

Лекція 2

Інформація та ентропія: кількісна міра

Вступ

до класичної теорії
інформації

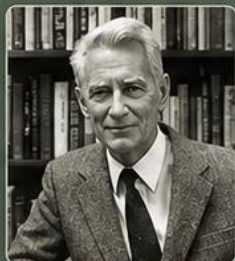
1. Кількісна міра
інформації

2. Ентропія та її
властивості

3. Надмірність інформації

$$H = - \sum_{i=1}^N p_i \log_2 p_i$$

$$I = \log_2 N$$



Клод Елвуд Шеннон



Ральф Вінтон
Лайон Хартлі



Провідні вчені та засновники



Джон фон Нейман
(1903–1957)



Клод Елвуд Шеннон
(1916–2001),
інженер, математик,
основоположник теорії
інформації



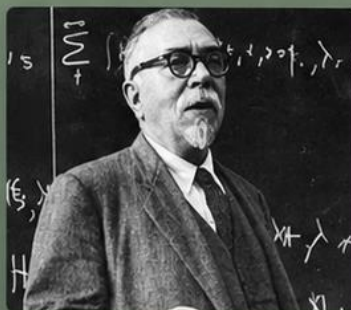
Ральф Вінтон
Лайон Хартлі
(1888–1970, США)



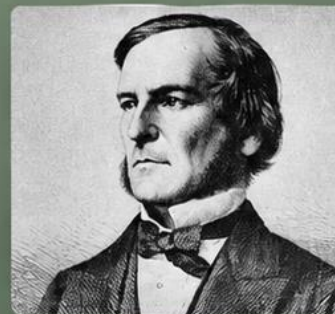
Гаррі Найквіст
(1889–1976),
швед



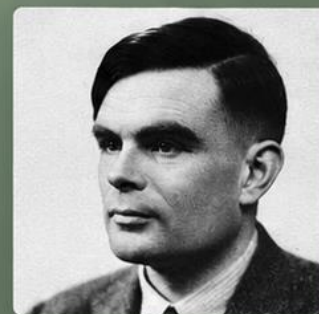
У. Р. Ешбі
(1903–1972),
Англія,
метод «Чорного
ящика»



Норберт Вінер
(1894–1964),
основоположник
кібернетики



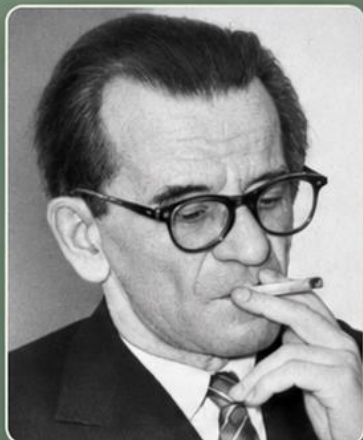
Джордж Буль
(1815–1864, Ірландія) —
англійський математик
і логік.



Тюрінг Алан
Матісон
(1912–1954),
математик



Українські піонери обчислювальної техніки



Сергій Олексійович Лебедєв (1902 – 1974)



академік АН УРСР

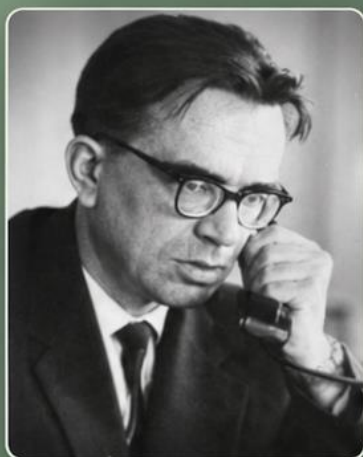


творець першого в континентальній Європі електронного комп'ютера



Керував з 1947 року розробкою у Києві (у передмісті Феофанія) першої в СРСР лампової обчислювальної машини МЕОМ (малої електронної обчислювальної машини).

