Математичні методи захисту інформації: курс лекцій. Частина 1

Завадська Л.О., Савчук М.М.

11 червня 2014 р.

Розділ 1

Задачі, напрямки та методи захисту інформації. Поняття про криптографічний захист інформації

1.1 Області застосування, мета, методи захисту інформації

Протягом свого існування людство пережило декілька інформаційних революцій: створення і розвиток мов, винахід писемності, винахід та широке застосування друкарства, створення комп'ютера та новітніх електронних технологій, що радикально змінили суспільство в усіх галузях, на всіх рівнях суспільного розвитку. Кількість інформації, що була доступна та використовувалась, постійно зростала, а за періоди інформаційних революцій — на декілька порядків. Володіння інформацією і в минулому і нині давало можливість досягти швидкого розвитку і успіху у різних галузях як у глобальному масштабі, так і в конкретних справах. Сьогодні світ переживає період, коли накопичено колосальний об'єм знань, що дозволяє перейти до здійснення справді революційних технологічних рішень. Основою розвитку нині може бути, перш за все, процес пізнання, і він посильний тільки високоосвіченому суспільству, в якому праця приймає все більш інтелектуальні форми. Технологіям майбутнього потрібні широко освічені люди, які здатні орієнтуватися в нових умовах дійсності, що стрімко змінюється.

Однією з галузей, що найбільш динамічно змінюються протягом останніх десятиліть, є технології електронної обробки інформації, телекомунікацій, комп'ютерних мереж, технології захисту інформації.

В Україні широкий попит на методи і засоби захисту інформації почав виявлятися у другій половині 80-х років XX ст. З часом виникла нагальна потреба використання криптографічних та технічних методів захисту також у приватному секторі. Сьогодні велика кількість конфіденційної інформації передається в електронному вигляді, на електронних носіях, між

ЕОМ звичайними лініями зв'язку. Інформація може продаватися та купуватися, мати ціну, що незрівнянно перевищує ціну матеріального носія. Часто володіння інформацією дає переваги, ціну яких неможливо підрахувати, наприклад, у військовій справі. Термін збереження секретності інформації може коливатися від декількох годин до багатьох десятиліть. Тому вкрай потрібні спеціалісти, які володіють криптографічними, технічними, комплексними методами захисту, знають відповідні стандарти, здатні використовувати (або розробляти) програмне й апаратне забезпечення для гарантування таємності та цілісності конфіденційної інформації. Криптографічні методи захисту вважаються одними з найбільш надійних та ефективних.

- Області застосування захисту інформації (відповідно — види таємниці):

  1. Військова;
  2. Дипломатична;
  3. Фінансова;
  4. Банківська;
  5. Комерційна;
  6. Промислова;
  7. Наукова;
  8. Юридична;
  9. Медична;
  10. Особиста таємниця.

- Мета і головні задачі захисту інформації:

  1. Конфіденційність (секретність) інформації;
  2. Цілісність інформації;
  3. Автентичність інформації;
  4. Доступність інформації.

- Напрямки, аспекти, методи і засоби захисту інформації:

  1. Юридичні, правові;
  2. Методично-нормативні;
  3. Організаційні;
  4. Безпосередні (фізичні);
  5. Технічні — захист від витоку по технічним каналам:
     (a) електромагнітному
     (b) оптичному
     (c) акустичному
     (d) віброакустичному;
  6. Стеганографічні;
  7. Криптографічні. Методи математичного захисту інформації;
  8. Методи квантової криптографії;
  9. Морально-етичні норми.

1.2 Перші поняття криптографічного захисту інформації

Визначення 1.1 (Криптографічний захист інформації)**.** Криптографічний захист інформації — це різновид захисту інформації, який реалізується за допомогою криптографічних перетворень, спеціальних ключових даних з метою приховування та відновлення змісту інформації, підтвердження достовірності, авторства, запобігання несанкціонованому використанню тощо.

Визначення 1.2 (Криптографічне перетворення)**.** Криптографічне перетворення — це перетворення інформації відповідно до певних правил (логічних, математичних) з метою забезпечення функціонування криптографічних протоколів.

Визначення 1.3 (Криптографічний ключ)**.** Криптографічний ключ — це параметр, який використовується в криптографічному алгоритмі для вибору конкретного криптографічного перетворення; ключі можуть бути таємними або відкритими.

