УДК 681.32 **Сушко С.О., Фомичова Л.Я., Киценко П.С.**

ІНДЕКС ЗБІГУ ДЛЯ ОСМИСЛЕНИХ ТЕКСТІВ УКРАЇНСЬКОЮ МОВОЮ

Як відомо, злам шифру модульного гамирування, для якого гама є періодичною послідовністю знаків абетки, проводиться в два етапи [1–4], на першому з яких обчислюється період гами, а на другому – знаходиться сама гама. Для визначення періоду гами застосовується тест Казіскі[5], оснований на періодичності потоку ключів і частому повторюванні певних буквосполучень у природних мовах. Це означає, що коли в криптограмі зустрічаються два однакових відрізка, то з великою імовірністю відстань між ними кратна довжині періоду гами. Для уточнення періоду гами і її подальшого визначення використовують уведений у практику криптоаналізу У.Фридманом [1] індекс збігу – ймовірність збігу двох навмання вибраних букв у рядку довжини n, складеному з букв абетки.

Якщо ототожнити букви абетки з кільцем лишків $Z_m = \{0,1,...,m-1\}$ за модулем m, то індекс збігу $I_{зб.}(x)$ для рядка визначається за формулою:

$$I_{зб.}(x) = \frac{\sum_{i=0}^{m-1} f_i(f_i - 1)}{n(n-1)}, \qquad (1)$$

де f_i – частота появи букви a_i у рядку, $i \in Z_m$.

Для природної мови можна обчислити взаємний індекс збігу як суму квадратів ймовірностей p_i появи символів абетки у відкритих текстах:

$$I_{зб.мови}(x) = \sum_{i=0}^{m-1} p_i^2 . \qquad (2)$$

Так, індекси збігу деяких європейських мов дорівнюють: 0,0529 – для російської, 0,0662– для англійської, 0,0778 – для французької, 0,0762– для німецької, 0,0738 – для італійської, 0,0775 – для іспанської. Для української мови подібні розрахунки досі не проводились.

Якщо N - період гами, то для рядка, складеного з кожної n-ої букви криптограми, взаємний індекс збігу має залишитися близьким до індексу збігу мови, на якій написано відкритий текст. Якщо послідовно перебрати можливі періоди та обчислити для них індекси збігу, то той період, для якого індекс буде максимально близьким до індексу збігу мови, є шуканим істинним періодом гами шифру.

Оцінимо індекс збігу української мови. При такій оцінці, зазвичай, за ймовірності p_i приймають еталонні частоти повторюваності букв у відкритих текстах. Для української мови частоти повторюваності букв наведені у роботах [6, 7], але обчислювались вони або за вибіркою невеликого об'єму, або з врахуванням апострофу, який зазвичай не зберігається при шифруванні. Автори перевірили розрахунки частот повторюваності букв в українській мові на вибірці об'ємом понад дев'ятсот тисяч слів. Використовувались навмання вибрані тексти, що належали до п'яти стилів сучасної української мови: розмовно-побутового, художнього, наукового, публіцистичного та ділового. Обчислені частоти записані у табл. 1.

Табл. 1. Середньостатистичні частоти букв та пропуску між словами в українській мові

□□□	0,138	**і**	0,044	**д**	0,027	**г**	0,013	**ж**	0,007
о	0,086	**р**	0,043	**л**	0,027	**ч**	0,011	**ю**	0,008
н	0,068	**е**	0,042	**п**	0,025	**х**	0,011	**є**	0,005

а	0,064	**с**	0,037	**з**	0,020	**ї**	0,010	**щ**	0,004
и	0,055	**к**	0,033	**я**	0,019	**ц**	0,010	**ф**	0,003
в	0,046	**м**	0,029	**ь**	0,016	**ш**	0,005	**ґ**	0,000
т	0,045	**у**	0,027	**б**	0,013	**й**	0,009		

Це дає змогу визначити наближене значення індексу збігу української мови:

$$I_{зб.мови}(x) = \sum_{i=0}^{33} p_i^2 \approx 0,0575. \tag{3}$$

Очевидно, якщо текстовий рядок є реалізацією незалежних випробувань випадкової величини з рівномірним розподілом, то індекс збігу неосмисленого рядка дорівнював $I_{зб.}(x) = \frac{1}{m}$. У випадку використання української абетки при врахуванні пропуску між словами $I_{зб.}(x) \approx 0,029$.

