ПРОВЕРКА ГИПОТЕЗЫ О РАВНОМЕРНОМ РАСПРЕДЕЛЕНИИ ЗНАЧЕНИЙ ХЭШ-ФУНКЦИИ

Лебеденко Е.В.[1], Рябоконь В.В.[2]

В статье рассматриваются вопросы выбора и оценки свойств хэш–функции, использующейся для идентификации массивов бинарных данных с применением метода независимых перестановок. Предъявлены требования к хэш–функциям, использующимся для метода независимых перестановок. Отмечена необходимость проверки равномерности распределения значений хэш–функции. По результатам анализа алгоритмов некриптографических хэш–функций в качестве базовой для метода независимых перестановок выбрана хэш–функция, основанная на линейном конгруэнтном методе. Рассмотрены варианты проверки выполнения требования по равномерности и предложен подход с использованием метода статистических испытаний и критерия согласия Пирсона. Проведена экспериментальная проверка и подтверждена гипотеза о равномерности распределения значений выбранной хэш–функции.

***Ключевые слова**: независимые перестановки, массивы бинарных данных, хэш–функции, равномерное распределение, критерий Пирсона.*

Введение

Метод независимых перестановок (англ. min-wise independent permutations), применяемый для получения оценки сходства массивов бинарных данных, основан на использовании набора независимых хэш-функций и получения минимальных хэш-значений [1]. При этом получаемая оценка сходства подчиняется биномиальному распределению в случае, когда служащие для перестановок хэш-функции помимо независимости обладают свойством равномерного распределения результатов по всем возможным значениям [2].

Таким образом, используемая для независимых перестановок хэш-функция должна:

- быть достаточно быстрой;
- легко модифицироваться для получения набора независимых функций;
- давать как можно более равномерное распределение хэш-значения.

Следует отметить, что криптографические хэш-функции, хорошо исследованные на равномерность распределения, использовать нецелесообразно вследствие их сравнительно большой вычислительной сложности. Равномерность распределения простых хэш-функций, таких как FAQ6, Rot13, Ly, Rs и т. п., исследована слабо, кроме того, в качестве тестовых данных обычно применялись сформированные под конкретные задачи словари [3].

Хэш-функция

По результатам анализа алгоритмов некриптографических хэш-функций в качестве базовой для метода независимых перестановок выбрана функция, основанная на линейном конгруэнтном методе:

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

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 $seed[i]$.

Определим понятие хэш-функции h_i. Пусть S – это множество возможных блоков на входе хэш-функции. В общем случае размер блока выбирается исходя из общего размера сравниваемых данных и специфики решаемой задачи.

Определим хэш-функцию как отображение:

$$h^m: S \rightarrow \{d \in N,\ 0 \le d \le m - 1,\ m \in N\}. \qquad (2)$$

При использовании 4 байт для хранения значения хэш-функции модуль может принимать значения $m \le 2^{32}$. С учетом требований к линейному конгруэнтному методу целесообразно в качестве модуля принять наиболее близкое к максимальному значению простое число m = 4 294 967 291.

1 Лебеденко Евгений Викторович, кандидат технических наук, Академия ФСО России, г. Орел, lebedenko_eugene@mail.ru

2 Рябоконь Владимир Владимирович, Академия ФСО России, г. Орёл, mimicria@mail.ru

Очевидно, что при заданных размерах входных и выходных данных хэш-функция h_i не будет являться совершенной [4], поскольку она не инъективна на $D \subset S$ [5]. Таким образом, согласно принципу Дирихле, неизбежно наличие коллизий хэш-функции. Однако с учетом некриптографического применения хэш-функции для независимых перестановок вопросы стойкости к коллизиям и необратимости не рассматриваются ввиду отсутствия злоумышленника и возможности дискредитации информации.

В самой простой постановке выполнение требования по равномерности распределения значений хэш-функции можно проверить, перебрав все возможные блоки на входе хэш-функции из множества S, и построить ряд частот хэш-значений. При идентификации массивов бинарных данных для примера будет использоваться размер блока $W = 32$ (байта), таким образом, $S = \{s \in N,\ 0 \le s \le 2^{256}\}$. Очевидно, что такое количество вариантов на входе хэш-функции слишком велико для полного перебора, а составление словарей из имеющейся выборки массивов бинарных данных представляется нецелесообразным.

Создание индексного массива из m элементов для построения ряда частот хэш-значений также нецелесообразно вследствие большого размера модуля и избыточности выделяемой памяти.

Таким образом, для проверки равномерности распределения хэш-значений предлагается использовать метод статистических испытаний и критерий согласия с нулевой гипотезой о равномерности распределения полученной выборки.