Визначення 1.4 (Криптографічний протокол)**.** Криптографічний протокол — це послідовність узгоджених дій згідно з деякими правилами, у відповідності з якими відбувається обмін інформацією між сторонами або учасниками протоколу та її перетворення з використанням криптографічних методів і засобів. Простий приклад криптографічного протоколу – це зашифрування та розшифрування повідомлення.

Визначення 1.5 (Криптографія)**.** Криптографія — науково-технічна дисципліна, яка вивчає принципи, методи і засоби криптографічного захисту інформації і інформаційних технологій, предметом якої є розробка криптографічних систем.

Визначення 1.6 (Криптоаналіз)**.** Криптоаналіз — це науково-технічна дисципліна, яка вивчає методи, способи і засоби аналізу криптографічного захисту інформації: криптографічних систем, криптографічних алгоритмів, протоколів з метою знайти способи їх розкриття без знання секретних ключів і, можливо, будови криптосистем, знайти способи несанкціонованого доступу, підробки даних тощо. Криптоаналіз оцінює складність таких способів розкриття (злому) і стійкість криптографічного захисту інформації. Фахівця, який займається криптоаналізом будемо називати криптоаналітиком.

Визначення 1.7 (Криптологія)**.** Криптологія за найбільш поширеною сучасною термінологією, об'єднує в собі дисципліни криптографію і криптоаналіз.

Зауваження 1. *Не всі країни дотримуються останньої термінології щодо дисциплін. Так, наприклад, в Росії назва „криптографія" об'єднує в собі власне криптографію (криптосинтез) у вище наведеному розумінні і криптоаналіз, а криптологія розглядається як галузь криптографії, що вивчає математичні моделі криптографічних систем, і також поділяється на криптосинтез та криптоаналіз.*

1.3 Етапи розвитку технологічних засобів криптографії

1. “Ручна” криптографія (із давнини до середини-кінця XIX століття)

   Основні види шифрів — заміни і перестановки.

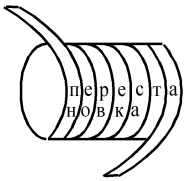

Рис. 1.1: Шифр Скитала

Перший шифр перестановки, застосування якого зафіксоване у військовій справі, (Спарта, V ст. до н.е.) – шифр Скитала (рис. 1.1). Таємний ключ – діаметр барабана.

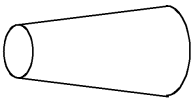

Рис. 1.2: Засіб криптоаналізу шифру Скитала

Для шифрування на стрічці, що намотувалась на барабан (скитал), писалось вздовж барабана повідомлення. Після знімання з барабану на стрічці була зовні випадкова послідовність літер – шифроване повідомлення. Криптоаналіз шифру Скитала запропонував Арістотель за допомогою барабана змінного діаметру (рис. 1.2): якщо на намотаній на нього стрічці з шифрованим повідомленням у деякому місці вгадувались якісь частини слів, то цьому місцю відповідав діаметр справжнього барабану.

Прикладом шифру заміни є шифр Цезаря — заміна кожної букви повідомлення на букву циклічно віддалену в алфавіті на фіксоване число позицій.

2. Застосування телеграфу для шифрування і кодування (з середини XIX ст.)
3. Використання механічних машин (кінець XIX ст. – 20і роки XX ст.)
4. Електромеханічні машини (з 20-х років XX ст. – середина XX ст.)

   Приклад — ENIGMA — основна шифрувальна машина Вермахту у Другій світовій війні.
5. Електронні машини (з кінця 40-х років XX ст.)
6. Напівпровідникові криптосистеми
7. Криптосистеми, засновані на мікросхемах

8. Використання комп'ютерної техніки для криптографічного захисту

9. Квантова криптографія

1.4 Про розвиток теоретичної криптографії

До епохи Відродження криптографією займалися, можна сказати, не професіонали. Найчастіше шифри уявляли собою деякі головоломки. Але слід зазначити, що іноді винаходилися шифри, які не були розкриті на протязі 100 і навіть більше років! Хоча за сучасними поняттями, враховуючи залучення до криптоаналізу ЕОМ, загалом це були слабкі, нестійкі шифри.

