УДК 004.056.53

ДОСЛІДЖЕННЯ ВЗАЄМНОГО ВПЛИВУ ПРОЦЕДУР ШИФРУВАННЯ ТА СТИСНЕННЯ ІНФОРМАЦІЙНОГО ПОВІДОМЛЕННЯ

Юдін О. К., д-р техн. наук, проф., ***Курінь К. О.***

Національний авіаційний університет

kszi@ukr.net

Проведено аналіз залежності ефективності методів захисту інформації від статистичних характеристик вихідних повідомлень. Розглянуто ентропійні характеристики повідомлень: умовні ентропії відкритого тексту відносно шифротексту, які характеризують ненадійність криптосистеми; ентропії нестиснених повідомлень, які визначають їх надмірність. Створено математичну модель процедури дослідження впливу комбінування процедур стиснення та шифрування на стійкість відкритого тексту та ступінь стиснення. Сформовану модель реалізовано програмно засобами Cryptool та Mathcad. Проведено розрахунок статистичних характеристик вихідних та модифікованих текстів.

Ключові слова: шифрування даних, стиснення даних, ентропія, умовні ентропії.

The analysis of dependence of efficiency of methods of data protection from statistical descriptions of messages is provided. The entropy characteristics of messages are considered: the entropy of plaintext of related to the ciphertext, which characterize the unreliability of cryptosystem; entropy of the uncompressed message, which determine its surplus. The mathematical model of procedure of research of influence of combination of proce-?ures of compression and enciphering on firmness of plaintext and degree of compression is created. The formed model is realized with the programmatic facilities of Cryptool. The statistical characteristics of initial and modified texts are calculated.

Keywords: ciphering of information, compression of data, entropy, conditional entropy.

Вступ

У зв'язку з лавиноподібним поширенням комп'ютерних систем та їх взаємодією через мережі спостерігається все більша залежність як організацій, так і окремих користувачів від інформації, що передається мережею та зберігається у таких системах.

Інформація стає все більш уразливою за таких причин:

- зростають обсяги даних, що зберігаються та передаються;
- розширюється коло користувачів, що мають доступ до ресурсів систем, програм і даних;
- ускладнюються режими експлуатації обчислювальних систем.

Це, у свою чергу, змушує усвідомити необхідність захисту даних та ресурсів від можливості несанкціонованого доступу, важливість використання спеціальних засобів для забезпечення достовірності отриманих даних, а також запобігання мережним атакам.

Специфіка сучасних підходів до захисту інформації визначається необхідністю забезпечення коректного функціонування об'єднаних мереж. Основні вимоги до безпеки комунікацій та обміну даними в мережі можуть бути представлені чотирма термінами: «конфіденційність», «автентифікація», «збереження цілісності» та «відсутність спотворень».

Захист інформації це сукупність заходів, методів і засобів, що забезпечують:

- убезпечення несанкціонованого доступу (НСД) до ресурсів систем, програм і даних;
- перевірку цілісності інформації;
- виключення несанкціонованого використання програм (захист програм від копіювання).

Очевидна тенденція до переходу на цифрові методи передачі і зберігання інформації дає змогу застосовувати уніфіковані методи і алгоритми для захисту дискретної (текст, факс, телекс) і безперервної (мова) інформації.

Один з найефективніших методів захисту інформації від НСД — шифрування (криптографія). Шифруванням називають процес перетворення відкритих даних у зашифровані (шифротекст) згідно з певними правилами із застосуванням ключів [1].

За допомогою криптографічних перетворень забезпечують:

- шифрування інформації;
- реалізацію електронного підпису;
- розподіл ключів шифрування;
- захист від випадкової або умисної зміни інформації.

Криптографічні заходи широко застосовуються при організації захищеної передачі даних з використанням сучасних засобів мережного захисту: системи *Kerberos*, сертифікатів X.509 *PGP*, *S/MIME*, *IP Security*, *SSl*, *SET* тощо.

