ОСНОВИ КРИПТОЛОГІЇ
КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту. Криптоаналіз шифру Віженера

Мета роботи

Засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела. Здобуття навичок роботи з поточними шифрами адитивного типу на прикладі шифру Віженера. Засвоєння методів частотного криптоаналізу.

Необхідні теоретичні відомості

1. Визначення ентропії імовірнісних ансамблів

Розглянемо множину символів $Z = \{z_1, z_2, ..., z_n\}$, в якій кожному символу x_i приписана імовірність $p_i = p(z_i)$. Пару $\langle Z, P \rangle$, де $P = \{p_1, p_2, ..., p_n\}$, ми називаємо *імовірнісним ансамблем*, або просто *ансамблем*.

Ентропія імовірнісного ансамблю $\langle Z, P \rangle$ – це величина

$$H(Z) = -\sum_{i=1}^{n} p_i \log p_i \ .$$

Значення ентропії різниться в залежності від того, яку основу логарифму використовувати. Тут і надалі вважається, що всюди використовується логарифм за основою 2; в цьому випадку одиницею виміру ентропії є біт.

Ентропія приймає значення із проміжку $0 \le H(Z) \le \log n$. Значення $H(Z) = 0$ досягається лише для вироджених розподілів, коли одне значення приймається із імовірністю 1, а всі інші – із імовірністю 0. Значення $H(Z) = \log n$ (максимальне можливе значення для ансамблю розміром *n*) приймається лише для рівноімовірного розподілу.

Розглянемо два ансамблі $\langle X, P \rangle$ та $\langle Y, Q \rangle$; побудуємо спільний ансамбль на множині $X \times Y$. Для цього необхідно задати сукупний розподіл імовірностей $p_{ij} = p(x_i, y_j)$, який повинен задовольняти вимогам нормування: $\sum_{i,j} p_{ij} = 1$, $\sum_{j} p_{ij} = p_i$, $\sum_{i} p_{ij} = q_j$.

Для побудованого таким чином ансамблю можна також визначити ентропію; вона має назву *сукупної ентропії*:

$$H(X,Y) = -\sum_{i,j} p_{ij} \log p_{ij} \ .$$

Разом із сукупною ентропією розглядають також *умовну ентропію*:

$$H(X \mid Y) = -\sum_j q_j H(X \mid y_j) = -\sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_j}.$$

Сукупна ентропія говорить нам, скільки інформації містять обидва ансамблі (із урахуванням взаємних залежностей). Умовна ентропія говорить нам, скільки інформації залишиться в ансамблі *X*, якщо поведінку ансамблю *Y* буде однозначно визначено.

Ансамблі називаються *незалежними*, якщо виконується умова: $\forall i, j : p_{ij} = p_i q_j$. В цьому випадку $H(X,Y) = H(X) + H(Y)$ і $H(X \mid Y) = H(X)$.

2. Методи оцінювання ентропії джерел символів

Текстом ми називаємо довільну послідовність символів $x_1, x_2, x_3, \ldots$ з деякого скінченного алфавіту *Z*, яку формує деяке джерело (генератор текстів). *n-грамою* називається відрізок тексту з *n* послідовних символів $(x_{i+1}, x_{i+2}, \ldots, x_{i+n})$, $i \geq 0$. Джерело вважається повністю описаним, якщо для кожної *n*-грами $(n \geq 1)$ заданий розподіл імовірностей $P(x_{i+1} = z_1, x_{i+2} = z_2, \ldots, x_{i+n} = z_n)$, $i \geq 0$, $z_j \in Z$, $1 \leq j \leq n$. Джерело називається *стаціонарним*, якщо

$$P(x_{i+1} = z_1, x_{i+2} = z_2, \ldots, x_{i+n} = z_n) = P(x_1 = z_1, x_2 = z_2, \ldots, x_n = z_n)$$

для всіх $i \geq 0$ і довільних $n \geq 1$, $z_j \in Z$, $1 \leq j \leq n$, тобто якщо розподіл всіх *n*-грам не залежить від зсуву за часом (або, що те ж саме, від місця появи в тексті).