Велике просування в криптографії відбулося після залучення до вирішення її проблем відомих математиків періоду відродження, серед яких були Ф. Вієт, Д. Кардано та ін. Пізніше почали створюватись спеціалізовані державні служби шифрування і дешифрування (розкриття). Такі служби створили, наприклад, Кромвель у Англії, кардинал Рішельє у Франції, Петро I в Росії.

У 1883 році була опублікована книга Керкгоффа «Військова криптографія», в якій були вперше сформульовані деякі вимоги до криптосистем, правила щодо утворення, експлуатації, стійкості криптографічних пристроїв. Частина цих правил і зараз вважаються обов'язковими.

Наукою у повному розумінні цього слова теоретичну криптографію стали визнавати з 1949 року після публікації у відкритому друці статті К. Шеннона "Теорія зв'язку в секретних системах". А з 1976 року завдяки ідеям статті Діффі і Хеллмана "Нові напрямки в криптографії" почався новий етап у розвитку криптографії – застосування криптосистем з відкритим ключем, яке дало можливість успішно вирішити низку назрілих проблем криптографічного захисту інформації.

1.5 Класифікація сучасних криптосистем

1. *Симетричні (одноключові, з секретним ключем).* Підрозділяються на блокові й потокові. У відправника та одержувача повідомлення один і той самий секретний ключ, вони знаходяться у рівних (симетричних) умовах, можуть як зашифрувати повідомлення так і розшифрувати за допомогою таємного ключа.

2. *Асиметричні (двохключові, з відкритим ключем, із загальнодоступними ключами).* Відомі з 1976 року, активно використовуються на практиці з 1978 року. У найпростішому випадку мають два ключі: один (відкритий) — у відправника для шифрування, інший — у одержувача (секретний) для розшифрування.

3. *Квантові* (знаходяться у стадії експерименту та розвитку)

1.6 Контрольні питання

1. Які мета та задачі захисту інформації?

2. Назвіть напрямки та методи захисту інформації.

3. Що таке криптографічний ключ?

4. З яких частин складається криптологія?

5. Чи були відомі способи захисту інформації шифруванням до нашої ери?

6. Коли використовувались роторні шифрувальні машини?

7. Яка дата вважається початком розвитку криптографії з відкритим ключем?

8. Чим відрізняються симетричні криптосистеми від асиметричних?

Розділ 2

Моделі джерел відкритого тексту. Ентропія на символ джерела

При шифруванні текст перетворюється таким чином, щоб зробити його зміст незрозумілим для того, хто не знає секретного ключа. Для побудови математичної теорії криптографічних систем шифрування потрібно насамперед дати математичний опис (або математичну модель) тексту та перетворень, які відбуваються з ним під час шифрування.

Визначення 2.1 (Алфавіт)**.** Надалі вважатимемо, що алфавіт є скінченним. Позначимо алфавіт як $Z_m = \{z_1, \ldots, z_m\}$, де $z_1, \ldots, z_m$ — букви (символи) алфавіту. Елементами алфавіту можуть бути власне букви; букви та цифри; букви, цифри та знаки пунктуації, взагалі скінченний набір будь-яких символів, наприклад, танцюючі чоловічки. Як правило, ми будемо розглядати український, російський чи латинський алфавіт (малі букви) зі знаком пропуску, що вважається буквою, або без нього, або ж двійковий алфавіт, що складається з двох символів: 0 та 1.

Визначення 2.2 (Текст)**.** Під текстом будемо розуміти послідовність букв деякого алфавіту.

Визначення 2.3 (Відкритий текст)**.** Відкритий текст (ВТ) — це текст, що підлягає шифруванню.

Визначення 2.4 (Шифрований текст)**.** Шифрований текст (ШТ) — це текст, що утворюється в результаті шифрування.

Відкритий та шифрований тексти можуть бути записані як у одному й тому ж, так і у різних алфавітах (більш докладно поняття ВТ та ШТ розглядаються у лекції 3).

Визначення 2.5 (n-грама)**.** n-грамою називається послідовність n символів тексту, що стоять підряд. При n=2 це біграма, при n=3 — триграма.