Більшість описаних схем захисту поряд з криптографічними перетвореннями застосовують також попередню процедуру стиснення даних.

Алгоритми стиснення даних дуже добре підходять для сумісного використання з криптографічними алгоритмами. На це є дві причини:

1) у разі зламування шифру криптоаналітик спирається на надлишковість, властиву будь-якому відкритому тексту. Стиснення допомагає позбавитися від цієї надлишковості. Через те, що стиснене повідомлення має меншу надмірність порівняно з оригінальним відкритим текстом, криптоаналіз стає більш ускладненим.

2) шифрування даних є досить трудомісткою операцією. Під час стиснення зменшується довжина відкритого тексту, і тим самим скорочується час, який витрачатиметься на його шифрування.

Разом з тим застосування процедури стиснення після попереднього шифрування даних є небажаним, оскільки статистичні характеристики шифротексту максимально наближені до статистичних характеристик випадкового набору символів. Через це надмірність такого тексту буде мінімальною, та ефективне стиснення спостерігатися не буде.

Постановка задачі

Як було зазначено, методи шифрування даних, так само як і засоби стиснення інформаційного потоку даних, є важливим інструментом організації захищеного та коректного сеансу передачі інформації комунікаційними мережами. Тому цілком справедливо можна говорити про необхідність дослідження результатів синтезу процедур шифрування та компресії.

Мета статті — дослідження комплексного перетворення даних, що забезпечується використанням процедур шифрування та стиснення, з виконанням перестановки даних процедур.

Крім того, до задач дослідження належать створення програмно-математичної моделі, що реалізує зазначені перетворення, та розрахунок кількісних характеристик вихідного повідомлення, які визначають його придатність до процесів компресії та шифрування.

Критерії оцінки інформативності вихідного повідомлення

У основу сучасної теорії інформації покладено кількісне визначення інформації через імовірнісні характеристики процесу утворення даного виду інформації.

Розглянемо випадок дискретної інформації, де повідомлення, предмет передачі через канал зв'язку, складається з послідовності дискретних символів, кожен з яких вибраний з деякої скінченої множини (алфавіту) букв повідомлення.

Нехай задана ймовірнісна схема

$$A=\begin{pmatrix} 1 & 2 & \dots & n \\ p(1) & p(2) & \dots & p(n) \end{pmatrix},$$

де $1, 2, ..., a, ..., n$ — виходи (букви, повідомлення) ймовірнісної схеми з множини алфавіту; $A=\{1, 2, ..., a...n\}$, $p(1), p(2), ..., p(a), ..., p(n)$ — імовірності цих повідомлень, такі, що $\sum_{a=1}^{n} p(a)=1$.

Основоположником теорії чисельної оцінки міри невизначеності ймовірнісних схем є американський інженер і математик Клод Шеннон. Нижче наводяться його основні ідеї в розв'язанні даної задачі.

Якщо є така міра (позначимо її через $H=H(A)=H(p(1), p(2),...p(a),..., p(n))$), то слід аксіоматично потребувати, щоб вона характеризувалася такими властивостями [2]:

- H має бути безперервною відносно $p(a), a \in A$;
- H має бути симетрична щодо своїх аргументів (тобто $H(p(1), p(2),...p(a),..., p(n)) = H(p(i_1), p(i_2),...p(i_n))$ для будь-якої перестановки індексів);
- якщо вибір розпадається на два послідовних вибори, то первинна H має бути зваженою сумою індивідуальних значень.

Тобто при $(p(1)+p(2))>0$ справедлива рівність:

$$H(p(1)p(2), p(3),..., p(n)) + \\ +(p(1)+p(2))H\left(\frac{p(1)}{p(1)+p(2)}, \frac{p(2)}{p(1)+p(2)}\right).$$

Зміст третьої властивості ілюструє рис. 1.