Ентропія на символ стаціонарного джерела визначається як

$$H_\infty = \lim_{n \to \infty} H_n, \text{ де } H_n = \frac{1}{n} H(x_1, x_2, \ldots, x_n),$$

а $H(x_1, x_2, \ldots, x_n)$, в свою чергу, – ентропія *n*-грами відкритого тексту $(x_1, x_2, \ldots, x_n)$:

$$H(x_1, x_2, \ldots, x_n) = -\sum_{z_1, z_2, \ldots, z_n} P(x_1 = z_1, \ldots, x_n = z_n) \cdot \log_2 P(x_1 = z_1, \ldots, x_n = z_n).$$

Величина H_n називається *питомою ентропією на символ n-грами*, а H_∞ – *ентропією джерела* або *ентропією мови*. Максимальне значення H_∞ приймає в тому випадку, коли всі символи тексту незалежні і рівноімовірні. Тоді $H_\infty = H_0 = \log_2 m$, де *m* – кількість букв в алфавіті *Z*.

Для реальних джерел відкритого тексту (таких, як природні мови) значення H_∞ набагато менше за H_0 через нерівноімовірність букв алфавіту в тексті та залежність між ними. Як послідовні наближення до H_∞ можна розглядати значення $H_1, H_2, H_3, \ldots$, які враховують відповідно імовірності букв алфавіту в мові, зв'язок букв всередині біграм, триграм і т.д. Проте експериментальна оцінка H_n при достатньо великих *n* нездійсненна з огляду на величезне число можливих значень *n*-грам. Тому розроблені підходи, що дозволяють непрямо оцінити значення H_∞ за допомогою деяких статистичних дослідів.

Один з таких підходів спирається на той факт, що H_∞ може бути також визначена як границя умовних ентропій:

$$H_\infty = \lim_{n\to\infty} H^{(n)}, \text{ де } H^{(n)} = H(x_n \mid x_1, x_2, ..., x_{n-1}).$$

Величина $H^{(n)}$ називається *умовною ентропією джерела*; вона визначає, скільки інформації про наступний символ ми матимемо із значень $(n-1)$ попередніх. Метод оцінки $H^{(n)}$ полягає в тому, що експериментатор за випадково вибраною $(n-1)$-грамою вгадує наступну за нею *n*-ту букву тексту. Нехай $q_1^{(n)}, q_2^{(n)}, ..., q_m^{(n)}$ – імовірності того, що буква буде правильно вгадана з 1-ої, 2-ої, …, *m*-тої спроби (природно, число спроб не може бути більше *m* – числа букв в алфавіті). Тоді має місце нерівність

$$\sum_{i=1}^{m-1} i(q_i^{(n)} - q_{i+1}^{(n)}) \log_2 i + m q_m^{(n)} \log_2 m \le H^{(n)} \le -\sum_{i=1}^{m} q_i^{(n)} \log_2 q_i^{(n)}$$

Ця нерівність дає не вельми точну оцінку H_∞, так як ліва і права його частини не прямують до єдиної границі при $n \to \infty$ і, крім того, через неможливість врахування експериментатором усіх закономірностей мови наведена оцінка буде завищеною. Проте з огляду на простоту реалізації наведений метод є цілком придатним для навчальних потреб.

Надлишковість джерела відкритого тексту (мови) дорівнює $R = 1 - \dfrac{H_\infty}{H_0}$ і характеризує величину можливого ущільнення тексту без втрати його змісту.

3. Шифр Віженера та його криптоаналіз

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 $X = x_0 x_1 x_2 ... x_n$, а ШТ через $Y = y_0 y_1 y_2 ... y_n$. 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

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

Кожен фрагмент Y_i фактично зашифрований шифром Цезаря з ключем k_i, $i = \overline{0, r-1}$. Звідси маємо, що значення імовірностей (або, точніше, частот) символів у цих фрагментах будуть приблизно 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.

1. Для невеликих значень r (приблизно $1 \le r \le 5$) можна скористатись поняттям 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* ($r \ge 2$).
2) Обчислити індекс відповідності даного шифтексту.
3) Порівнюючи обчислене значення із індексами I_r, зробити висновок щодо довжини ключа.

В 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. В другому алгоритмі для маленьких *r* (приблизно $2 \le r \le 5$) значення індексів I_r при різних *r* помітно відрізняються; але для великих розбіжності стають несуттєвими. Звідси бачимо, що застосування даного методу для довгих періодів не ефективне.

2. При великих *r* можна застосувати метод, що 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 співпадінь символів:

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

де $\delta(a,b)$ – символ Кронекера. 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, $i = \overline{1,r}$; знайти цей ключ можна, поклавши $k = (y^* - x^*) \bmod m$, 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»). Якщо ключ вгадано невірно, замість x^* треба брати другу, третю і т.д. за імовірністю літеру.

При розшифруванні деякі фрагменти будуть встановлені неправильно, але можливі помилки легко виправляються при аналізі розшифрованого тексту в цілому.