Перший електронний комп'ютер у континентальній Європі



Віктор Михайлович Глушков (1923 – 1982)



вітчизняний вчений, академік



один із основоположників вітчизняної обчислювальної техніки та інформатики



У 1996 р. IEEE Computer Society присвоєно академіку В. М. Глушкову звання «Computer Pioneer» за його внесок у розвиток автоматики та обчислювальної техніки.

Внесок у світове визнання української школи інформатики

Вступ до класичної теорії інформації: задачі, постулати, базові поняття, модель зв'язку



ТЕОРІЯ ІНФОРМАЦІЇ як наука

ТЕОРІЯ ІНФОРМАЦІЇ, як наука, вивчає питання оцінки кількості інформації, аналізу інформаційних характеристик джерел повідомлень і каналів зв'язку, можливості кодування та декодування повідомлень, щоб забезпечити максимальну швидкість їх передачі каналами зв'язку при наявності та відсутності завад.



Основні задачі:

1. оцінка кількості інформації;
2. математичний опис сигналів інформації;
3. перетворення сигналів інформації в процесах модуляції, демодуляції, фільтрації, аналого-цифрового та цифро-аналогового перетворень;
4. кодування та декодування сигналів.



Важливо знати

Задачі теорії інформації вирішуються методами математичного аналізу, теорії ймовірностей і випадкових функцій, неперервного та дискретного перетворень Лапласа, Фур'є та інших розділів математики.

Постулати теорії інформації

1 Вибір повідомлення

Джерело здійснює вибір повідомлення з деякої множини з певною ймовірністю.

2 Кодування та декодування

Повідомлення можуть передаватися каналом зв'язку в закодованому вигляді. Множина кодованих повідомлень є взаємно однозначним відображенням множини повідомлень. Правило декодування відоме декодеру та записане в його програмі.

3 Критерій правильного приймання

Повідомлення прийняте правильно, якщо після декодування його можна точно відновити. При цьому не враховуються час від передавання до завершення декодування та складність операцій кодування й декодування.

Лекція 2

Основні поняття

ІНФОРМАЦІЯ — продукт взаємодії даних і адекватних їм методів

ДАНІ — отримані відомості, подані у формалізованому вигляді (літерами, цифрами, символами тощо)

ПОВІДОМЛЕННЯ — дані, що підлягають передачі

СИГНАЛ — матеріальний носій повідомлення

Дані — сировина для створення інформації: перетворення й обробка даних дозволяє (з точки зору теорії інформації) витягнути новизну та перетворити її на знання.

СИГНАЛИ



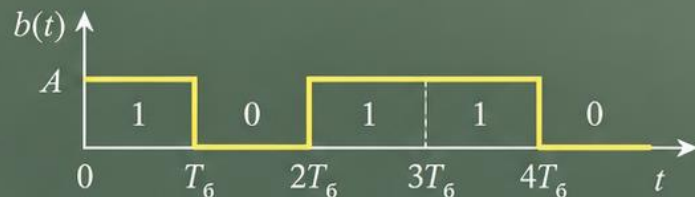
ПОВІДОМЛЕННЯ



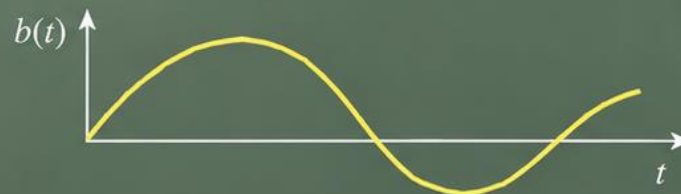
ДАНІ



ІНФОРМАЦІЯ



а) цифровий сигнал



б) неперервний сигнал

Рисунок 1 – Первинні сигнали

Лекція 2

Інформація та ентропія: кількісна міра

Інформація існує у вигляді:



документів, креслень,
рисуноків, текстів



світлових або звукових
сигналів



радіохвиль



електричних та нервових
імпульсів



магнітних записів



жестів та міміки



запахів та смакових відчуттів



хромосом, за допомогою яких
передаються по спадку ознаки
і властивості організмів

Найбільш важливими властивостями інформації є:



