Н. Выбор управляемых параметров алгоритма идентификации массивов бинарных данных // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 8, №3 (2016) http://naukovedenie.ru/PDF/108TVN316.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.

УДК 004.67

Лебеденко Евгений Викторович
ФГКВОУ ВО «Академия Федеральной службы охраны Российской Федерации», Россия, Орёл[1]
Кандидат технических наук, сотрудник
E-mail: lebedenko_eugene@mail.ru

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

Игнатов Юрий Николаевич
ФГКВОУ ВО «Академия Федеральной службы охраны Российской Федерации», Россия, Орёл
Сотрудник
Кандидат технических наук, доцент
E-mail: june959@mail.ru

Выбор управляемых параметров алгоритма идентификации массивов бинарных данных

Аннотация. В статье рассматриваются вопросы выбора параметров разработанного авторами алгоритма идентификации массивов бинарных данных с применением метода независимых перестановок. По результатам анализа алгоритмов некриптографических хэш-функций в качестве базовой для метода независимых перестановок выбрана хэш-функция, основанная на линейном конгруэнтном методе. Представлены упрощённые схемы разработанных алгоритмов выработки и сравнения идентификаторов с использованием набора базовых хэш-функций. Выбраны и обоснованы подходящие значения для размера блоков, на которые будет разбиваться массив бинарных данных, и количества базовых хэш-функций, используемых для независимых перестановок. Авторами проведены экспериментальные проверки и представлены результаты измерений оценки сходства для различных параметров алгоритма идентификации.

Ключевые слова: информационные объекты; недекларированные возможности; массивы бинарных данных; идентификация; независимые перестановки; коэффициент Жаккара; алгоритм; оценка сходства; хэш-функции; размер блока

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

Введение

Идентификация представляет собой совокупность двух взаимосвязанных элементов: присвоения идентификатора и сравнения предъявленного идентификатора с перечнем присвоенных идентификаторов. К настоящему времени опубликовано большое число теоретических и экспериментальных исследований, посвященных выявлению нечетких дубликатов в процессе идентификации [1-3]. При идентификации массивов бинарных данных процедуры присвоения и сравнения идентификаторов должны обладать низкой вычислительной сложностью, что обусловлено большими объёмами исходных данных для сравнения и ограничениями на временной ресурс. В работе [4] показано, что основные существующие и перспективные подходы к идентификации обладают неудовлетворительной, несмотря на полиномиальность, вычислительной сложностью применительно к контролю информационных объектов по требованиям РД НДВ[2]. Предложенный способ идентификации массивов бинарных данных базируется на методе независимых перестановок (англ. min-wise independent permutations), при этом для получения оценки сходства массивов бинарных данных используются минимальные хэш-значения набора независимых хэш-функций [5, 6]. Полученные в ходе моделирования [4] экспериментальные данные являются основой для синтеза алгоритма идентификации массивов бинарных данных и формирования области его применения и ограничений.

Алгоритм присвоения и сравнения идентификаторов

При использовании метода независимых перестановок идентификатором массива бинарных данных является вектор минимальных сигнатур хэш-функций (выражение 1), каждый элемент которого вычисляется с помощью выражения 2.

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

где

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

По результатам анализа алгоритмов некриптографических хэш-функций [7] в качестве базовой для алгоритма идентификации массивов бинарных данных выбрана функция, основанная на линейном конгруэнтном методе, представленная выражением 3.

$$h_i(s_j) = (seed[i] \cdot h_i(s_{j-1}) + s_j) \bmod m, \quad (3)$$

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

Данный выбор обусловлен низкой вычислительной сложностью (1 умножение, 1 сложение, 1 взятие по модулю), а также возможностью получения набора хэш-функций для независимых перестановок с помощью различных значений коэффициента функции $seed[i]$. Очевидно, что выбранная хэш-функция не является совершенной [8], однако коллизии хэш-функции не принимаются в расчёт ввиду отсутствия злоумышленника и возможности дискредитации информации.

Упрощённая схема алгоритма выработки (присвоения) идентификатора показана на рисунке 1.

[2] Руководящий документ. Защита от несанкционированного доступа к информации. Часть 1. Программное обеспечение средств защиты информации. Классификация по уровню контроля отсутствия недекларированных возможностей. М.: Гостехкомиссия России, 1999.