Будь-який текст має певну статистичну структуру. Для опису цієї структури використовуються різноманітні ймовірнісні моделі мови.

Визначення 2.6 (Джерело відкритого тексту)**.** Джерело відкритого тексту генерує послідовність символів алфавіту $x_1, x_2, \ldots, x_n, \ldots$ випадковим чином. Джерело визначається алфавітом та ймовірностями появи n-грам: $\mathbb{P}(x_{i+1} = z_1, x_{i+2} = z_2, \ldots, x_{i+n} = z_n)$ для будь-яких цілих $n \geq 1$, $i \geq 0$ (тут $x_1, x_2, \ldots, x_n$ — випадкові величини, а $z_1, \ldots, z_n$ — букви алфавіту), які мають задовольняти умовам:

1. Вихід джерела ВТ є випадковим процесом з дискретним часом та множиною станів Z_m

$$\sum_{z_1, z_2, \ldots, z_n \in Z_m} \mathbb{P}(x_{i+1} = z_1, \ldots, x_{i+n} = z_n) = 1$$

2. Умова узгодженості скінченновимірних розподілів виходу джерела ВТ: Для будь-якого цілого $s \geq 1$

$$\sum_{z_1, z_2, \ldots, z_n \in Z_m} \mathbb{P}(x_{i+1} = z_1, \ldots, x_{i+n} = z_n, \ldots, x_{i+n+s} = z_{n+s}) = \\ = \mathbb{P}(x_{i+1} = z_1, \ldots, x_{i+n} = z_n)$$

Визначення 2.7 (Стаціонарне джерело відкритого тексту)**.** Джерело називають стаціонарним, якщо для будь-яких цілих $n \geq 1, 1 \leq i_1 < \cdots < i_n, j \geq 0$ і будь-якого набору букв алфавіту $z_1, \ldots, z_n$ виконується рівність:

$$\mathbb{P}(x_{i_1+j} = z_1, x_{i_2+j} = z_2, \ldots, x_{i_n+j} = z_n) = \mathbb{P}(x_{i_1} = z_1, x_{i_2} = z_2, \ldots, x_{i_n} = z_n)$$

У подальшому будемо розглядати лише стаціонарні джерела, тобто такі, у яких немає залежності від зсуву j. Для стаціонарних джерел достатньо задати ймовірності $\mathbb{P}(x_1 = z_1, \ldots, x_n = z_n)$ для $n \geq 1$.

В залежності від властивостей сумісних розподілів $\mathbb{P}(x_1 = z_1, \ldots, x_n = z_n)$, $n \geq 1$ можна побудувати різні моделі джерела ВТ.

Найбільш уживаними є описані нижче чотири моделі, з яких кожна наступна все більш адекватно відображає структуру мови. Перша з них є простою й менш за всі враховує реальні статистичні властивості мови. Назвемо її моделлю М0.

Визначення 2.8 (Модель М0)**.** У цій моделі джерело у кожен момент часу генерує символи із Z_m незалежно та рівноімовірно:

$$\mathbb{P}(x_i = z) = \frac{1}{m}, z \in Z_m, i = 1, 2, \ldots$$

Всі n-грами в моделі М0 є рівноймовірними для будь-яких цілих $n \geq 1$ та $z_1, \ldots, z_n \in Z_m$

$$\mathbb{P}(x_1 = z_1, \ldots, x_n) = z_n = \frac{1}{m^n}$$

Модель М0 має допоміжний характер, бо вона не враховує навіть найпростіших властивостей мови.

Наступна модель ВТ враховує частоти, з якими окремі букви зустрічаються у мові.

Визначення 2.9 (Модель М1)**.** Символи тексту $x_1, x_2, \ldots, x_n, \ldots$ є незалежними, але вони генеруються із різними ймовірностями

$$\mathbb{P}(x_i = z) = \mathrm{p}(z), z \in Z_m, i = 1, 2, \ldots$$

Розподіл ймовірностей $\mathrm{p}(z)$ відповідає частотам появи букв $z \in Z_m$ у мові. У цій моделі ймовірність появи n-грами має вигляд:

$$\mathbb{P}(x_1 = z_1, x_2 = z_2, \ldots, x_n = z_n) = \prod_{i=1}^{n} \mathrm{p}(z_i)$$

Зазначимо, що ймовірності букв у природних мовах значно відрізняються. Наприклад, найбільшу частоту в українській та російській мові має буква "о", в англійській — буква "е". В російській мові "о" зустрічається майже в 50 раз частіше букви "ф", що має найменшу частоту.*

Наступна модель враховує залежність в мові між двома буквами, що стоять поряд.

