ІНФОКОМУНІКАЦІЙНІ ТЕХНОЛОГІЇ ТА ЕЛЕКТРОННА ІНЖЕНЕРІЯ
INFORMATION AND COMMUNICATION TECHNOLOGIES, ELECTRONIC ENGINEERING

№ 2 (1), 2022

https://doi.org/10.23939/ictee2022.01.096

ДОСЛІДЖЕННЯ ЕНТРОПІЇ ДВІЙКОВИХ ПОСЛІДОВНОСТЕЙ, ФРАГМЕНТОВАНИХ НА ПІДПОСЛІДОВНОСТІ СТАЛОЇ ДОВЖИНИ

Р. Політанський, В. Качур

Чернівецький національний університет ім. Ю. Федьковича, вул. Коцюбинського, 2, Чернівці, 58012, Україна

Відповідальний за рукопис: Василь Качур (e-mail: kachur.vasyl.v@chnu.edu.ua).

(Подано 24 липня 2022)

Здійснено дослідження структурної ентропії послідовностей, які фрагментовані на двійкові підпослідовності (символи) заданої довжини, основані на частотах входження цих символів у послідовність. Досліджено послідовності, згенеровані за логістичним відображенням із початковим значенням послідовності із проміжку [0;1] та кроком 0.1. Найбільша довжина символу (підпослідовності) становила 10 біт. Порівняння розрахованих значень ентропії із її максимальним значенням показало, що спостерігається відхилення від рівномірного розподілу для символів, довжина яких 8 і більше біт, але значення ентропії поступово наближається до максимального зі збільшенням довжини підпослідовності. Встановлено також, що такий параметр генерування, як початкове значення, на ентропію не впливає. Дослідження показують також, що алгоритм ефективний завдяки високій швидкодії та не потребує використання значних обчислювальних потужностей.

Ключові слова: *легка криптографія; Інтернет речей; псевдовипадкові послідовності; логістичне відображення.*
УДК: 621.126

1. Вступ

Двійкові послідовності застосовують у найрізноманітніших галузях сучасної науки: від фізичних дворівневих систем (набору елементарних частинок із двома можливими значеннями магнітних моментів) до актуальних задач теорії інформації, які виникають у процесах кодування [1]. У телекомунікаціях дослідження двійкових послідовностей також набувають важливого значення у зв'язку із упровадженням новітніх технологій передавання інформації, що стають стандартом зв'язку в системах Інтернету речей. Хоч уже розроблено низку методів генерування псевдовипадкових послідовностей та досліджено їх властивості [2], пошуки нових методів та способів у цьому напрямі продовжуються [3]. Метою цих досліджень є розв'язання оптимізаційної задачі пошуку способів генерування, що поєднують можливість порівняно нескладної апаратної реалізації із використанням засобів, що набули значного поширення (Arduino, стандарти для RFID [4]), та високу ефективність їх застосування у заданих експлуатаційних умовах.

2. Аналіз та постановка завдання

У [5] вирішено завдання дослідження періодичності псевдовипадкових послідовностей, основаного на застосуванні алгоритму булевого гіперкуба, у якому використано відображення дійсних чисел псевдовипадкової послідовності на багатовимірний масив булевих значень, розмірність якого залежить від розрядності дійсного числа та кількості цифр у кожному розряді. Це точний метод, який дає можливість виявити повторення чисел заданої розрядності в аналізованій послідовності. Разом із тим, метод булевого гіперкуба потребує значної кількості обчислювальних ресурсів, оскільки масив булевих змінних займає доволі великий обсяг пам'яті, який становить 10^n біт оперативної пам'яті, де n – це розрядність чисел у двійковій послідовності.

Дослідження ентропії послідовності також може вирішити питання щодо періодичності послідовності. Пояснимо це на прикладі обчислення ентропії двійкової періодичної послідовності невеликої довжини 36 біт (1):

$$0101001000\ 0101010010\ 0001010100\ 100001\ . \tag{1}$$

Періодом цієї послідовності є підпослідовність 010100100001, довжина якої 12 біт. Розглянемо питання залежності ентропії від довжини послідовності та від її фрагментації на елементарні символи, що утворені на основі елементів двійкового алфавіту {0,1}. Алфавіти, що утворені цими символами, мають такий вигляд: {0, 1}; {00, 01, 10, 11}; {000, 001, 010, 011, 100, 101, 110, 111} і т. д.

Структурна ентропія цієї послідовності визначається за відомою формулою, основаною на імовірностях кожного символу (2):

$$H = -\sum_{s_i \in M} p(s_i) \log_2 p(s_i), \tag{2}$$

де сума розраховується на усій множині M символів алфавіту s_i.