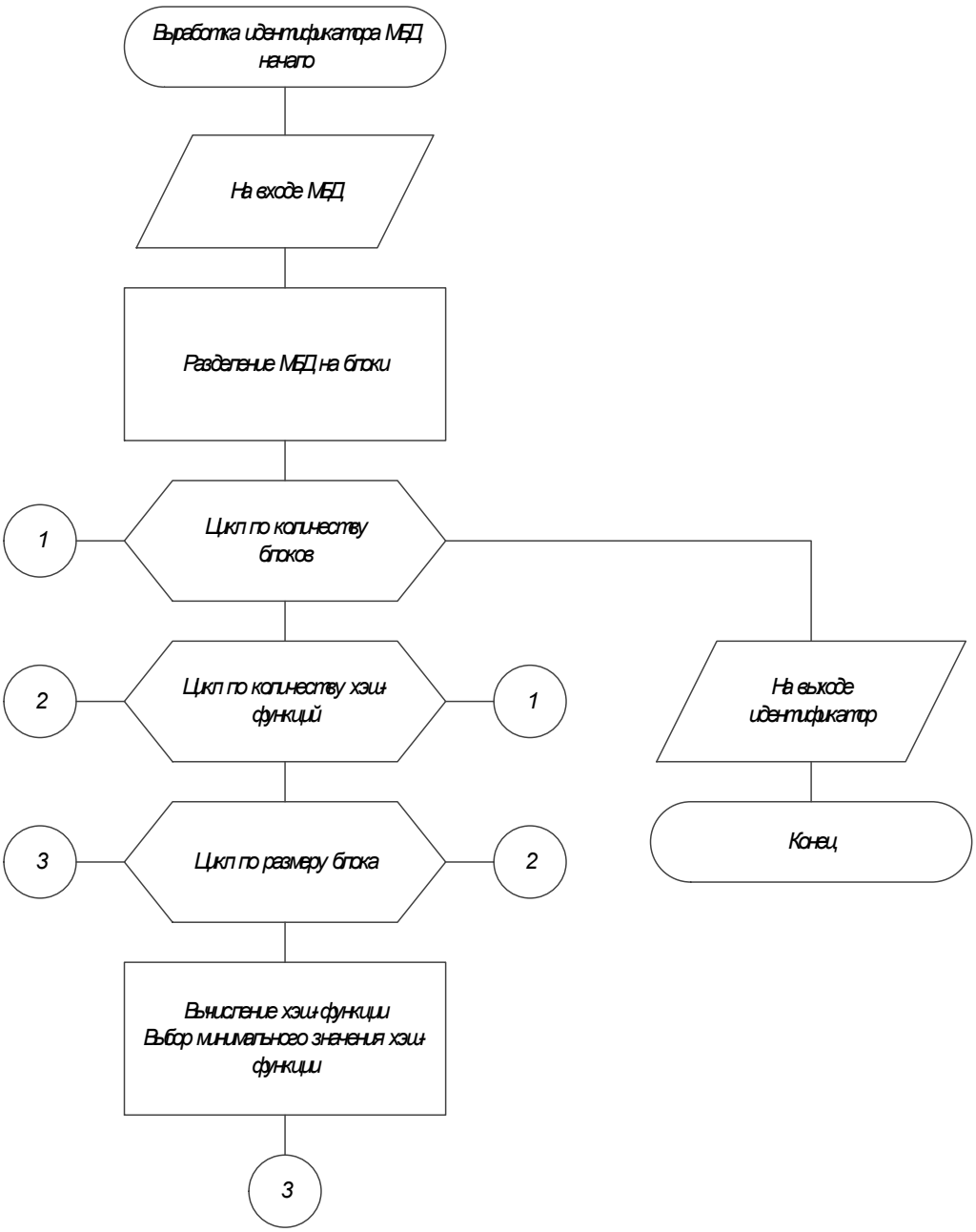

Рисунок 1. *Упрощённая схема алгоритма выработки идентификатора*

При этом для массивов бинарных данных сравнение отдельных байтов массивов (как минимальных элементов) бессмысленно, и возникает необходимость разделения массива на блоки. К каждому блоку применяется набор из *n* независимых хэш-функций. Вектор размера *n* формируется путём выбора минимальной сигнатуры для каждой хэш-функции из набора.