Визначення 2.10 (Модель М2)**.** Джерело генерує біграми $x_1x_2, x_3x_4, x_5x_6, \ldots$ і т.д. незалежно одну від одної. Тобто на множині всіх біграм заданий розподіл імовірностей $\mathrm{p}(z_i, z_j), i, j = 1, \ldots, m$, і кожна нова біграма джерела генерується незалежно від інших.

Більш складні залежності мови враховуються за допомогою марковської моделі.

Визначення 2.11 (Однорідний ланцюг Маркова)**.** Однорідний ланцюг Маркова — це послідовність випадкових величин $\{x_i\}, i = 1, 2, \ldots$, що приймають значення у дискретній множині Z, така, що для будь-якого $n \geq 2$

$$\begin{aligned}\mathbb{P}(x_n = z_n \mid x_1 = z_1, x_2 = z_2, \ldots, x_{n-1} = z_{n-1}) = \\ = \mathbb{P}(x_n = z_n \mid x_{n-1} = z_{n-1}) = p_{z_n z_{n-1}}\end{aligned}$$

Для неоднорідного марковського ланцюга імовірності переходу $p_{z_n z_{n-1}}$ залежали б від n, а в нашому випадку вони залежать лише від станів z_{n-1}, z_n.

Визначення 2.12 (Модель М3)**.** У цій моделі джерела ВТ послідовність $x_1, x_2, \ldots$ утворює однорідний ланцюг Маркова. Для завдання такого марковського ланцюга достатньо задати розподіл початкових станів $p_0(i), i \in Z_m$ та перехідні ймовірності $p_{ij} = \mathbb{P}(x_{n+1} = j \mid x_n = i), i, j \in Z_m$, які в силу однорідності не залежать від n.

При виконанні деяких умов на ланцюг Маркова (які не суперечать властивостям природних мов) існує граничний розподіл

$$\pi_i = \lim_{n \to \infty} \mathbb{P}(x_n = i \mid x_1 = j)$$

що не залежить від початкового стану j. Він називається стаціонарним розподілом ймовірностей марковського ланцюга. Ці ймовірності задовольняють наступним рівностям

*За умови, що твердий знак ототожнюється з м'яким.

$$\begin{cases} \sum\limits_{i \in Z} \pi_i = 1 \\ \sum\limits_{i \in Z} \pi_i p_{ij} = \pi_j, j \in Z_m \end{cases}$$

Ймовірність n-грами у моделі М3 у стаціонарному режимі можна записати у вигляді

$$\mathbb{P}\left(x_1 = z_1, \ldots, x_n = z_n\right) = \pi_{z_1} p_{z_1 z_2} p_{z_2 z_3} \ldots p_{z_{n-1} z_n}$$

Якщо існує стаціонарний розподіл і $p_0\left(i\right) = \pi_i, i \in Z_m$, то одноріний ланцюг Маркова — стаціонарний випадковий процес.

2.1 Деякі відомості з теорії інформації

Нехай $X = \{x_1, x_2, \ldots, x_m\}$ — скінченна множина, на якій заданий розподіл ймовірностей $P = \{\mathrm{p}\left(x_1\right), px_2, \ldots, px_m\}$.

Визначення 2.13 (Повідомлення)**.** Елементи x_i будемо називати повідомленнями

Визначення 2.14 (Скінченний ансамбль)**.** Пара (X, P) в теорії інформації називається ансамблем (скінченним).

Часто говорять просто про ансамбль X, розуміючи під цим пару (X, P).

Інтуїтивно зрозуміло, що малоймовірне повідомлення несе у собі більше інформації, ніж більш імовірне. К. Шеннон запропонував для виміру кількості інформації функцію, яка відповідає цьому інтуїтивному уявленню, і до того ж є зручною при обчисленнях.