Якщо після визначення періоду гами (позначимо його d) виписати символи криптограми рядками у d стовпців Y_1, Y_2,..., Y_d, то криптоаналіз шифру гамирування можна провести за допомогою взаємних індексів збігу $MI_{зб.}(Y_i, Y_j^s)$ стовпців при відносному зсуві символів на величину s:

$$MI_{зб.}(Y_i, Y_j^s) = \frac{\sum_{t=0}^{m-1} f_t' \cdot f''_{t-s(\mathrm{mod}\, m)}}{n' \cdot n''}, \tag{4}$$

де f_t' – частота повторюваності в стовпці Y_i букви, що має в абетці номер t; $f''_{t-s(\mathrm{mod}\, m)}$ – частоти повторюваності в стовпці Y_j букви, що має в абетці номер $t - s(\mathrm{mod}\, m)$; $0 \le s \le m-1$; $1 \le i < j \le d$; d – період гами; n' і n'' – кількість букв у стовпцях Y_i і Y_j відповідно; m – кількість букв абетки. Усього можна побудувати $m \cdot C_d^2$ різних взаємних індексів збігу.

Залежність взаємних індексів збігу від різних зсувів для української абетки демонструють дані з табл. 2 (зсуви s і $m - s$ дають однакові взаємні індекси збігу). Ненульовим зсувам відповідають взаємні індекси збігу від 0,021 до 0,037, а найбільше значення взаємного індексу збігу $0,057$ виникає при відсутності зсуву, тобто при $s = 0$.

Табл.2. Взаємні індекси збігу для української абетки при різних зсувах

зсув s	0	1	2	3	4	5	6	7	8
$MI_{зб.}$	0,057	0,037	0,026	0,031	0,026	0,023	0,025	0,026	0,023
зсув s	9	10	11	12	13	14	15	16	17
$MI_{зб.}$	0,021	0,025	0,034	0,034	0,026	0,024	0,036	0,036	0,030

Очевидно, взаємний індекс зсуву $MI_{зб.}(Y_i, Y_j^s)$ будь-яких двох стовпців криптограми буде близьким саме до значення 0,057 тільки за умови, що зсув s між цими рядками справжній і визначається гамою шифру.

Виконані експериментальні дослідження з обчисленням довжини періоду гами за допомогою метода Фрідмана показують, що при довгих гамах (більше 50, а інколи 30 знаків) метод дає збій і визначає збільшені періоди гами порівняно із реальними. Подібні результати описані в роботі [8]. Це обумовлено тим, що для збору інформації про довгі періоди гам потрібні достатньо довгі криптограми.

Список літератури

1. Friedman W.F. The Index of Coincidence and Its Application in Cryhtography. Riverbank Publication, 1920, № 22. Geneva IL: Riverbank Labs.
2. King J., Bahler D. An Implementation of Probabilistic Relaxation in the Cryptanalysis of Simple Substitution Ciphers. Cryptologia, 1992, № 16(3), p. 215 – 225.
3. Matthews R. An Empirical Method for Finding the Keylenght of Periodic Ciphers. Cryptologia, 1988, № 12(4), p. 220 – 224.
4. Конхейм А.Г. Основы криптографии. Москва, «Радио и связь», 1987, 412 С.
5. Kasiski, F.W. 1863. Die Geheimschriften und die Dechiffrir-Kunst. Berlin: E. S. Mittler und Sohn
6. Вербіцький О.В. Вступ до криптології. Видавництво науково-технічнї літератури, Львів 1998, 247 с.
7. Перебийніс В.І., Муравицька М.П., Дарчук Н.П. Частотні словники та їх використання. К.: Наукова думка, 1983.
8. Алферов А.П., Зубов А.Ю., Кузьмин А.С., Черемушки А.В. Основі криптографии. Москва, Гелиос, 2002, 480 с.

Рецензент: Жуков І.А.
Надійшла 30.09.2010