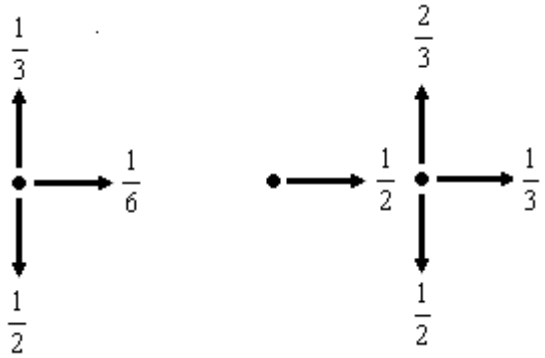

Рис. 1. Властивість ентропії

У цьому конкретному випадку згідно з властивістю 3 вимагають, щоб

$$H\left(\frac{1}{3}, \frac{1}{6}, \frac{1}{2}\right)=H\left(\frac{1}{2}, \frac{1}{2}\right)+\frac{1}{2}H\left(\frac{1}{3}, \frac{2}{3}\right).$$

Ваговий множник $\frac{1}{2}$ введений через те, що другий вибір проводиться в половині випадків.

Виявляється, що наведені три умови з точністю до постійного коефіцієнта c визначають функцію $H(p(1), p(2),..., p(n))$:

$$H(p(1), p(2),..., p(n)) = -c\sum_{a=1}^{n} p(a)\log_d\ p(a),$$

де $c = \text{const}$.

Вибір константи рівносильний вибору основи d логарифма, а її значення та значення основи можна трактувати як визначення масштабу одиниці кількості невизначеності.

Найчастіше як одиницю невизначеності обирають невизначеність, що міститься в альтернативній відповіді «да-ні», що відповідає ймовірнісній схемі $\left(p(1)=\frac{1}{2}, p(2)=\frac{1}{2}\right)$. У цьому випадку основа логарифма дорівнює двом:

$$H = -\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}\right)=1.$$

Цю одиницю називають *бітом*, вона в $3{,}32 = \log_2 10$ рази менша за десяткову одиницю вимірювання інформації.

Таким чином, остаточний вираз для H має вигляд:

$$H(p(1), p(2),..., p(n)) = -\sum_{a=1}^{n} p(a)\log_2 p(a).$$

Кількістю інформації (невизначеністю) в повідомленні $a \in A$ називається число $h(a)$, що визначається співвідношенням [2]:

$$h(a) = -\log_2\ p(a).$$

Середнє значення кількості інформації [3]

$$H(A) = -\sum_{a=1}^{n} p(a)\log_2\ p(a) \qquad (1)$$

називається ентропією кінцевої ймовірнісної схеми.

Розглянемо разом зі схемою

$$A = \begin{pmatrix} 1 & 2 & ... & n \\ p(1) & p(2) & ... & p(n) \end{pmatrix}$$

схему

$$B = \begin{pmatrix} 1 & 2 & ... & m \\ q(1) & q(2) & ... & q(m) \end{pmatrix}$$

Схеми A і B можуть бути залежні між собою, тобто для ймовірності $p(a,b)$ одночасного виконання подій a,b виконується нерівність: $p(a,b) \neq p(a)p(b)$. Розглянемо об'єднану схему вигляду

$$C = \begin{pmatrix} 11 & 12... & ab & nm \\ p(11) & p(12)... & p(ab) & p(nm) \end{pmatrix};$$

$$p(a) = \sum_{b=1}^{m} p(ab); q(b) = \sum_{a=1}^{n} p(ab);$$

Відповідно до визначення (1), ентропія

$$H(AB) = -\sum_{a,b} p(ab)\log_2\ p(ab)$$

називається ентропією об'єднаної схеми AB.

Позначаючи через $p(b/a) = p(ab)/\ p(a), b \in B$ умовні вірогідності повідомлень схеми B за умови повідомлення a, можна розглянути ряд «умовних» імовірнісних схем

$$B/a = \begin{pmatrix} 1 & 2 & ... & nm \\ p(1/a) & p(2/a) & ... & p(m/a) \end{pmatrix}.$$

Для кожної з цих схем згідно з виразом (1) можна визначити ентропію

$$H(B/a) = -\sum_{b=1}^{m} p(b/a)\log_2\ p(b/a).$$

Середнє значення

$$H(B/A) = -\sum_{a=1}^{n} p(a)\sum_{b=1}^{m} p(b/a)\log_2\ p(b/a). \qquad (2)$$

називається ентропією схеми B за умов схеми A [2].