Визначення 2.15 (Власна інформацієя повідомлення)**.** Власною інформацією повідомлення x_i називається величина

$$I\left(x_i\right) = -\log p\left(x_i\right) \geq 0$$

Усереднимо власну інформацію за всіма повідомленнями ансамблю

$$H(X) = -\sum_{i=1}^{m} \mathrm{p}\left(x_i\right) \cdot \log \mathrm{p}\left(x_i\right)$$

Визначення 2.16 (Ентропія ансамблю)**.** Величина $H\left(X\right)$ називається ентропією ансамблю X.

$H\left(X\right)$ може бути інтерпретована як невизначеність експерименту, в якому з ансамблю X вибирається одне повідомлення, причому повідомлення x_i може бути вибране з ймовірністю $\mathrm{p}\left(x_i\right)$, $i = 1, 2, \ldots$

Найчастіше в означенні ентропії беруть логарифм за основою 2, тоді інформація та ентропія вимірюються в бітах. Якщо логарифм десятковий (lg), то — в дитах, якщо логарифм натуральний (ln), то — в натах.

Розглянемо декартовий добуток скінченних множин X та Y, тобто множину пар $\left(x, y\right), x \in X, y \in Y$. Нехай на множині пар задано розподіл імовірностей $\mathrm{p}\left(x, y\right)$. Тоді говорять, що ансамблі X та Y задані сукупно. Розподіл $\mathrm{p}\left(x, y\right)$ індукує розподіли на X та Y:

$$\begin{cases} \mathrm{p}(x) & = \sum\limits_{y_j \in Y} \mathrm{p}(x,y) \\ \mathrm{p}(y) & = \sum\limits_{x_j \in X} \mathrm{p}(x,y) \end{cases}$$

тобто X та Y можна також розглядати і як окремі ансамблі.

Визначення 2.17 (Сукупна ентропія)**.** Сукупною ентропією ансамблів X та Y називається величина

$$H(XY) = -\sum_{x,y} \mathrm{p}(x,y) \cdot \log \mathrm{p}(x,y)$$

Визначення 2.18 (Незалежні ансамблі)**.** Сукупно задані ансамблі X та Y називаються незалежними, якщо

$$\forall (x,y) : \mathrm{p}(x,y) = \mathrm{p}(x) \cdot \mathrm{p}(y)$$

Розглянемо декартовий добуток скінченних множин X та Y, тобто множину пар

$$(x,y)\,,\ x \in X, y \in Y$$

Визначення 2.19 (Сукупно задані ансамблі)**.** Нехай на множині пар задано розподіл імовірностей $\mathrm{p}(x,y)$. Тоді говорять, що ансамблі X та Y задані сукупно.

Розподіл $\mathrm{p}(x,y)$ індукує розподіли на X та Y:

$$\mathrm{p}(x) = \sum_{y_j \in Y} \mathrm{p}(x,y)\,; \qquad \mathrm{p}(y) = \sum_{x_j \in X} \mathrm{p}(x,y)$$

тобто X та Y можна також розглядати і як окремі ансамблі.

Визначення 2.20 (Сукупна ентропія)**.** Сукупною ентропією ансамблів X та Y називається величина

$$H(XY) = -\sum_{x,y} \mathrm{p}(x,y) \cdot \log \mathrm{p}(x,y)$$

Визначення 2.21 (Незалежні ансамблі)**.** Сукупно задані ансамблі X та Y називаються незалежними, якщо

$$\forall (x,y) :\ \mathrm{p}(x,y) = \mathrm{p}(x) \cdot \mathrm{p}(y)$$

З курсу теорії ймовірностей відомо визначення умовної ймовірності:

$$\mathrm{p}(x \mid y) = \frac{\mathrm{p}(x,y)}{\mathrm{p}(y)},\ \mathrm{p}(y) \neq 0$$

Нехай відомий результат y експерименту Y. Тоді можна визначити умовну ентропію таким чином:

$$H(X|y) = -\sum_{x \in X} \mathrm{p}(x|y) \cdot \log \mathrm{p}(x|y)$$

Усереднивши $H(X \mid y)$ за всіма y, отримаємо умовну ентропію ансамблю X відносно ансамблю Y:

$$H(X \mid Y) = -\sum_{y \in Y} \mathrm{p}(y) \cdot \sum_{x \in X} \mathrm{p}(x \mid y) \cdot \log \mathrm{p}(x \mid y) = \\ = -\sum_{\substack{x \in X \\ y \in Y}} \mathrm{p}(x, y) \cdot \log \mathrm{p}(x \mid y)$$