Сравнение идентификаторов осуществляется с использованием функции сравнения (выражение 4), результатом сравнения является оценка сходства массивов (выражение 5).

$$F_i^{AB} = \begin{cases} 1, h_i^{\min}(A) \oplus h_i^{\min}(B) = 0 \\ 0, h_i^{\min}(A) \oplus h_i^{\min}(B) \neq 0 \end{cases} \quad (4)$$

$$\hat{R} = \sum_i F_i^{AB} \quad (5)$$

Упрощённая схема алгоритма сравнения идентификаторов показана на рисунке 2.

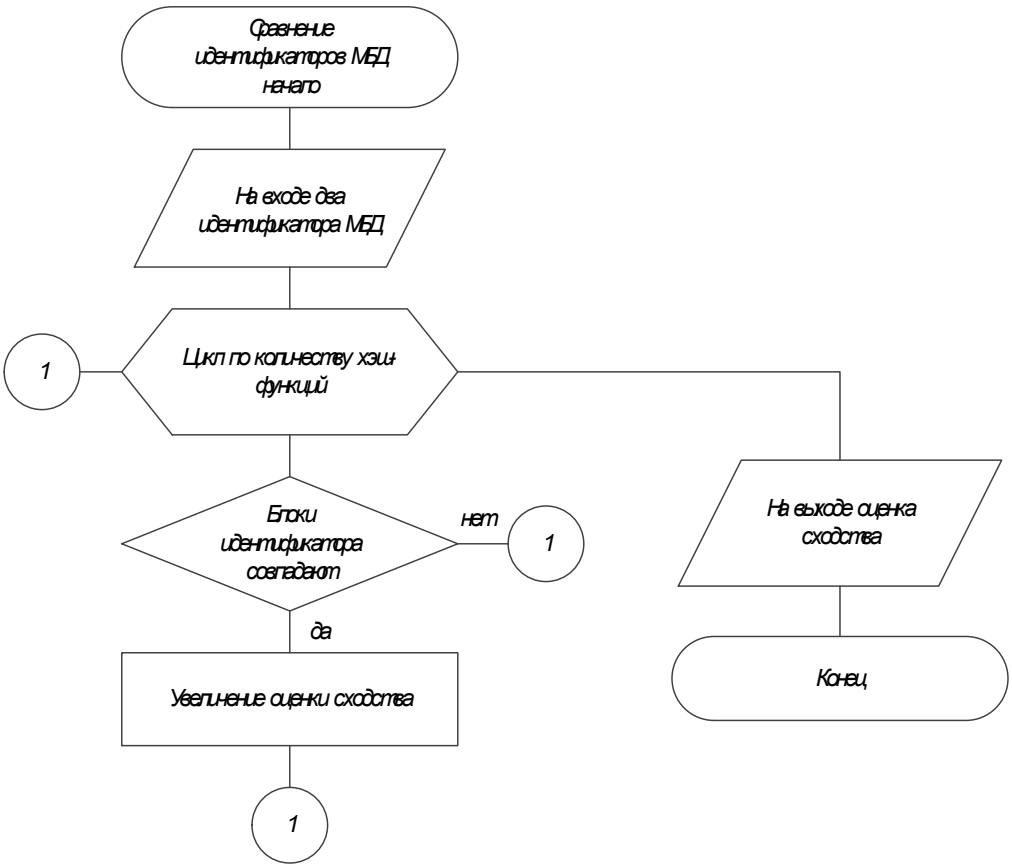

Рисунок 2. *Упрощённая схема алгоритма сравнения идентификаторов*

Выбор значений параметров

Для реализации алгоритма идентификации необходимо выбрать подходящие значения для размера блоков W, на которые будет разбиваться массив бинарных данных, и количество хэш-функций n, используемых для независимых перестановок.

Зафиксируем размеры массивов определённой величиной $A_b = B_b = 2048$ (байт), количество хэш-функций $n = 100$ и количество совпадающих байт $S_b = 0.5 \cdot A_b = 1024$ (байта). Тогда при изменяющемся значении размера блока $W_b = 1 \ldots 100$ результаты измерений оценки сходства на каждом шаге представлены на рисунке 3.

