СИМЕТРИЧНА КРИПТОГРАФІЯ
КОМП'ЮТЕРНИЙ ПРАКТИКУМ №2

Криптоаналіз шифру Віженера

Мета роботи

Засвоєння методів частотного криптоаналізу. Здобуття навичок роботи та аналізу потокових шифрів гамування адитивного типу на прикладі шифру Віженера.

Необхідні теоретичні відомості

Нехай $A=\{a_0, a_1, ..., a_{m-1}\}$ – алфавіт відкритого (ВТ) та шифрованого (ШТ) текстів, що складається з m букв. Природнім чином можна замінити символи алфавіту їх номерами і перевести множину A у кільце $Z_m=\{0, 1, ..., m-1\}$ із відповідними операціями додавання та множення.

Шифр Віженера є прикладом поліалфавітної підстановки. Ключем цього шифру є послідовність r букв алфавіту $(k_0, k_1, ..., k_{r-1})$, яку підписують під ВТ, повторюючи стільки разів, скільки потрібно. Часто в якості ключа використовують якусь фразу або уривок тексту. Число r називається *періодом шифру Віженера*.

Позначимо ВТ через $X=x_0x_1x_2...x_n$, а ШТ через $Y=y_0y_1y_2...y_n$. Шифрування відбувається шляхом додавання букв ВТ до підписаних під ними букв ключа за модулем m, тобто

$$y_i=(x_i+k_{i \bmod r}) \bmod m,\ i=\overline{0,n}.$$

Криптоаналіз шифру Віженера починають з визначення періоду r. Зробити це можна тому, що шифр Віженера зберігає деякі статистичні властивості мови. Дійсно, розіб'ємо шифртекст Y на блоки

$$Y_0=y_0, y_r, y_{2r}, ...$$
$$Y_1=y_1, y_{r+1}, y_{2r+1}, ...$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$Y_{r-1}=y_{r-1}, y_{2r-1}, y_{3r-1}, ...$$

Кожен фрагмент Y_i фактично зашифрований шифром Цезаря з ключем k_i, $i=\overline{0,r-1}$. Звідси маємо, що значення частот символів у цих фрагментах будуть очікувано співпадати із значеннями імовірностей символів мови з точністю до перестановки. Це зауваження дозволяє побудувати розпізнавач періоду шифру Віженера, причому існує щонайменше два методи знаходження періоду.

Перший метод ґрунтується на понятті індексу відповідності.
Індексом відповідності тексту Y називається величина

$$I(Y)=\frac{1}{n(n-1)}\sum_{t\in Z_m} N_t(Y)(N_t(Y)-1),$$

де $N_t(Y)$ – кількість появ букви t у шифртексті Y. Якщо вважати, що текст Y обирається із множини можливих відкритих текстів випадково та рівноімовірно, то індекс відповідності буде випадковою функцією, а його математичне очікування дорівнюватиме $MI(Y) = \sum_{t \in Z_m} p_t^2$, де p_t – імовірність появи літери t в мові. Однак, якщо Y є шифртекстом, одержаним в результаті роботи шифру Віженера, то величина індексу відповідності та його математичне очікування буде стрімко падати із ростом довжини ключа r.

Для знаходження істинного значення r за допомогою індексу відповідності пропонується два можливих алгоритми. Перший алгоритм виглядає так:

1) Для кожного кандидата $r = 2,3,...$ розбити шифртекст Y на блоки $Y_1, Y_2, ..., Y_r$.
2) Обчислити значення індексу відповідності для кожного блоку.
3) Якщо сукупність одержаних значень схиляється до теоретичного значення I для даної мови, то значення r вгадане вірне. Якщо сукупність значень схиляється до значення $I_0 = \frac{1}{m}$, що відповідає мові із рівноімовірним алфавітом, то значення r вгадане неправильно.

Другий алгоритм використовує інший підхід.

1) Одержати оцінки індексу відповідності I_r для шифртекстів, що були зашифровані шифром Віженера із різними періодами r ($r \geq 2$).
2) Обчислити індекс відповідності даного шифтексту.
3) Порівнюючи обчислене значення із індексами I_r, зробити висновок щодо довжини ключа.