Властивості ентропії

1. $H(X) \geq 0$.
2. $H(X) = 0$ тоді і тільки тоді, коли деяке $p_i = 1$, а всі інші $p_j = 0, j \neq i$. ($\log 0$ не визначений, та оскільки $p \cdot \log p \xrightarrow[p \to 0]{} 0$, то за неперервністю довизначимо такі доданки в $H(X)$ як нульові).
3. $H(X)$ приймає максимальне значення тоді, коли всі x_i є рівноймовірними: $p(x_i) = \dfrac{1}{m}$. В цьому випадку
$$H(X) = -\sum_{i=1}^{m} \frac{1}{m} \cdot \log \frac{1}{m} = \log m$$
4. Якщо X та Y — незалежні, то
$$H(XY) = H(X) + H(Y)$$
5. $$H(X \mid Y) \geq 0$$
6. Інтуїтивно: сукупна невизначеність експериментів X та Y дорівнює невизначеності X плюс невизначеність Y, що залишилася після того, як результат експерименту X став відомим. X та Y входять у формулу симетрично
$$H(XY) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$
7. Додаткові знання не можуть збільшити невизначеність
$$H(X) \geq H(X \mid Y) \geq H(X \mid YZ) \geq \dots$$

Визначення 2.22 (Взаємна інформація)**.** Взаємною інформацією ансамблів X та Y називається величина

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

При незалежних ансамблях X та Y

$$I(X;Y) = 0$$

Якщо множини X та Y співпадають, то декартовий добуток $X \times Y$ позначимо як X^2. Аналогічно, якщо всі $X_i = X$, то

$$X_1 \times X_2 \times \dots \times X_n = X^n$$

2.2 Ентропія на символ джерела

Якщо Z_m — алфавіт, то n-грама $(z_1, \ldots, z_n) \in Z_m^n$ і n ансамблів задані сукупно розподілом n-грам $\mathbb{P}\{x_1 = z_1, \ldots, x_n = z_n\}$ (джерело стаціонарне: немає залежності від розташування n-грами в тексті). Ентропія n-грами

$$H(Z_m^n) = -\sum_{\substack{z_i \in Z_m \\ i=\overline{1,n}}} \mathrm{p}(z_1, \ldots, z_n) \cdot \log \mathrm{p}(z_1, \ldots, z_n) \tag{2.1}$$

Усереднена ентропія на один символ n-грами дорівнює

$$H_n = \frac{H(Z_m^n)}{n}$$

Для стаціонарних джерел існує границя цієї величини (у теорії інформації доводиться, що послідовність H_n є монотонно незростаючою):

$$\lim_{n\to\infty} H_n = \lim_{n\to\infty} \frac{H(Z_m^n)}{n} = H_\infty$$

Визначення 2.23 (Ентропія на символ джерела)**.** Границя H_∞ називається ентропією на символ джерела.

Розглянемо, чому буде дорівнювати ентропія на символ джерела для різних моделей ВТ, що введені раніше.

M0: $H(Z_m^n) = n \cdot H(Z_m) = n \cdot \log m$ (третя та четверта властивості ентропії)

$$\frac{H(Z_m^n)}{n} = \log m = H_\infty$$

M1: Внаслідок незалежності букв у тексті

$$H_\infty = \lim_{n\to\infty} \frac{H(Z_m^n)}{n} = \lim_{n\to\infty} \frac{n \cdot H(Z_m)}{n} = \\ = H(Z_m) = -\sum_{z\in Z} \mathrm{p}(z) \cdot \log \mathrm{p}(z)$$

M2: Будемо розглядати тексти довжиною $2 \cdot n$:

$$H_\infty = lim_{n\to\infty} \frac{H(Z_m^{2\cdot n})}{2 \cdot n} = \frac{n \cdot H(Z_m^2)}{2 \cdot n} = \frac{H(Z_m^2)}{2} = H_2$$

Результат саме такий, оскільки біграми незалежні.