об'єктивність та
суб'єктивність



повнота



достовірність



адекватність



доступність



цінність



своєчасність



актуальність

Що можна робити з інформацією

01 Створювати

02 Передавати

03 Сприймати

04 Використовувати

05 Запам'ятовувати

06 Приймати

07 Копіювати

08 Формалізувати

09 Розповсюджувати

10 Перетворювати

11 Комбінувати

12 Обробляти

13 Ділити на частини

14 Спрощувати

15 Збирати

16 Зберігати

17 Шукати

18 Вимірювати

19 Знищувати

Модель зв'язку та передавання повідомлень

СИСТЕМА ЕЛЕКТРИЧНОГО ЗВ'ЯЗКУ



ЗАГАЛЬНА МОДЕЛЬ ЗВ'ЯЗКУ ЗА К. ШЕННОНОМ



Таким чином, теорія інформації, як наука, вивчає:

- 1 Кількісну міру інформації.
- 2 Інформаційні характеристики джерел повідомлень і каналів зв'язку.
- 3 Кодування в дискретних та неперервних каналах.
- 4 Коди, що виявляють та виправляють помилки.
- 5 Ефективність кодування та передачі інформації.

Лекція 2

1. Кількісна міра інформації

1.

Статистичний підхід



У реальних системах для кількісної оцінки інформації використовують **статистичний підхід**, у рамках якого інформація розглядається як сукупність повідомлень про поведінку деякої системи.

2.

Стан системи та невизначеність



Система може випадково знаходитись в одному з кількох можливих **станів**. Вона має певну ступінь **невизначеності (ентропію)**, і її фактичний стан до отримання повідомлення залишається невідомим.

3.

Повідомлення як інформація



Повідомлення про фактичний стан системи, яку ми розглядаємо, є **інформацією** про неї. Якщо стан системи визначений і не може змінюватись, або всі можливі зміни станів відомі до появи повідомлення, то таке повідомлення не несе нічого нового і не дає інформації.

4.

Чим більша невизначеність — тим більше інформації



Чим більша невизначеність системи (більша ентропія), тобто чим більша кількість можливих станів (різноманіття), які вона може приймати, тим більше інформації несе повідомлення про її стан.



Для ілюстрації сказаного розглянемо приклад:
про 4 стани системи.

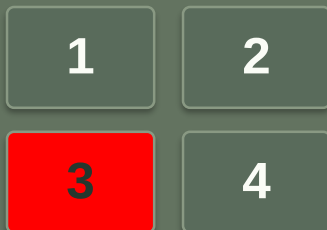


Приклад 1: дві зони, чотири сектори

Дві географічні зони поділені на чотири сектори кожна. Потрібно оцінити кількість інформації у повідомленнях спостерігачів.

Перша зона

«Пожежа можлива тільки у секторі 3».

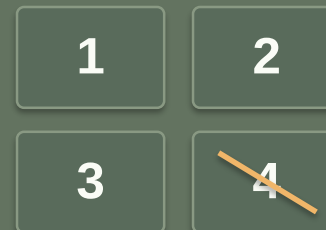


Невідомих
станів: 0

$$I_1 = H - H'_1 = 4$$

Друга зона

«У секторі 4 пожежа неможлива».



Невідомих
станів: 3

$$I_2 = H - H'_2 = 4 - 3 = 1$$

Попередня оцінка у вихідному прикладі: $H = N = 4$.

Однак число станів саме по собі ще не є зручною мірою інформації.

Приклад 1. Дві географічні зони поділені на чотири сектори кожна. Спостерігач першої зони доповів, що пожежа може виникнути тільки у секторі 3. Спостерігач другої зони доповів, що у секторі 4 пожежа неможлива. Потрібно оцінити кількість інформації у повідомленнях спостерігачів.

Припустимо, що кожна зона – це система, яка має 4 стани ($N = 4$), тому що пожежа може виникнути у кожному з 4-х секторів. Вважаючи кожний стан рівноімовірним, ступінь невизначеності (ентропії) H можна прийняти рівною кількості станів: $H = N = 4$ одиниці.