У першому алгоритмі для великих періодів починає, з одного боку, суттєво зменшуватись кількість статистики, а з іншого, росте кількість параметрів для порівняння, що приводить до різкого падіння точності; однак, якщо розміри блоків залишаються достатньо великими, цей метод дозволяє знаходити довжину ключа доволі точно. Замість розглядання великої сукупності індексів відповідності по кожному блоку на практиці зазвичай розглядають їх усереднене значення.

У другому алгоритмі для маленьких r (приблизно $2 \leq r \leq 5$) значення індексів I_r при різних r помітно відрізняються, тому довжина ключа може бути знайдена; але для великих розбіжності стають несуттєвими. Звідси бачимо, що застосування даного алгоритму для довгих періодів не ефективне.

Другий метод визначення довжини ключа шифру Віженера використовує такий факт: в шифртексті на відстанях, що кратні періоду, однакові символи будуть зустрічатись частіше, ніж на будь-яких інших. Цей факт пояснюється тим, що у введених вище блоках Y_i однакові символи будуть зустрічатись із тією самою імовірністю, що й у відкритому тексті, а на інших відстанях потрібно, щоб співпадали значення відповідних сум $x_i + k_i$, що виконується із меншою імовірністю.

Отже, в цьому випадку пропонується такий порядок дій для знаходження істинного значення r: для кожного кандидата $r = 6,7,...$ обчислити значення статистики збігів символів:

$$D_r = \sum_{i=1}^{n-r} \delta(y_i, y_{i+r}),$$

де $\delta(a,b)$ – символ Кронекера. Іншими словами, D_r дорівнює кількості пар однакових літер шифртексту, які знаходяться на відстані r символів. Для кандидатів, що рівні та кратні істинному періоду, значення D_r будуть істотно більшими за інші одержані значення.

Після встановлення значення періоду шифру подальше його розшифрування зводиться до серії розшифрувань шифрів Цезаря. Дійсно, кожен фрагмент Y_i зашифрований шифром Цезаря з ключем k_i, $i=\overline{1,r}$. Найпростіший спосіб знаходження ключа полягає в обчисленні $k_i=(y^*-x^*)\bmod m$, де y^* – буква, що частіше за всіх зустрічається у фрагменті Y_i, а x^* – найімовірніша буква у мові, якою написано відкритий текст (для російської мови це буква «о», для англійської – буква «е» тощо). Цей метод на практиці дозволяє визначити більшу частину літер достатньо довгого ключа. Якщо деяку літеру ключа було вгадано невірно (що визначається за спотворенням відкритого тексту після дешифрування), у відповідному блоці замість x^* треба брати другу, третю і т.д. за імовірністю літеру, або коригувати значення ключа відповідно до реконструкції тексту за правильно розшифрованими фрагментами. При розшифруванні деякі фрагменти будуть встановлені неправильно, але можливі помилки легко виправляються при аналізі розшифрованого тексту в цілому.

Більш надійний метод визначення ключа полягає в наступному. Для кожного блоку Y_i обчислюється функція

$$M_i(g)=\sum_t p_t N_{t+g}(Y_i),$$

де $N_x(Y_i)$ – кількість появ букви *x* у шифротексті, p_t – імовірність появи літери *t* в мові. Те значення g, на якому функція $M_i(g)$ буде досягати максимуму, дорівнює значенню літери ключа k_i. Цей метод враховує увесь розподіл частот літер у блоці, тому він дозволяє відновити літери ключа майже безпомилково.

Порядок виконання роботи

0. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.

1. Самостійно підібрати текст для шифрування (2-3 кб) та ключі довжини r = 2, 3, 4, 5, а також довжини 10-20 знаків. Зашифрувати обраний відкритий текст шифром Віженера з цими ключами.

2. Підрахувати індекси відповідності I_r для відкритого тексту та всіх одержаних шифротекстів і порівняти їх значення.

3. Використовуючи наведені теоретичні відомості, розшифрувати наданий шифртекст (згідно свого номеру варіанта). Зокрема, необхідно:

– визначити довжину ключа, використовуючи або метод індексів відповідності, або статистику співпадінь D_r (на вибір);

– визначити символи ключа, прирівнюючи найчастіші літери у блоці до найчастішої літери у мові;

– визначити символи ключа за допомогою функції $M_i(g)$;

– розшифрувати текст, використовуючи знайдений ключ; в разі необхідності скорегувати ключ.