Для виконання розрахунків із генерованими послідовностями в (2) імовірності символів замінюють їхніми частотами у послідовності. Це частково проілюстровано у табл. 1 для символів із довжинами 1, 2 і 3 біти.

Таблиця 1

Частоти входження символів у двійкову послідовність (1)

Довжина послідовності	Частоти символів													
	0	1	00	01	10	11	000	001	010	011	100	101	110	111
1	1	0												
2	0,50	0,50	0	1	0	0								
3	0,67	0,33					0	0	1	0	0	0	0	0
4	0,50	0,50	0	1	0	0								
5	0,60	0,40												
6	0,67	0,33	0,33	0,66	0	0	0	0	0,50	0	0,50	0	0	0
7	0,57	0,43												
8	0,625	0,375	0,25	0,50	0,25	0								
9	0,67	0,33					0	0	0,33	0	0,67	0	0	0
10	0,70	0,30	0,40	0,40	0,20	0								
11	0,73	0,27												
12	0,67	0,33	0,33	0,50	0,17	0	0	0,25	0,25	0	0,50	0	0	0
…														
33	0,67	0,33					0	0,18	0,27	0	0,54	0	0	0
34	0,70	0,30	0,35	0,47	0,18	0								
35	0,73	0,27												
36	0,67	0,33	0,33	0,5	0,17	0	0	0,25	0,25	0	0,50	0	0	0

В ідеальному випадку розподіл символів будь-якої довжини (із урахуванням верхньої межі, яка зумовлена довжиною усієї послідовності) повинен наближатися до рівномірного. Аналіз ентропії один із найкращих у цьому сенсі, тому що за рівномірного розподілу ентропія набуває максимального, заздалегідь відомого значення, яке можна обчислити за формулою (3):

$$H_{\max} = \log_2 2^n = n, \tag{3}$$

де n – довжина символу, біт.

Метою дослідження є визначення залежності ентропії від початкового значення послідовності та ступеня її відхилення від її максимального значення, а також зміни цієї характеристики зі збільшенням довжини послідовності.

3. Опис розробленого алгоритму

Розрахунок ентропії послідовності, фрагментованої на символи різної довжини, реалізовано у програмному середовищі DEV C++. Опишемо роботу цього алгоритму. Одним із вхідних параметрів програми є максимальна довжина підпослідовності, для якої обчислюється ентропія. Програма розраховує ентропію для усіх множин, утворених символами меншої довжини, аж до окремих символів “0” і “1”. Усі значення ентропії розраховуються одночасно, протягом генерування усієї послідовності.

Обчислення значень ентропії усіх множин символів довжиною від одиниці до максимального значення здійснюється для окремих фрагментів послідовності, після завершення процесу їх генерування. Довжина цих фрагментів дорівнює найменшому спільному множнику усіх досліджуваних довжин підпослідовностей:

$$L = НСД\{1,2,3,...,n_{\max}\}, \tag{4}$$

де $n_{\max}$ – найбільша довжина аналізованого символу.

Тоді фрагмент міститиме цілу кількість символів усіх досліджуваних довжин, а кількість символів можна знайти, поділивши довжину згенерованого фрагмента на довжину символів.

Виконаємо розрахунки довжини циклу, на які розділена початкова двійкова послідовність, якщо найбільша довжина аналізованих символів становить $n_{\max} = 10$. Найменший спільний дільник чисел *{1,2,3,4,5,6,7,8,9,10}* дорівнює 2520. Кількість підпослідовностей різних довжин можна знайти, поділивши довжину згенерованого фрагмента на довжини двійкових символів (табл. 2).

Таблиця 2

Кількість символів, для яких розраховано частоти входження після генерування 2520 біт послідовності

Довжина символів	Кількість символів	Розмірність масиву для розрахунку частот входження символів
1	2520	array1[2]
2	1260	array2[4]
3	840	array3[8]
4	630	array4[16]
5	504	array5[32]
6	420	array6[64]
7	360	array7[128]
8	315	array8[256]
9	280	array9[512]
10	252	array10[1024]

Всередині кожного циклу виконуються обчислення частот входження усіх досліджених символів, одного за одним. Схему алгоритму наведено на рис. 1.

Для тестування ефективності розробленого алгоритму досліджено такі параметри роботи програми, як час компіляції коду та час її виконання, залежно від максимальної довжини символів, на які фрагментована початкова двійкова послідовність.

Результати тестування наведено у табл. 3 для алгоритму, що реалізує аналіз ентропії послідовності, розділеної на символи, максимальна довжина яких становить від 5 до 10 символів.