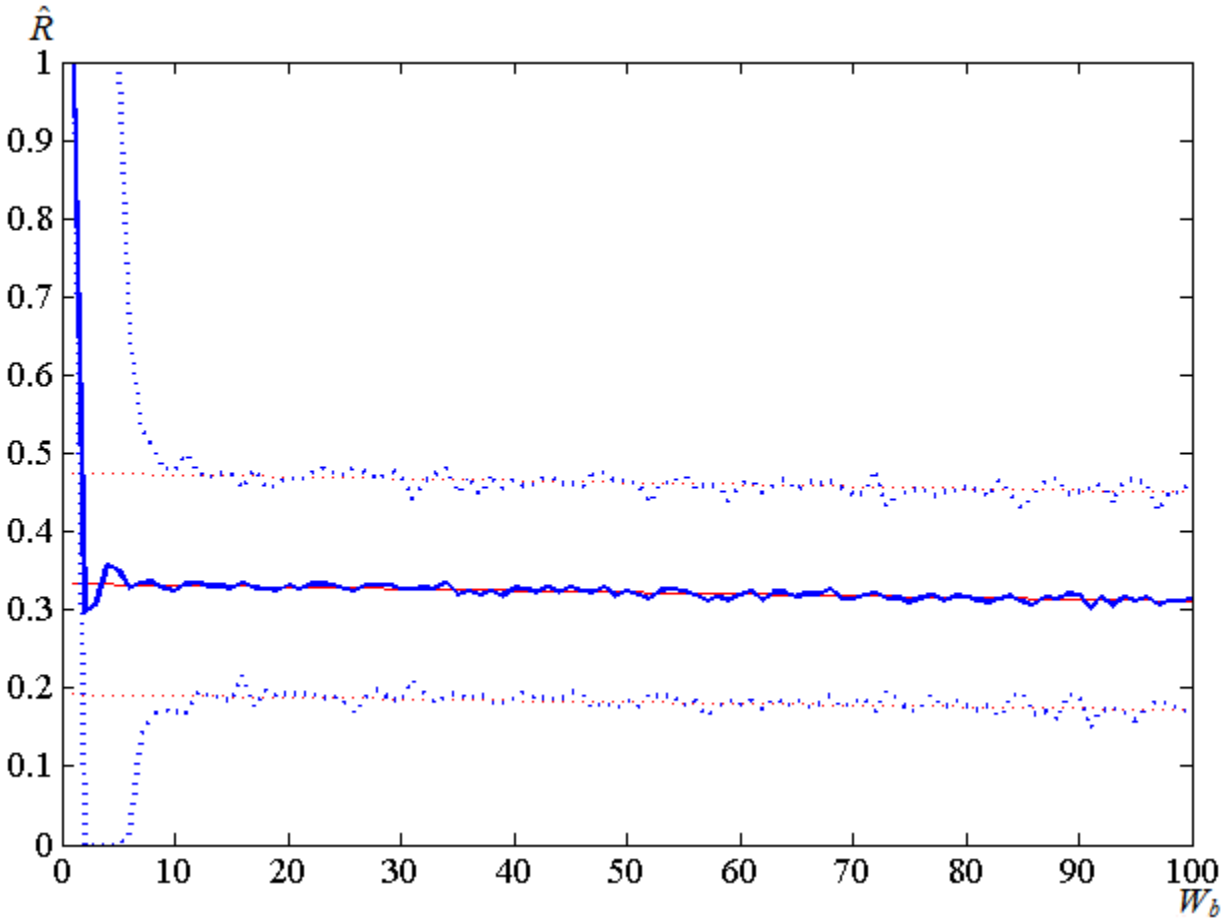

Рисунок 3. *Результаты измерений оценки сходства при изменяющемся размере блока*

При этом аналитически рассчитанные оценка сходства и доверительный интервал совпадают со своими статистическими значениями не на всём диапазоне значений. 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 оценки сходства массивов близко к единице. Это обусловлено высокой вероятностью совпадения коротких блоков байт массивов бинарных данных и низким перекрытием блоков. Таким образом, для идентификации массивов бинарных данных предлагается использовать размер блока W_b = **16** (байт). Как видно из графика на рис. 3 дальнейшее увеличение размера блока не влияет на точность получаемой оценки сходства, но приведёт к увеличению вычислительной сложности алгоритма.