Будемо оцінювати кількість інформації I у повідомленні кожного спостерігача як величину, на яку зменшилась невизначеність (ентропія H) системи в результаті отримання повідомлення.

Після повідомлення першого спостерігача система стала повністю відомою, тобто "пожежа у секторі 3", невідомих (невизначених) станів немає. Тому $N'_1 = 0$, а значить ентропія у даному випадку повністю знята, тобто $H'_1 = 0$. Отже, кількість інформації в доповіді першого спостерігача оцінюється так:

$$I_1 = H - H'_1 = 4, \quad (1.1)$$

Повідомлення другого спостерігача зменшило ентропію тільки на одиницю ("у 4-му секторі пожежі немає"), але залишилися невизначеними останні три стани системи ($N'_2 = 3$).

Тому ентропія знята не повністю, $H'_2 = 3$ од. Кількість інформації I_2 прийме значення:

$$I_2 = H - H'_2 = 4 - 3 = 1 \text{ (од.)} \quad (1.2)$$

З прикладу 1 видно, що чим більша кількість можливих станів системи N , тим більша ентропія (невизначеність) H системи, тому більшу кількість інформації несе в собі повідомлення про стан системи.

Таким чином, кількісною мірою інформації (або кількісною мірою невизначеності) може бути число можливих станів системи N :

$$I = H - H' = N - N'.$$

? Однак...

Чому кількість станів — ще не міра

Якщо $N = 1$, система визначена: її ентропія має дорівнювати нулю.
Проста оцінка $H = N$ дає 1, а не 0.

Вимоги до кількісної міри інформації

Зв'язок із числом станів

$$I = f(N)$$

Нуль для визначеної системи

$$N = 1 \Rightarrow I = 0$$

Зростання разом із N

$$N \uparrow \Rightarrow I \uparrow$$

Невід'ємність

$$I \geq 0$$

Цим вимогам відповідає логарифмічна міра,
запропонована Р. Хартлі у 1928 році.

$$I = \log N$$

Інформація як зменшення ентропії

До повідомлення

N — кількість можливих станів системи.

$$H(x) = \log N$$

Після повідомлення

N' — кількість станів, що залишилися можливими.

$$H'(x) = \log N'$$

$$I = H(x) - H'(x) = \log N - \log N' = \log \frac{N}{N'}$$

За повних відомостей відомий один стан: $N' = 1$. Тоді невизначеність знято повністю.

2. Ентропія та її властивості.

Однак такий спосіб вимірювання кількості інформації незручний тим, що при $N = 1$ (система може знаходитись тільки в одному стані, тобто система визначена) маємо $N = H = 1$, коли ентропія H повинна дорівнювати нулю (для визначеної системи $H = 0$). Тому при оцінюванні кількості інформації (1.3) в повідомленнях невизначеність (ентропію) системи зручніше виражати не числом можливих невідомих станів її до отримання повідомлення ($N = H$) та після отримання ($N' = H'$), а деякою іншою величиною, яка була б функціонально зв'язана з N , але не суперечила б логічним представленням про невизначеність (ентропію) та інформацію.

Отже, вимогами до кількісної міри інформації є:

- функціональний зв'язок з числом можливих невідомих станів системи N ($I = f(N)$);
- рівність нулю ($I = 0$), коли система визначена, тобто $N = 1$;
- збільшення з ростом N ($I \uparrow = f(N \uparrow)$);
- приймати значення не менше нуля $I \geq 0$ (тому що повідомлення може або нести інформацію ($I > 0$), або ні ($I = 0$), але виносити інформацію не може).