Рис. 1. Схема алгоритму обчислення ентропії

Таблиця 3

Результати тестування швидкодії розробленого алгоритму та його програмної реалізації у середовищі DEV C++ із застосуванням чотириядерного процесора AMD із тактовою частотою 1,80 GHz

Довжина символів	Мінімальна довжина послідовності	Час компіляції коду програми, с	Час виконання для мінімальної довжини послідовності	Час виконання для послідовності із 1000000 двійкових символів
5	60	3,91	0,18	9,37
6	60	3,99	0,18	11,32
7	420	3,89	0,20	10,52
8	840	4,05	0,22	10,93
9	2520	3,95	0,22	10,83
10	2520	3,88	0,22	10,57

Результати, наведені у табл. 3, показують, що цей алгоритм високоефективний із погляду швидкодії, його легко застосовувати для послідовностей великої довжини із використанням обчислювальних ресурсів типу Arduino. Тому його можна використовувати у технологіях Інтернету речей.

4. Результати дослідження ентропії

Дослідження ентропії для двійкової послідовності, генерованої за логістичним відображенням, задано рекурентним виразом (5):

$$x_{n+1} = 4 \cdot x_n \left(1 - x_n\right). \tag{5}$$

Формування двійкової послідовності на основі (5) відбувалося із порівнянням дійсних чисел x_n, утворених на кожному кроці ітерації, із пороговим значенням 0,5. Якщо число, утворене на черговому кроці ітерації, перевищує порогове значення, то генерується біт 0, інакше – біт 1.

Здійснено обчислення ентропії у широкому діапазоні таких параметрів генерування послідовності, як початкове значення рекурентного виразу (5) та довжина послідовності.

Розглянемо спочатку питання про вплив початкового значення послідовності. З великим ступенем достовірності можна стверджувати, що ентропія залишається практично незмінною у разі його зміни. Такий висновок можна зробити тому, що найбільше значення дисперсії ентропії, обчисленої на множині початкових значень послідовності, становить лише 0,5 % для послідовності найменшої довжини 2520 біт, розділеної на символи завдовжки 10 біт. У разі збільшення довжини послідовності дисперсія істотно зменшується і становить лише 0,02 % для максимальної дослідженої довжини послідовності 200100 біт, розділеної на символи по 10 біт.

На рис. 2 наведено залежність ентропії двійкової послідовності від значення довжини символів, на які вона поділена. На цьому рисунку подано усереднені значення, які обчислені для широкого діапазону початкового параметра генерування та довжини усієї послідовності. Початкове значення змінювалося із кроком 0,1 на інтервалі [0;1], довжина послідовності змінювалася від 2520 до більш як 200000 біт. Також на рис. 2 позначено максимальну ентропію, яка спостерігається для послідовностей із рівномірно розподіленими символами й обчислюється за формулою (3).

Також ми дослідили залежність ентропії від довжини послідовності. Загалом виявлено тенденцію зростання ентропії до її максимального значення зі збільшенням довжини послідовності. Для символів малої довжини ентропія одразу має значення, що майже збігаються із максимальним її значенням, як видно із результатів досліджень, наведених на рис. 2. Для символів більшої довжини ентропія спочатку є меншою, а потім поступово зростає до її максимального значення. Результати розрахунків залежності ентропії від довжини послідовності подано на рис. 3.

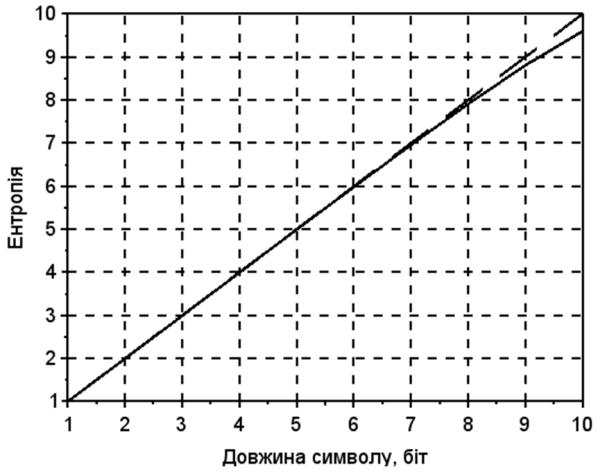

Рис. 2. Усереднені на множині початкових значень та довжин послідовностей значення ентропії. Штрихова лінія – максимальне значення ентропії