На выбор количества хэш-функций влияет необходимая точность получаемой оценки сходства. Для метода независимых перестановок данная оценка имеет биномиальное распределение с математическим ожиданием частоты появления события совпадения минимальных значений хэш-функций, равным коэффициенту Жаккара (выражение 6).

$$p = J(S_W) = \frac{S_W}{A_W + B_W - S_W}, \tag{6}$$

где: A_W, B_W – количество блоков массивов A и B, S_W – количество совпадающих блоков.

Дисперсия этой величины рассчитывается с использованием выражения 7.

$$\sigma^2 = \frac{p \cdot (1-p)}{n} \tag{7}$$

Необходимо отметить, что с ростом количества испытаний, т.е. с увеличением количества хэш-функций n, дисперсия стремится к нулю, а частота появления события в испытаниях – к истинной вероятности наступления события. При расчёте доверительного интервала с использованием границы Высочанского-Петунина +/-3·σ точность алгоритма идентификации будет возрастать пропорционально $\sqrt{n}$.

Зафиксируем размеры массивов определённой величиной $A_b = B_b = 2048$ (байт), размер блока $W_b = 16$ и количество совпадающих байт $S_b = 0.5 \cdot A_b = 1024$ (байта). Тогда при изменяющемся значении количества хэш-функций $n = 1…1000$ результаты измерений оценки сходства на каждом шаге представлены на рисунке 4.