Таким вимогам задовольняє логарифмічна міра інформації, яка була запропонована американським вченим Л. Хартлі у 1928 р.:

$$I = \log N. \quad (1.4)$$

Якщо позначити число можливих станів системи до повідомлення як N , а після повідомлення – як N' , ентропія H стану системи x до повідомлення $H(x) = \log N$, після повідомлення $H'(x) = \log N'$. Тоді кількість інформації можна оцінити як різницю ентропій стану системи до та після повідомлення:

$$I = H(x) - H'(x) = \log N - \log N' = \log \frac{N}{N'}. \quad (1.5)$$

При повних відомостях про стан системи, тобто якщо після отримання повідомлення точно відомий тільки один стан системи $N' = 1$ (повідомлення першого спостерігача у прикладі 1), то з формули (1.5) маємо

$$I = \log N$$

? Однак...

Ентропія за Шенноном

Хартлі передбачає рівноймовірні стани. Для реальних систем із різними ймовірностями К. Шеннон у 1948 році запропонував статистичну оцінку.

$$H(x) = - \sum_{i=1}^N p_i \log_2 p_i$$

Сума добутків ймовірностей станів на логарифми цих ймовірностей, узята зі знаком «мінус».

Визначена система

Один стан має ймовірність $p = 1$.
Невизначеність відсутня.

$$H(x) = -1 \cdot \log 1 = 0$$

Максимальна ентропія

Усі стани рівноймовірні: $p_1 = p_2 = \dots = 1/N$.

$$H_{\max}(x) = \log N$$

Формула Хартлі має один суттєвий недолік – усі можливі стани системи передбачаються рівнозначними, тобто ймовірності появи одного з можливих станів рівні:

$$p_1 = p_2 = p_3 = \dots = p_N.$$

Реальні системи, як правило, мають нерівнозначні стани, тому у 1948 р. американський математик Клод Ельвуд Шеннон запропонував оригінальний метод оцінки кількості інформації з використанням тієї ж логарифмічної функції, але з урахуванням ймовірності появи повідомлень про відповідний стан системи, що поклало початок розвитку статистичної теорії інформації.

За Шенноном, для визначення ентропії системи беруть суму добутків ймовірностей p_i появи кожного з i -х станів системи на логарифми цих ймовірностей:

$$H(x) = - \sum_{i=1}^N p_i \log p_i, \quad (1.7)$$

де знак мінус перед сумою враховує від'ємне значення логарифмічної функції, оскільки $0 \leq p_i \leq 1$, то $\log p_i < 0$.

Ентропія зручна для оцінки ступеня невизначеності системи через ряд властивостей

1) ентропія дорівнює нулю, якщо система може приймати один стан з ймовірністю $p = 1$, тобто коли стан системи визначений: $H(x) = -1 \log 1 = 0$;

2) максимальне значення ентропія досягає, коли всі стани системи рівноімовірні, тобто $p_1 = p_2 = \dots = p = \frac{1}{N}$. Із формули (1.7) маємо:

$$H_{\max}(x) = -Np \log p = -N \frac{1}{N} \log \frac{1}{N} = -1(\log 1 - \log N) = \log N.$$

Для аналізу інформаційних процесів в комп'ютерних системах використовують логарифм з основою 2 ($\log_2 N$), оскільки в принципах побудови і функціонування сучасної цифрової техніки закладена двійкова система числення. Тоді одиницею вимірювання ентропії, а також інформації називають "біт" (від англ. bit – binary digit, двійковий знак).

Таким чином, кількість інформації про деяку систему X дорівнює значенню, на яке зменшується ентропія цієї системи у результаті отримання повідомлення (1.5):

$$I = H(x) - H'(x).$$

Приклад 2: оцінка інформації у бітах

У комп'ютерних системах використовують логарифм з основою 2. Одиниця інформації та ентропії — біт (binary digit — двійковий знак).

ЧОТИРИ РІВНОЙМОВІРНІ СТАНИ

Перше повідомлення

Стан визначено: залишився один сектор.

$$I_1 = \log_2 4 - \log_2 1 = 2$$

Друге повідомлення

Виключено один сектор: можливі три стани.

$$I_2 = \log_2 4 - \log_2 3 \approx 0,4$$

Одиниця результатів — біт.