Проверка равномерности распределения

Наиболее распространенным критерием в задачах математической статистики является критерий Пирсона (χ^2 – Хи-квадрат) [6]. Пусть результаты испытания таковы, что все они могут быть разделены на N категорий. Проводится l независимых испытаний, т. е. таких, на исход каждого из которых не влияют результаты любого из остальных испытаний.

При проверке рассчитывается

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - n_i')^2}{n_i'}, \qquad (3)$$

где n_i и n_i' – эмпирические и теоретические частоты выпадения соответствующих i-х значений.

Считается, что гипотеза о равномерности распределения принимается с вероятностью не менее 95 %, если рассчитанное значение χ^2 не превысило критического значения $\chi^2_{\text{крит}}$ для доверительной вероятности 0,95 и $k = N - 2 - 1$ (k – число степеней свободы, 2 – число параметров, по которому оценивается равномерное распределение, 1 – константа в критерии Пирсона) [7].

Условием использования критерия Пирсона является достаточно большое количество испытаний l. Кроме того, статистические свойства критериев типа χ^2 зависят как от того, каким образом область значений случайной величины разбивается на интервалы, так и от выбора числа интервалов группирования N [8]. Рекомендуемое в различных источниках количество интервалов группирования, используемое при проверке статистических гипотез с помощью критерия Пирсона, колеблется в очень широких пределах.

Поскольку проверяется гипотеза о равномерном распределении значений хэш-функции, множество возможных хэш-значений D разбивается на равные интервалы. При этом количество интервалов можно определить исходя из минимизации ошибки разбиения $\varepsilon \ge 1$, неизбежно возникающей в случае простого значения модуля m. Возможные значения количества интервалов для минимальной ошибки $\varepsilon = 1$ представлены в таблице 1.

Таблица 1

Разбиение области хэш-значений на интервалы

Количество интервалов	Размер интервала	Количество испытаний
2	2 147 483 645	4
5	858 993 458	25
10	429 496 729	100
19	226 050 910	361
38	113 025 455	1444
95	45 210 182	9025
190	22 605 091	36 100

Таблица 2

Результаты расчетов по критерию χ^2

h_1	80,4	102,9	98,1	96,5	107,9
h_2	75,7	98,4	71,5	93,5	78,5
h_3	105,5	89,6	103,5	81,4	74,6
h_4	96,7	91,9	101,9	96,2	98,2
…					
h_{100}	97,7	102,2	85,5	101,3	70,3