Розглянемо, яким чином ці положення теорії інформацій застосовні до визначення характеристик вихідного повідомлення, які визначають його придатність до процесів компресії та шифрування.

Шифр (M, K, E, f) містить у собі два ймовірнісних вибори: вибір відкритого повідомлення $m \in M$ та вибір ключа $\chi \in K$.

Тим самим визначена ймовірнісна модель шифру шифрування. Кількість інформації, що створюється при виборі відкритого повідомлення, вимірюється величиною

$$H(M) = -\sum_{m \in M} p(m)\log_2\ p(m).$$

Аналогічно невизначеність, пов'язана з вибором ключа, задається виразом:

$$H(K) = -\sum_{m \in K} p(\chi)\log_2 p(\chi).$$

Невизначеність повідомлення, так само як і невизначеність ключа може змінюватися, коли є можливість спостерігати криптограму — шифрований текст $e \in E = f(M, K)$. Цю умовну невизначеність природно вимірювати умовною ентропією:

$$H(M/E) = -\sum_{(m,e)\in E\times M} p(m,e)\log_2\ p(m/e); \qquad (3)$$

$$H(K/E) = -\sum_{(\chi,e)\in E\times K} p(\chi,e)\log_2 p(\chi/e).$$

Введені умовні ентропії Шенон назвав ненадійністю відкритого повідомлення і ключа. Ці ненадійності використовуються як теоретична міра секретності. Як обґрунтування такого використання можна навести такі міркування. Якщо ненадійність повідомлення (ключа) дорівнює нулю, то звідси витікає, що лише одне повідомлення (один ключ) має одиничну апостеріорну

вірогідність, а всі інші – нульову. Цей випадок відповідає повній обізнаності криптоаналітика про повідомлення (ключі).

Дійсно, нехай

$$H(M/E) = -\sum_{(m,e)\in E\times M} p(m,e)\log_2 p(m/e) = 0.$$

Тоді за будь-яких $(m,e)\in(M\times E)$:

$$p(m,e)\log_2 p(m/e) = 0.$$

Виберемо такі (m,e), для яких існує ключ $\chi\in K$, такий, що $\chi m = e$. Для таких пар маємо $(m,e) > 0$. Отже, з попередньої рівності отримуємо $\log_2 p(m/e) = 0$, тобто $p(m/e) = 1$. Таким чином, з шифротексту e відкритий текст m відновлюється однозначно.

Згідно з теоремою Шенона шифр (M,K,E,f), де $E = f(M,K)$, вважається досконалим тоді, коли $H(M/E) = H(M)$.

На практиці ж $H(M/E) > H(M)$.

Алгоритми стиснення підвищують ефективність зберігання та передачі даних за допомогою скорочення їх надмірності. Алгоритм стиснення отримує на вході вихідний текст джерела та формує відповідний йому стислий текст, тоді як розгортаючий алгоритм виконує обернені перетворення. Більшість алгоритмів стиснення розглядають початковий текст як набір рядків, що складаються з букв алфавіту вихідного тексту.

Поняття ентропії в теорії стиснення застосоване наступним чином: під ентропією символа a, що характеризується ймовірністю $P(a)$, розуміють кількість інформації, яка міститься в даному символі і, дорівнює $-P(a)\log_2 P(a)$.

За допомогою ентропії визначається також кількість інформації, яка міститься в рядку символів алфавіту. Для рядка S, який складається з символів алфавіту $\{a_1, a_2, ..., a_n\}$, які характеризуються ймовірностями $\{P(a_1), P(a2), .., P(an)\}$, ентропія визначається за таким виразом:

$$H(S) = -\sum_{i=1}^{n} P(a_i)\log_2 P(a_i). \qquad (4)$$

Дане значення визначає найкраще стиснення цього рядка, тобто найменшу кількість бітів, необхідних для його представлення.