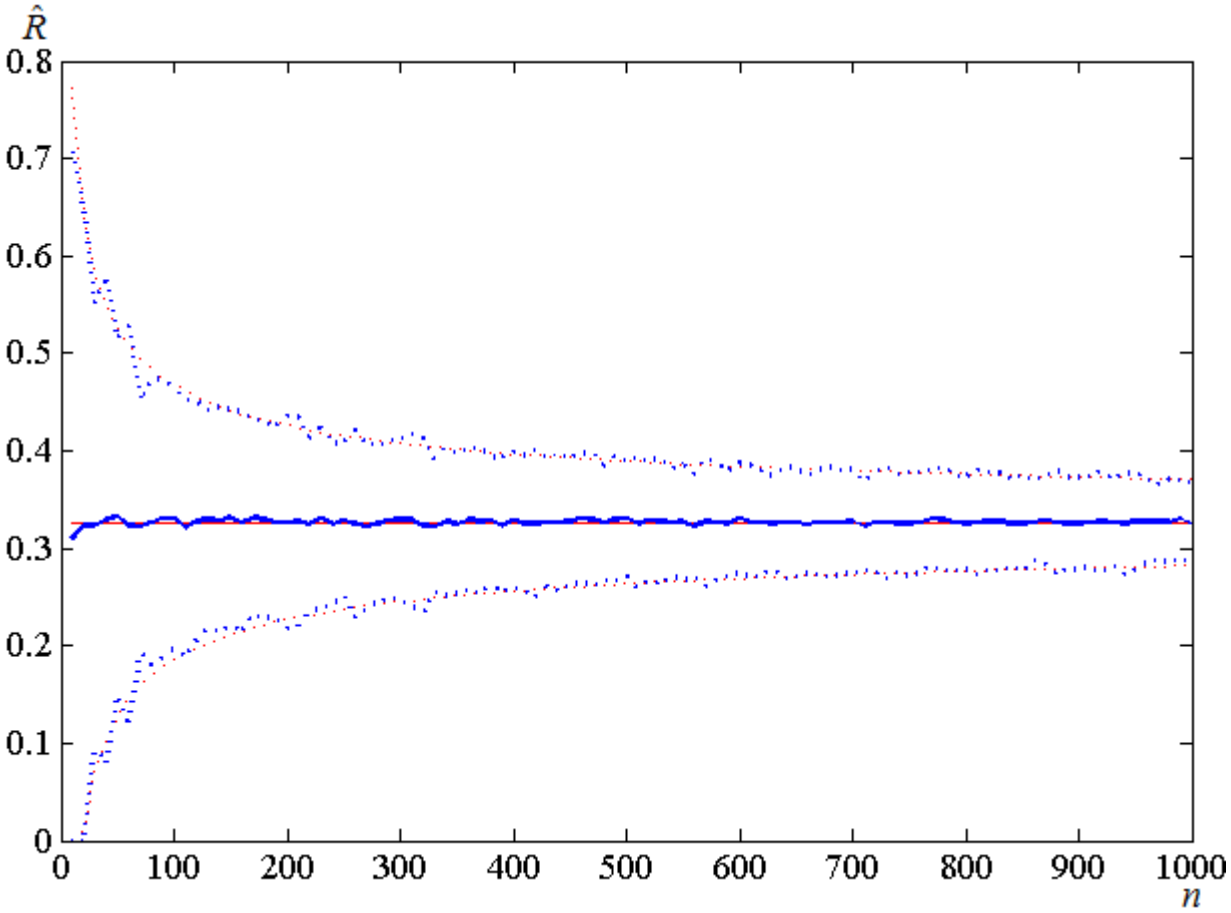

Рисунок 4. *Результаты измерений оценки сходства при изменяющемся количестве хэш-функций*

При сравнении Web-документов использовалось $n = 84$, в дальнейшем полученные значения группировались в более крупные структуры "мегашинглы" [9]. Для удобства расчётов при идентификации массивов бинарных данных с приемлемой точностью достаточно задать $n = 100$. Дальнейшее увеличение количества хэш-функций будет оказывать всё меньшее влияние на точность идентификации при существенном увеличении вычислительной сложности способа.

Заключение

Полученные результаты показывают, что набор хэш-функций, формируемый на базе линейного конгруэнтного метода, может применяться для идентификации массивов бинарных данных с использованием метода независимых перестановок. При этом для разработанного алгоритма идентификации аналитически рассчитанные оценка сходства и доверительный интервал совпадают со своими статистическими значениями в большом диапазоне значений управляемых параметров. Выбранные управляемые параметры алгоритма могут быть использованы для практической реализации разработанного алгоритма идентификации массивов бинарных данных. Направлением дальнейших исследований в этой области является оценка свойств разработанных алгоритмов и выбранного набора хэш-функций [10].

ЛИТЕРАТУРА

1. Chowdhury A., Frieder O., Grossman D., Mccabe C. Collection statistics for fast duplicate document detection // ACM Trans. Inform. Syst., 2002. Vol. 20, No. 2. P. 171-191.
2. Лебеденко Е.В., Рябоконь В.В. Автоматизация процесса проверки исходных текстов специального программного обеспечения на наличие бинарных вставок. - Информационные технологии моделирования и управления №4 (88). - 2014 г. - С. 328-333.
3. Пименов В.Ю. Метод поиска нечётких дубликатов изображений на основе выявления точечных особенностей. Труды РОМИП 2007-2008. - СПб.: НУ ЦСИ, 2008. - С. 145-158.
4. Рябоконь В.В. Моделирование идентификации массивов бинарных данных. – Системы управления и информационные технологии №3.1 (61). - 2015. - С. 172-178.
5. A. Broder, M. Charikar, A. Frieze, M. Mitzenmacher. Min-Wise Independent Permutations. [Электронный ресурс]. - Режим доступа: http://www.cs.princeton.edu/courses/archive/spring04/cos598B/bib/BroderCFM-minwise.pdf, свободный (дата обращения: 10.03.2016).
6. A. Broder. On the resemblance and containment of documents. [Электронный ресурс]. - Режим доступа: http://gatekeeper.dec.com/ftp/pub/dec/SRC/publications/broder/positano-final-wpnums.pdf, свободный (дата обращения: 10.03.2016).
7. Partow A. General Purpose Hash Function Algorithms. [Электронный ресурс]. - Режим доступа: http://www.partow.net/programming/hashfunctions/, свободный (дата обращения: 11.03.2016).
8. Pescio C. Minimal perfect hashing // Dr. Dobb's Journal. - № 249, 1996 [Электронный ресурс]. - Режим доступа: http://www.eptacom.net/pubblicazioni/pub_eng/mphash.html, свободный (дата обращения: 20.05.2016).
9. Зеленков Ю.Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для Web-документов. Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». - RCDL-2007. - Переславль-Залесский. - 2007.
10. Поляков Д.В., Попов А.И. Генератор монотонных хеш-функций для ассоциативного массива: Труды Нижегородского государственного технического университета им. Р.Е. Алексеева. - №2 (109), 2015. - С. 70.