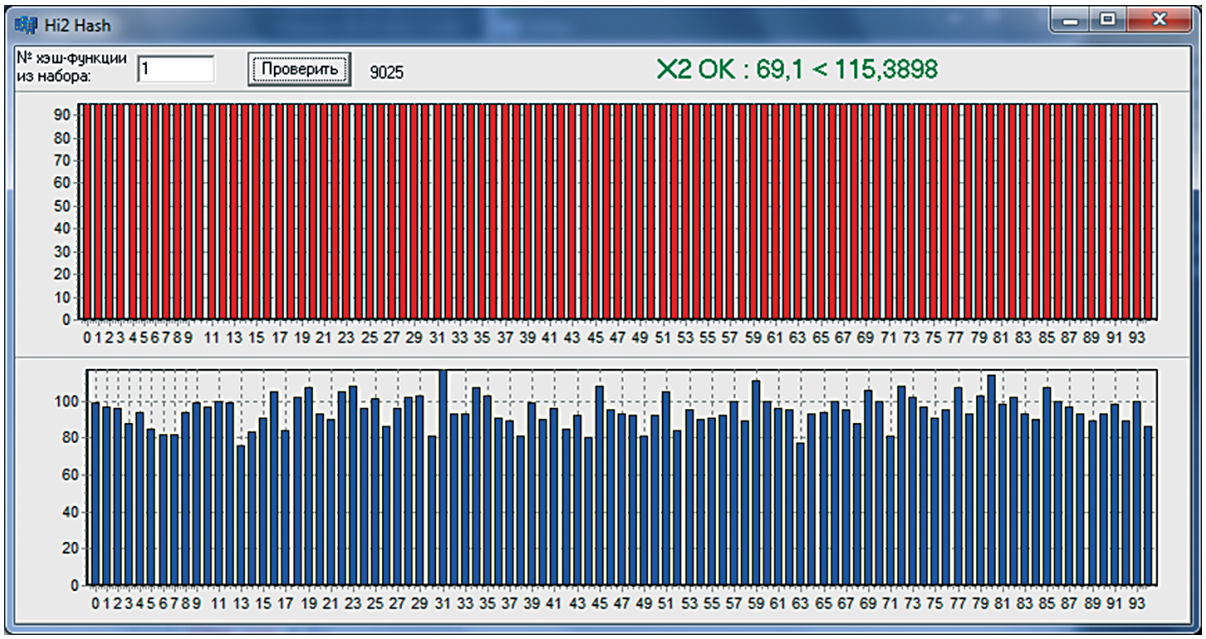

Рис. 1. *Пример программы расчета критерия χ^2*

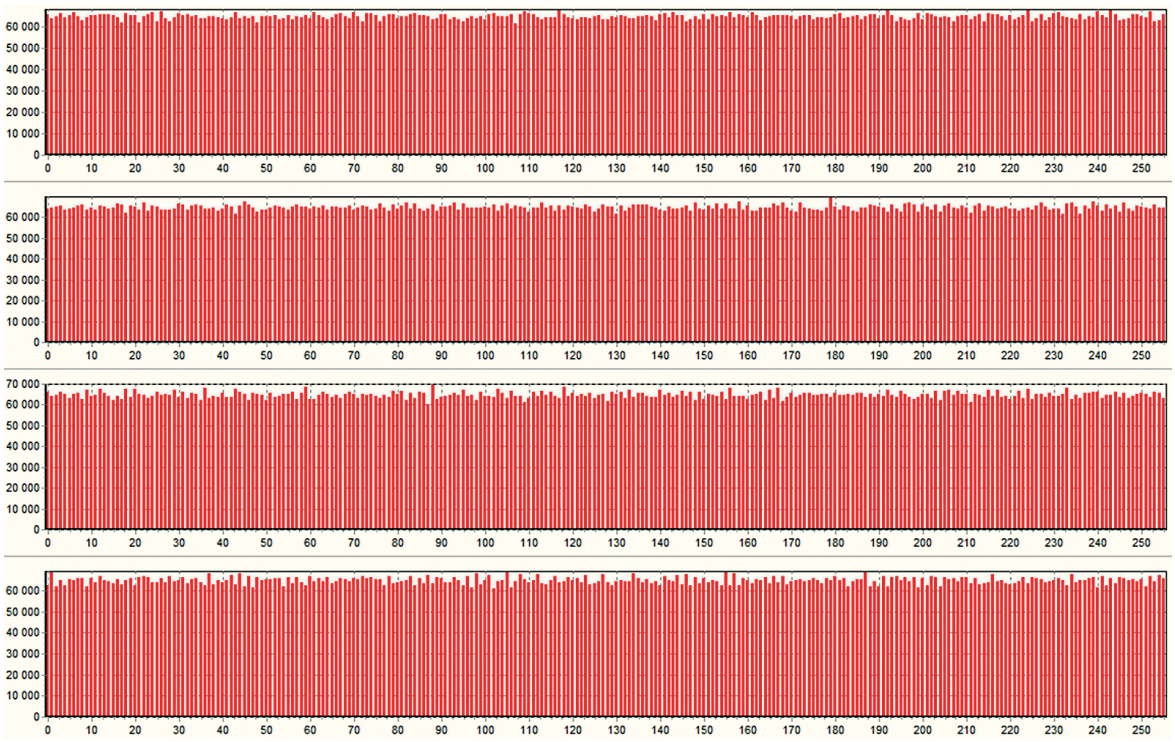

Рис. 2. *Распределение каждого байта хэш-значений*

При этом количество испытаний определялось с использованием выражения [9]:

$$l = N^2 \ . \quad (4)$$

Испытания количеством l = 9025 проводились не менее пяти раз для каждой хэш-функции h_i из набора, i = 1…100, результаты представлены в таблице 2, пример программы расчета показан на рисунке 1.

Для количества интервалов N = 95 критическое значение $\chi^2_{крит}$ = 115,3898, таким образом, гипотеза о равномерности распределения принимается с доверительной вероятностью 0,95.

Дополнительно был построен ряд распределения каждого байта хэш-значений в отдельности, результат представлен на рисунке 2.

Таким образом, для метода независимых перестановок можно использовать как полное 4-байтное значение хэш-функции, так и значения каждого байта в отдельности.

Заключение

Результаты показывают, что при достаточно большом количестве статистических испытаний l хеш-функция вида (1) обладает высокой равномерностью хеширования. Таким образом, набор хэш-функций, формируемый с помощью различных значений коэффициента, может быть использован для идентификации массивов бинарных данных с помощью метода независимых перестановок. Направлением дальнейших исследований в этой области являются оценка случайности получаемых хэш-значений и доказательство независимости полученного набора хэш-функций [10].