Надмірністю в представленні рядка S називається величина $D(S) = L(S) - Hbit(S)$, де $L(S)$ — довжина повідомлення в бітах, а $Hbit(S)$ — ентропія, також виражена в бітах. Зрозуміло, що чим більша надмірність повідомлення, тим краще стиснення. Алгоритмів, які могли б без втрати інформації стискати рядок до меншого числа битів, ніж складає його ентропія, не існує.

Ефективність процедури стиснення оцінюється характеристикою, що називається коефіцієнтом стиснення й чисельно визначається так:

$$k = \frac{L(S')}{L(S)}, \qquad (5)$$

де $L(S')$ — довжина повідомлення, отриманого на виході кодера в результаті стиснення рядка S, виражена в бітах.

Дослідження впливу процедур стиснення та шифрування на вихідне повідомлення

У текстовому редакторі було створено файл test.txt, що містить вихідне повідомлення (M). В програмному середовищі *Mathcad* було створено програмний модуль який реалізовує стиснення вихідного текстового рядка та шифротексту за допомогою статистичного алгоритму Хаффмена [4]. У цьому ж програмному модулі проводиться розрахунок коефіцієнта стиснення, що забезпечується кодером, та ентропії нестиснених повідомлень. Шифрування вихідного тексту та його стислого представлення, а також криптоаналіз отриманих на їх підставі шифротекстів проводили засобами програмного продукту *CryptTool* v1.4.30. Крім того, в програмному середовищі *Mathcad* було розраховано значення ненадійності вихідних повідомлень.

Для оцінювання ефективності комбінації процедур стиснення та шифрування, що застосовується до вихідного повідомлення M, виконаємо такі дії:

1. До вихідного повідомлення M (test.txt) застосовується процедура стиснення методом Хаффмена, в результаті чого отримується його стисле представлення C (test_compressed.txt). За формулою (4) розраховується ентропія вихідного повідомлення $H(M)$, за формулою (5) — значення коефіцієнта стиснення $k1$.

2. Отриманий стислий текст C шифрується за допомогою криптографічного алгоритму *RC*4 [5; 6] з використанням 24-бітового ключа (мала розмірність ключа вибрана задля того, щоб у прийнятні строки здійснити криптоаналіз зашифрованого повідомлення). У результаті шифрування формується шифротекст CE (test_compressed_crypt.txt). Згідно з формулою (3) розраховується ненадійність повідомлення, поданого на вхід процедури шифрування $H(C/CE)$.

3. Вихідне повідомлення M шифрується за допомогою знову ж таки алгоритму *RC*-4 при використанні ключа, застосованого в п. 2. Це пертворення забезпечує формування шифротексту E (test_crypt.txt). Ненадійність вихідного повідомлення $H(M/E)$ оцінюється за формулою (4).

4. Отриманий шифротекст стискається за допомогою алгоритму Хаффмена, в результаті цього перетворення на виході кодера отримуємо стислий текст *EC* (test_crypt_compressed.txt). Ефективність процедури стиснення визначається шляхом розрахунку коефіцієнту стиснення $k2$ згідно (4). Розраховується ентропія $H(E)$ повідомлення, яке подається на віхд кодера.

Описані перетворення ілюструє рис. 2.

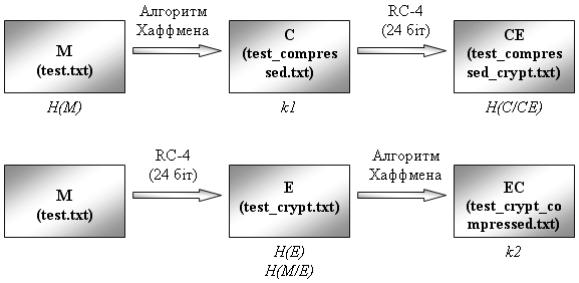

Рис. 2. Перетворення вихідного повідомлення

Результати розрахунків статистичних характеристик повідомлень, до яких застосовані процедури стиснення та шифрування, зведені в таблицю.