Приклад: нерівноймовірні стани

ІМОВІРНОСТІ З ВИХІДНОГО ПРИКЛАДУ

$$p_1 = 0,2$$

$$p_2 = 0,3$$

$$p_3 = 0,4$$

$$p_4 = 0,1$$

$$I_1(x) = - \sum_{i=1}^4 p_i \log_2 p_i = 1,85$$

$$I_2(x) = - \sum_{i=1}^4 p_i \log_2 p_i - \left(- \sum_{i=1}^3 p_i \log_2 p_i \right)$$

$$I_2(x) = 1,85 - 1,51 = 0,34$$

Одиниця результатів — біт.

Приклад 2. Кількість інформації у повідомленнях (при використанні ПЕОМ) можна визначити так:

$$I_1(x) = H_1(x) - H_1^1(x) = \log_2 4 - \log_2 1 = 2 \text{ біт},$$

$$I_2(x) = H_2(x) - H_2^1(x) = \log_2 4 - \log_2 3 = 2 - 1,6 = 0,4 \text{ біт}.$$

Якщо стани системи мають різну ймовірність появи ($p_1 \neq p_2 \neq p_3 \neq \dots \neq p_N$), то для оцінки кількості інформації використовують формулу Шеннона. Наприклад, якщо $p_1 = 0,2$; $p_2 = 0,3$; $p_3 = 0,4$; $p_4 = 0,1$, то для повідомлення прикладу маємо:

$$I_1(x) = -\sum_{i=1}^4 p_i \log_2 p_i = 1,85 \text{ біт},$$

$$I_2(x) = -\sum_{i=1}^4 p_i \log_2 p_i - \left(-\sum_{i=1}^3 p_i \log_2 p_i \right) = 1,85 + (0,2 \log_2 0,2 + 0,3 \log_2 0,3 + 0,4 \log_2 0,4) = 1,85 - 1,51 = 0,34 \text{ біт}.$$

Із порівняння формул (1.4) і (1.7), підхід запропонований Хартлі (для систем із рівноймовірними станами) є частковим рішенням формули Шеннона. Для систем, які мають різні ймовірнісні стани, розрахунки ентропії та кількості інформації за Хартлі дадуть завищені результати.

Статистичний підхід до кількісної оцінки інформації (формула Шеннона) широко використовується в наукових дослідженнях та при розв'язуванні багатьох важливих практичних питань. Наприклад, задачі оптимального кодування повідомлень в системах передачі інформації, визначення пропускної спроможності каналів з завадами, розрахунок ємності запам'ятовуючих пристроїв для зберігання інформації тощо. При цьому кількість інформації можна вимірювати в бітах.

Формула Хартлі для повідомлення

За повної незалежності та рівноймовірності символів формула визначає максимально можливу кількість інформації в повідомленні заданої довжини n .

$$I_n = n \cdot \log_2 N = -n \cdot \log_2 P_N$$

Кількість інформації в повідомленні, біт

$$H = \frac{I_n}{n} = \log_2 N = -\log_2 P_N$$

Середня кількість інформації на один символ

n — кількість символів у повідомленні; N — кількість можливих символів; P з індексом N у позначеннях — ймовірність появи одного символу.

Лекція 2

Якщо імовірності появи символів відрізняються

$$P_1 \neq P_2 \neq \dots \neq P_N$$

1. Кількість інформації в повідомленні зменшується (формула Шеннона)

$$I_n = -n \cdot \sum_{i=1}^N P_i \cdot \log_2 P_i \quad (\text{біт})$$

Якщо імовірності появи символів відрізняються, кількість інформації в повідомленні зменшується.

2. Середня кількість інформації на один символ (ентропія)

$$H = \frac{I_n}{n} = - \sum_{i=1}^N P_i \cdot \log_2 P_i \quad (\text{біт/символ})$$

3. Чому інформації стає менше?



Причиною зменшення ентропії (а відповідно і кількості інформації) є нерівномірність імовірностей появи символів, а також наявність статистичної залежності між символами (кореляції) у повідомленні. Це призводить до зменшення кількості інформації в реальних повідомленнях.