M3: Можна довести, що для джерела, що описується однорідним ланцюгом Маркова, який має стаціонарний розподіл $\{\pi_i\}$ та ймовірності переходу $p_{ij},\ i, j \in Z_m$

$$H_\infty = -\sum_{i,j\in Z_m} \pi_i \cdot p_{ij} \cdot \log p_{ij}$$

Величина H_n є n-м наближенням до H_∞. Зазначимо, що перші наближення H_1, H_2, H_3 ще дуже відрізняються від H_∞ (див. табл. 2.1). В той же час обчислити H_n при великих значеннях n практично неможливо через величезну кількість можливих n-грам. З іншого боку, можна розглянути умовну ентропію n-го символу тексту при умові, що відомі $n-1$ попередніх: $H^{(n)} = H(x_n \mid x_1, \ldots, x_{n-1})$. В теорії інформації доводиться, що послідовність $H^{(n)}$ має границю при $n \to \infty$ і ця границя співпадає з границею послідовності H_n. Отже

$$H_\infty = \lim_{n\to\infty} \cdot H(x_n \mid x_1, \ldots, x_{n-1})$$

Це є друге визначення ентропії на символ джерела. Воно використовується для експериментальної оцінки H_∞ шляхом вгадування людиною наступної букви тексту.

	H_0	H_1	H_2	H_3	H_5	H_8
Англійська мова	4.76	4.03	3.32	3.10	2.1	1.9
Російська мова	5	4.35	3.52	3.01	—	—

Табл. 2.1: Експериментальні оцінки H_n при деяких значеннях n

Використовуючи друге визначення H_∞, А. Н. Колмогоров експериментально оцінив для російської мови $H^{(n)}$ при великих значеннях n. Виявилося, що після $H^{(30)}$ значення $H^{(n)}$ вже практично не змінюються, тобто дорівнюють H_∞, в той час як H_{15} ще істотно відрізняється від H_∞. Аналогічні результати було отримано і для інших європейських мов (роботи К. Шеннона та ін.).

2.3 Надлишковість на символ джерела

Визначення 2.24 (Надлишковість на символ джерела)**.** Надлишковістю на символ джерела (з алфавітом довжиною m) називається величина

$$R = 1 - \frac{H_\infty}{H_0},\ H_0 = \log_2 m$$

H_0 дорівнює максимальній кількості інформації, яку може нести в собі один символ джерела, а H_∞ — кількості інформації, яку насправді несе в собі один символ. Таким чином, для достатньо великих n величина $R \cdot n$ є середньою кількістю "зайвих" символів у тексті довжиною n, після втрати яких теоретично можна відновити текст. Надлишковість європейських мов знаходиться десь на рівні 60-80%. Але це не означає, що після випадкового видалення 60% символів (букв) завжди залишиться можливість відновити текст. Відкидати букви треба вибірково, використовуючи всі закономірності мови, і відновлювати також, використовуючи всі ці закономірності. Але практично врахувати всі закономірності неможливо. Експериментально встановлено: тільки до 25% букв можна видалити випадковим чином, щоб при цьому текст залишився придатним для відновлення.

Якби букви в тексті були незалежними та рівномірно розподіленими, то H_∞ дорівнювало б H_0, а надлишковість R була б нульовою. Але тоді

будь-яка послідовність букв була б змістовним текстом, і навіть найменша помилка при передачі призводила б до іншого змістовного тексту й не могла б бути визначена. При усному мовленні ми “ковтаємо” частину звуків, або виголошуємо їх нечітко, на письмі інколи робимо орфографічні помилки, та завдяки надлишковості мови все одно розуміємо один одного. Тож надлишковість — це природний механізм, що сприяє розумінню та протидіє помилкам.

2.4 Контрольні питання

1. Що таке джерело відкритого тексту?
2. Чому розглядаються стаціонарні джерела і що це означає?
3. Які моделі відкритого тексту ви знаєте?
4. Дайте визначення ентропії ансамблю.
5. Назвіть найважливіші властивості ентропії.
6. Які визначення ентропії на символ джерела ви знаєте?
7. Скільки доданків у правій частині формули (2.1)?
8. Що таке надлишковість на символ джерела? Чому приблизно дорівнює надлишковість європейських мов?

Зміст