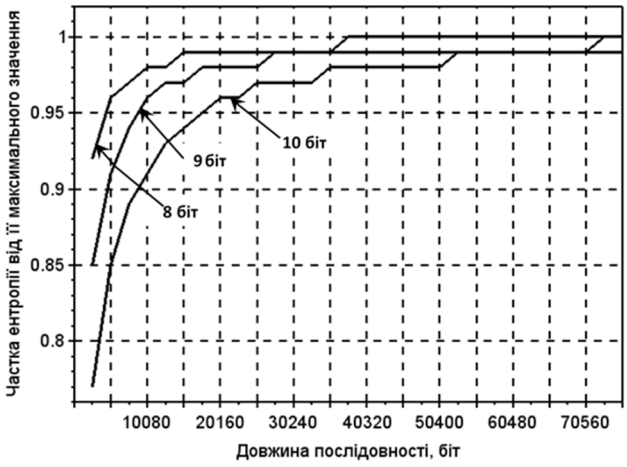

Рис. 3. Графік залежності ентропії, вираженої у частці від її максимального значення, від довжини послідовності

На графіках, наведених на рис. 3, ентропію відображено у вигляді частки від її максимального значення. Подано графіки для символів із 8, 9 та 10 біт, для яких найбільше її відхилення від максимального значення за невеликих значень довжини послідовності. Отже, зі збільшенням довжини символів рівномірний розподіл досягається тільки за деякого порогового значення довжини послідовності. Цю властивість ентропії можна використати для аналізу статистичної зв'язаності та корельованості псевдовипадкової двійкової послідовності.

Висновки

У роботі описано алгоритм обчислення ентропії двійкової послідовності, розділеної на символи завдовжки 1–10 біт. Алгоритм реалізовано у програмному середовищі DEV C++. За допомогою розробленої програми досліджено ентропію двійкових послідовностей, генерованих за допомогою

логістичного відображення на усій множині початкових значень цього відображення, взятих із кроком 0,1. На основі результатів дослідження ми дійшли висновку, що початкове значення не виявляє впливу на ентропію послідовності для усіх досліджених символів, на які її можна розділити. До того ж значення ентропії практично збігаються із її максимальним значенням, яке відзначається для рівноймовірного розподілу символів у послідовності. Це означає, що досліджені послідовності мають хороші статистичні властивості, тому їх можна використати в алгоритмах "легкої криптографії", які набувають істотного поширення у зв'язку із запровадженням протоколів передавання інформації у технологіях Інтернету речей. Дослідження швидкодії алгоритму дають підстави стверджувати, що він може слугувати ефективним засобом дослідження якості генерованих послідовностей та апробації нових алгоритмів їх генерування.

Список використаних джерел

[1] Nuno, J. and Munoz, F. (2022), "Entropy-Variance Curves of Binary Sequences Generated by Random Substitutions of Constant Length", Entropy, Vol. 24, No. 2, 290.

[2] Soto, J. and Bassham, L. (2000). "Randomness Testing of the Advanced Encryption Stadard Finalist Candidates". Available at:https://www.researchgate.net/publication/2611325_Randomness_Testing_of_the_Advanced_Encryption_Standard_Finalist_Candidates.

[3] Orúe López, A. B. et al. (2017), "Lightweight Pseudorandom Number Generator for Securing the IoT", IEEE Access, Vol. 5, pp. 27800–27806.

[4] EPC Global (2015) "UHF Air Interface Protocol Standard Generation2/V2.0.1". Available at: http://www.gs1.org/epcrfid/epc-rfid-uhf-airinterface-protocol/latest (accessed on Sept, 27, 2017).

[5] Політанський Р. Л. (2020) "Дослідження періодичності псевдовипадкових послідовностей методом булевого гіперкубу", Вчені записки ТНУ ім. В. І. Вернадського. Серія: технічні науки, том 31(70), ч. 1, No. 2, pp. 145–151.

INVESTIGATION OF THE ENTROPY OF BINARY SEQUENCES FRAGMENTED INTO FIXED-LENGTH SUBSEQUENCES OF SYMBOLS

R. Politanskyi, V. Kachur

Yuriy Fedkovych Chernivtsi National University, 2, Kotsyubynsky, Chernivtsi, 58012, Ukraine

The article investigates the structural entropy of sequences that are fragmented into binary subsequences (symbols) of a given length, based on the frequency of these symbols in the total sequence. The sequences generated by the logistic mapping with the initial value of the sequences from the interval [0;1] and the step 0.1 are investigated. The maximum length of a symbol (subsequence) is 10 bits. A comparison of the calculated entropy values with its maximum value shows that there is a deviation from a uniform distribution for symbols with a length of 8 or more bits, but the entropy value gradually approaches the maximum with an increase in the length of the subsequence. It is also established that such a generation parameter as the initial value does not affect the entropy. The conducted studies also show that the algorithm is effective in terms of high speed and does not require the use of significant computing power.

Key words: *light cryptography; Internet of things; pseudorandom sequences; logistic mapping.*