Порядок виконання роботи

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H_1 та H_2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H_1 та H_2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H_1 та H_2 на тому ж тексті, в якому вилучено всі пробіли.

2. За допомогою програми CoolPinkProgram оцінити значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$.

3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

4. 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.

5. Підрахувати індекси відповідності для відкритого тексту та всіх одержаних шифртекстів і порівняти їх значення.

6. Використовуючи наведені теоретичні відомості, розшифрувати наданий шифртекст (згідно свого номеру варіанта).

Методичні вказівки

Звичайні текстові файли містять багато символів окрім власне літер; для обчислення значень ентропій вони повинні пройти попередню фільтрацію: всі символи, окрім текстових, повинні вилучатись або замінюватись на пробіли; прописні літери – замінюватись на відповідні стрічні; послідовність пробілів (або інших розділових знаків,

наприклад, символів кінця рядку) повинна трактуватись як один пробіл або вилучатись, якщо пробіл не входить до алфавіту.

При підрахунку частот біграм треба розглядати як пари букв, що перетинаються, так і пари букв, що не перетинаються (тобто рухатися вздовж тексту з кроком 2). Одержані результати не повинні суттєво відрізнятись, однак в першому випадку використовується більше статистики, а тому чисельні дані більш точні. Таблицю частот символів потрібно подавати відсортованою за спаданням частот. Таблицю частот біграм зручно подавати у вигляді квадратної матриці, індексованої першою та другою літерами біграм.

Програма CoolPinkProgram використовує текст, що лежить у допоміжному файлі text. Цей текст написаний російською мовою без знаків пунктуації та великих літер; буква «ё» заміщена буквою «е», а «ъ» – буквою «ь». Пробіл також вважається буквою. Таким чином, кількість букв алфавіту $m = 32$. При підрахунку $H^{(10)}$, $H^{(20)}$, $H^{(30)}$ необхідно виконати не менш ніж 50 експериментів.

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

Для оцінки теоретичного значення індексу відповідності користуйтесь значеннями частот символів мови, одержаних під час виконання першої частини практикуму.

При пошуку періоду шифру Віженера потрібно обчислювати значення статистики D_r щонайменше до $r = 30$.

Оформлення звіту

Звіт має містити:

1. Мету лабораторної роботи.
2. Постановку задачі та варіант завдання.
3. Хід роботи, опис труднощів, що виникали, та шляхів їх розв'язання.
4. Тексти всіх відповідних програм.
5. Таблиці частот букв і біграм тексту, одержані значення H_1 та H_2, оцінки для $H^{(10)}$, $H^{(20)}$, $H^{(30)}$ (включно із відповідними скріншотами).
6. Оцінку надлишковості R.
7. Обчислені значення індексів відповідності для вказаних значень r (подати у вигляді таблиці та діаграми).
8. Обчислену послідовність D_r (подати у вигляді діаграми).
9. Шифрований та відповідний розшифрований тексти (відповідно до варіанту завдання), знайдене значення ключа.
10. Висновки.

Контрольні запитання

1) Які два визначення ентропії на символ джерела ви знаєте?

2) Порівняйте одержані значення H_1, H_2, $H^{(10)}$, $H^{(20)}$, $H^{(30)}$. Зробіть висновки.

3) Що таке надлишковість джерела? Яка надлишковість російського письмового тексту згідно ваших даних?

4) Які моделі відкритих текстів розглядаються у криптографії?

5) Яка різниця між блочними та поточними шифрами?

6) Що таке шифри моно- та поліалфавітної підстановки?

7) Що таке шифр Віженера? Опишіть процес зашифрування та розшифрування.

8) Що таке індекс відповідності?

9) Чому не потрібно підраховувати індекс відповідності для шифртексту з $r = 1$? Чому він дорівнює?

10) Яка модель відкритого тексту розглядається при криптоаналізі шифру Віженера?

11) Завдяки чому можливий криптоаналіз шифру Віженера?

12) Що таке частотний аналіз?

Оцінювання практикуму

За виконання комп'ютерного практикуму студент може одержати до 18 рейтингових балів; зокрема, оцінюються такі позиції:

– реалізація програм – до 7 балів (в залежності від правильності та швидкодії);
– теоретичний захист роботи – до 8 балів;
– оформлений звіт – 1 бал;
– своєчасне виконання роботи – 2 бали (згідно графіку складання);
– несвоєчасне виконання роботи – (-2) бали за кожні два тижні пропуску.