Лекція 2

2. Надмірність інформації

Надмірність джерела повідомлень — це його властивість видавати інформацію більшою кількістю знаків, ніж можна було б.

$$R = \frac{H_{\max} - H}{H_{\max}} = 1 - \frac{H}{H_{\max}} = 1 - \frac{H}{\log_2 N}$$

Завдяки надмірності повідомлення захищені від перешкод.
Це використовується при **завадостійкому кодуванні**.

кодові комбінації для передачі інформації про 1 із 4-х станів	контрольний надмірний елемент (сума по модулю 2 початкової комбінації)	код з виявленням одиничної помилки
00	0	000
01	1	011
10	1	101
11	0	110

Коефіцієнти надмірності

За однакової кількості інформації повідомлення тим довше, чим менша його ентропія.

$$r = \frac{H_2 - H_1}{H_2} = 1 - \frac{H_1}{H_2} = 1 - \mu$$

Чим більша ентропія, тим більшу кількість інформації містить у середньому кожний елемент повідомлення.

Нехай ентропії двох джерел повідомлень $H_1 < H_2$, а кількість інформації, що одержується від них, однакова, тобто $I = n_1 H_1 = n_2 H_2$, де n_1 і n_2 – довжина повідомлення від першого та другого джерел. Позначимо

$$\mu = \frac{n_2}{n_1} = \frac{H_1}{H_2}$$

При передачі однакової кількості інформації ПОВІДОМЛЕННЯ тим ДОВШЕ, чим МЕНША його ЕНТРОПІЯ.

Коефіцієнт стиску μ – величина, що характеризує ступінь укорочення повідомлення при переході до кодування станів елементів, що характеризується більшою ентропією.

Коефіцієнт надмірності r – частка зайвих елементів при переході до кодування станів елементів, що характеризуються більшою ентропією.

$$r = \frac{H_2 - H_1}{H_2} = 1 - \frac{H_1}{H_2} = 1 - \mu$$

Продуктивність джерела інформації

Продуктивність — кількість інформації, що виробляється джерелом за одиницю часу.

$$\overline{H} = \frac{H}{\tau}$$

$$\tau = \sum_{i=1}^M p(x_i) \tau_i$$

τ — середня тривалість символу;
 τ_i — тривалість символу x_i .

Надмірність дискретного джерела з алфавітом потужності M

$$R = \frac{H_{\max} - H}{H_{\max}} = \frac{\log_2 M - H}{\log_2 M} = 1 - \frac{H}{\log_2 M}$$



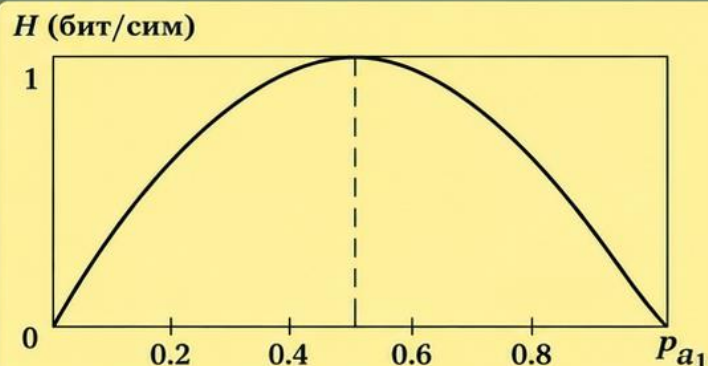
Висновки:

1. Ентропія є мірою невизначеності, непередбачуваності ситуації. Зменшення ентропії, що відбулось завдяки деякому повідомленню, точно збігається з кількістю інформації, яка міститься в цьому повідомленні.

2.

Ентропія дорівнює нулю, якщо повідомлення відоме завчасу.

3. Ентропія максимальна, якщо всі знаки (елементи) алфавіту рівноймовірні.



4.

Ентропія, продуктивність та надмірність — інтегральні інформаційні характеристики дискретного джерела інформації.



ІНФОРМАЦІЯ
ЗМІНЮЄ
СВІТ