Ентропійні характеристики

$(H(M))$	0,0067	$H(C/CE)$	7,307	$k1$	1,796
$H(E)$	0,012	$H(M/E)$	4,581	$k2$	1,011

Отримані результати дають змогу зробити такі висновки:

— кількість інформації, що міститься в зашифрованому повідомленні більша за відповідну характеристику відкритого повідомлення (за даних умов — майже в два рази), що свідчить про те, що надмірність тексту, який піддається шифруванню, зменшується. При цьому очевидно знижується ефективність процедури стиснення, що ілюструється зменшенням значення коефіцієнта стиснення на 68 %;

— умовна ентропія повідомлення, яке попередньо було стиснене з використанням статистичного методу стиснення, в 1,595 разу перевищує відповідну характеристику нестисненого тексту, що свідчить про вищу криптостійкість криптосистеми, в якій перед шифруванням попередньо застосовується стиснення. На практиці такі ж самі висновки були отримані після спроби криптоаналізу закритих повідомлень *CE* та *E* засобами програми *CryptTool*. Спроба зламати шифротекст, отриманий з нестислого повідомлення, виявилась вдалою (нагадаємо, що розмірність ключа була для наглядності вибрана малою), у випадку аналізу стисненого повідомлення відновити відкритий текст не вдалося, оскільки метод криптоаналізу грунтується на виявленні певних статистичних закономірностей відкритого тексту, які були усунені кодером під час стиснення.

Висновки

Проведено аналіз залежності ефективності методів захисту інформації від статистичних характеристик вихідних повідомлень. Розглянуто ентропійні характеристики повідомлень:

- умовні ентропії відкритого тексту відносно шифротексту, які характеризують ненадійність криптосистеми;
- ентропії нестисненого повідомлення, які визначають його надмірність.

Створено математичну модель процедури дослідження впливу комбінування процедур стиснення та шифрування на стійкість відкритого тексту та ступінь стиснення. Сформовану модель реалізовано програмно засобами *Cryptool* та *Mathcad*.

Проведено розрахунок статистичних характеристик вихідних та модифікованих текстів. Аналіз отриманих результатів дав змогу зробити такі висновки:

- криптостійкість відкритого повідомлення при попередньому його шифруванні підвищується більш ніж у 1,5 разу, завдяки тому, що в результаті шифрування усуваються статистичні ознаки, властиві відкритому тексту, на підставі яких проводиться криптоаналіз;
- ефективність процедури стиснення зменшується майже в два рази за умови попереднього шифрування вихідного тексту, оскільки кодер сприймає шифротекст як набір випадкових символів, у якому майже відсутня інформаційна надмірність.

ЛІТЕРАТУРА

1. *Столингс Вильям*. Криптография и защита сетей: принципы и практика. — 2-е изд. / Вильям Столингс. — М.: Издательский дом «Вильямс», 2001. — 672 с.

2. *Бабаш А. В.* Криптография: аспекты защиты / А. В. Бабаш, Г. П. Шанкин. — М. : ЖСОЛОН-Р, 2002. — 512 с.

3. *Юдін О. К.* Кодування в інформаційно-комунікаційних мережах: монографія. — К. : НАУ, 2007. — 308 с.

4. *Селомон Д.* Стиснення даних, зображень і звуку / Д. Селомон. — М. :Техносфера, 2006. — 386 с.

5. *Коробейников А. Г.* Математические основы криптологии: учеб. пособ. / А. Г. Коробейников, Ю. А. Гатчин. — СПб. : СПб.ГУИТМО, 2004. — 106 с.

6. *Петров А. А.* Компьютерная безопасность. Криптографические методы защиты / А. А. Петров. — М. : ДМК, 2000. — 448 с.

Стаття надійшла до редакції 07.12.2010.