Lebedenko Eugene Viktorovich
The Academy of the Federal guard service of the Russian Federation, Russia, Orel
E-mail: lebedenko_eugene@mail.ru

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

Ignatov Yury Nikolaevich
The Academy of the Federal guard service of the Russian Federation, Russia, Orel
E-mail: june959@mail.ru

Controllable parameters selection for binary data arrays identification algorithm

Abstract. The article discusses the questions of controllable parameters selection for authors developed algorithm of binary data arrays identification using the method of independent permutations. According to the non-cryptographic hash functions algorithms analysis the hash function based on linear congruential generator is selected as a base for the method of independent permutations. The simplified schemes of developed algorithms for identifiers generation and comparison using a set of basic hash functions are presented. An appropriate values for the block size which will split the binary data array, and the number of basic hash functions used for independent permutations are selected and proved. The authors carried out experimental tests and the results of similarity evaluation metering for various identification algorithm parameters are shown.

Keywords: information objects; undeclared capabilities; binary data arrays; identification; independent permutations; Jaccard index; algorithm; similarity value; hash functions; block size

REFERENCES

1. Chowdhury A., Frieder O., Grossman D., Mccabe C. Collection statistics for fast duplicate document detection // ACM Trans. Inform. Syst., 2002. Vol. 20, No. 2. P. 171-191.
2. Lebedenko E.V., Ryabokon' V.V. Avtomatizatsiya protsessa proverki iskhodnykh tekstov spetsial'nogo programmnogo obespecheniya na nalichie binarnykh vstavok. - Informatsionnye tekhnologii modelirovaniya i upravleniya №4 (88). - 2014 g. - S. 328-333.
3. Pimenov V.Yu. Metod poiska nechetkikh dublikatov izobrazheniy na osnove vyyavleniya tochechnykh osobennostey. Trudy ROMIP 2007-2008. - SPb.: NU TsSI, 2008. - S. 145-158.
4. Ryabokon' V.V. Modelirovanie identifikatsii massivov binarnykh dannykh. – Sistemy upravleniya i informatsionnye tekhnologii №3.1 (61). - 2015. - S. 172-178.
5. A. Broder, M. Charikar, A. Frieze, M. Mitzenmacher. Min-Wise Independent Permutations. [Elektronnyy resurs]. - Rezhim dostupa: http://www.cs.princeton.edu/courses/archive/spring04/cos598B/bib/BroderCFM-minwise.pdf, svobodnyy (data obrashcheniya: 10.03.2016).
6. A. Broder. On the resemblance and containment of documents. [Elektronnyy resurs]. - Rezhim dostupa: http://gatekeeper.dec.com/ftp/pub/dec/SRC/publications/broder/positano-final-wpnums.pdf, svobodnyy (data obrashcheniya: 10.03.2016).
7. Partow A. General Purpose Hash Function Algorithms. [Elektronnyy resurs]. - Rezhim dostupa: http://www.partow.net/programming/hashfunctions/, svobodnyy (data obrashcheniya: 11.03.2016).
8. Pescio C. Minimal perfect hashing // Dr. Dobb's Journal. - № 249, 1996 [Elektronnyy resurs]. - Rezhim dostupa: http://www.eptacom.net/pubblicazioni/pub_eng/mphash.html, svobodnyy (data obrashcheniya: 20.05.2016).
9. Zelenkov Yu.G., Segalovich I.V. Sravnitel'nyy analiz metodov opredeleniya nechetkikh dublikatov dlya Web-dokumentov. Trudy 9-oy Vserossiyskoy nauchnoy konferentsii «Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii». - RCDL-2007. - Pereslavl'-Zalesskiy. - 2007.
10. Polyakov D.V., Popov A.I. Generator monotonnykh khesh-funktsiy dlya assotsiativnogo massiva: Trudy Nizhegorodskogo gosudarstvennogo tekhnicheskogo universiteta im. R.E. Alekseeva. - №2 (109), 2015. - S. 70.