***Рецензент**: Баранов Владимир Александрович, доктор технических наук, доцент, сотрудник Академии ФСО России, baranov.va@mail.ru*

Литература:

1. 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: 1.05.2016).
2. Рябоконь В.В. Моделирование идентификации массивов бинарных данных // Системы управления и информационные технологии. 2015. Т. 61. № 3-1. С. 172-178.
3. Partow A. General Purpose Hash Function Algorithms. – URL: http://www.partow.net/programming/hashfunctions/ (дата обращения: 1.05.2016).
4. Pescio C. Minimal perfect hashing. Dr. Dobb's Journal. 1996. № 249.
5. Верещагин Н.К., Шень А. Лекции по математической логике и теории алгоритмов. Ч. 1. Начала теории множеств. – 2-е изд., испр. – М.: МЦНМО, 2002. – С. 128.
6. Акимова Г.П., Пашкина Е.В., Соловьев А.В. Методологический подход к оценке качества случайных чисел и последовательностей // Труды Института системного анализа Российской академии наук. 2008. Т. 38. С. 156-167.
7. Виноградов И.М. Математическая энциклопедия. – М.: Сов. Энциклопедия, 1984. – С. 1216.
8. Лемешко Б.Ю., Чимитова Е.В. О выборе числа интервалов в критериях согласия типа χ2 // Заводская лаборатория. Диагностика материалов. 2003. Т. 69. № 1. С. 61-67.
9. Heinhold I., Gaede K.W. Ingeniur statistic. – München; Wien, Springler Verlag. – 1964. – S. 352.
10. Поляков Д.В., Попов А.И. Генератор монотонных хеш-функций для ассоциативного массива // Труды НГТУ им. Р.Е. Алексеева. 2015. № 2 (109). С. 70-81.

HYPOTHESIS VERIFICATION ABOUT THE UNIFORM DISTRIBUTION OF HASH VALUES

Lebedenko E.V.[3], Ryabokon V.V.[4]

The article discusses the selection and evaluation of hash function properties used to identify binary data arrays with the method of independent permutations. Requirements for hash functions used for the method

3 Eugene Lebedenko, Ph.D., The Academy of Federal Security Guard Service of the Russian Federation, Orel, lebedenko_eugene@mail.ru

4 Vladimir Ryabokon, The Academy of Federal Security Guard Service of the Russian Federation, Orel, mimicria@mail.ru

of independent permutations are shown. The necessity of checking the uniformity of the distribution of hash values is considered. According to the analysis of non–cryptographic hash functions algorithms selected hash function based on the linear congruential method as the base for the method of independent permutations. Variants of uniformity checking are considered and proposed an approach using the method of statistical tests and goodness–of–fit Pearson. The experimental verification confirmed the hypothesis about uniform distribution of selected hash values is shown.

***Keywords**: independent permutations, binary data arrays, hash functions, uniform distribution, the Pearson criterion.*

References:

1. 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.
2. Ryabokon' V.V. Modelirovanie identifikatsii massivov binarnykh dannykh, Sistemy upravleniya i informatsionnye tekhnologii. 2015. T. 61. No 3-1, pp. 172-178.
3. Partow A. General Purpose Hash Function Algorithms. – URL: http://www.partow.net/programming/hashfunctions/.
4. Pescio C. Minimal perfect hashing. – Dr. Dobb's Journal. – No 249, 1996.
5. Vereshchagin N.K., Shen' A. Lektsii po matematicheskoy logike i teorii algoritmov. Ch. 1. Nachala teorii mnozhestv. – 2-e izd., ispr. – M. : MTsNMO, 2002. – P. 128.
6. Akimova G.P., Pashkina E.V., Solov'yev A.V. Metodologicheskiy podkhod k otsenke kachestva sluchaynykh chisel i posledovatel'nostey, Trudy Instituta sistemnogo analiza Rossiyskoy akademii nauk. 2008. V. 38, pp. 156-167.
7. Vinogradov I.M. Matematicheskaya entsiklopediya. – M. : Sov. Entsiklopediya, 1984. – P. 1216.
8. Lemeshko B.Yu., Chimitova E.V. O vybore chisla intervalov v kriteriyakh soglasiya tipa X2, Zavodskaya laboratoriya. Diagnostika materialov. 2003. T. 69. No 1, pp. 61-67.
9. Heinhold I., Gaede K.W. Ingeniur statistic. – München; Wien, Springler Verlag. – 1964. – P. 352.
10. Polyakov D.V., Popov A.I. Generator monotonnykh khesh-funktsiy dlya assotsiativnogo massiva, Trudy NGTU im. R.E. Alekseeva. 2015. No 2 (109), pp. 70-81.