Методичні вказівки

Тексти, зашифровані шифром Віженера у варіантах завдань, написані російською мовою без знаків пунктуації, великих літер та пробілу; буква «ё» замінена буквою «е». Загальна кількість літер у алфавіті $m = 32$.

Для оцінки теоретичного значення індексу відповідності та для обчислення функцій $M_i(g)$ користуйтесь значеннями частот символів мови, одержаних під час виконання першого комп'ютерного практикуму.

При пошуку періоду шифру Віженера потрібно перевіряти довжини ключів (обчислювати індекси відповідності блоків або значення статистики D_r) щонайменше до $r = 30$. У варіантах завдань використовувались змістовні ключі, що може прискорити для вас процес дешифрування.

Оформлення звіту

Звіт до комп'ютерного практикуму оформлюється згідно зі стандартними правилами оформлення наукових робіт, за такими винятками:

- дозволяється використовувати шрифт Times New Roman 12pt та одинарний інтервал між рядками;
- для оформлення фрагментів текстів програм дозволяється використовувати шрифт Courier New 10pt та друкувати тексти в дві колонки;
- дозволяється не починати нові розділи з окремої сторінки.

До звіту можна не включати анотацію, перелік термінів та позначень та перелік використаних джерел. Також не обов'язково оформлювати зміст.

Звіт має містити:

- мету комп'ютерного практикуму;
- постановку задачі та варіант завдання;
- хід роботи, опис труднощів, що виникали, та шляхів їх розв'язання;
- обчислені значення індексів відповідності I_r для вказаних значень r (подати у вигляді таблиці та діаграми);
- обчислену послідовність D_r або набори значень індексів відповідності, одержаних при встановленні довжини ключа шифру Віженера (подати у вигляді таблиці та діаграми);
- значення ключа, одержане шляхом співставлення найчастіших літер блоків найчастішій літері мови;
- значення ключа, одержане із використанням функції $M_i(g)$;
- скореговане значення ключа (за необхідності);
- фрагмент шифрованого тексту (відповідно до варіанту завдання) та результати його розшифрування усіма знайденими варіантами ключа – 5-10 рядочків;
- висновки.

Тексти всіх програм здаються викладачеві в електронному вигляді для перевірки на плагіат. До захисту комп'ютерного практикуму допускаються тільки ті студенти, які оформили звіт та пройшли перевірку програмного коду.

Контрольні запитання

1) На які види поділяються класичні шифри? У чому між ними відмінність?
2) Що таке шифри моно- та поліалфавітної підстановки?
3) Що таке шифр Віженера? Опишіть процеси зашифрування та розшифрування.
4) Що таке індекс відповідності?
5) Чому не потрібно підраховувати індекс відповідності для шифртексту з $r = 1$? Чому він дорівнює?
6) Яка модель відкритого тексту розглядається при криптоаналізі шифру Віженера?
7) Завдяки чому можливий криптоаналіз шифру Віженера?
8) Що таке частотний аналіз?
9) Яким чином визначається довжина ключа шифру Віженера при дешифруванні?
10) Яким чином визначається значення ключа шифру Віженера при дешифруванні?

Оцінювання практикуму

За виконання комп'ютерного практикуму студент може одержати до 7 рейтингових балів; зокрема, оцінюються такі позиції:

- реалізація програм – до трьох балів (в залежності від правильності та швидкодії);
- теоретичний захист роботи – до трьох балів;
- несвоєчасне виконання роботи – (-1) бал за кожні два тижня пропуску дедлайну.

Програмний код, створений під час виконання комп'ютерного практикуму, перевіряється на наявність неправомірних запозичень (плагіату) за допомогою сервісу *Stanford MOSS Antiplagiarism*. У разі виявлення в програмному коді неправомірних запозичень реалізація програм оцінюється у 0 балів, а за виконання практикуму студент одержує штраф (–10) балів.

Студенти допускаються до теоретичного захисту тільки за умови оформленого звіту з виконання практикуму та проходження перевірки програмного коду.