Лекція 2

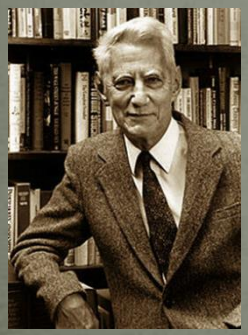
Інформація та ентропія: кількісна міра

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 Ше́ннон

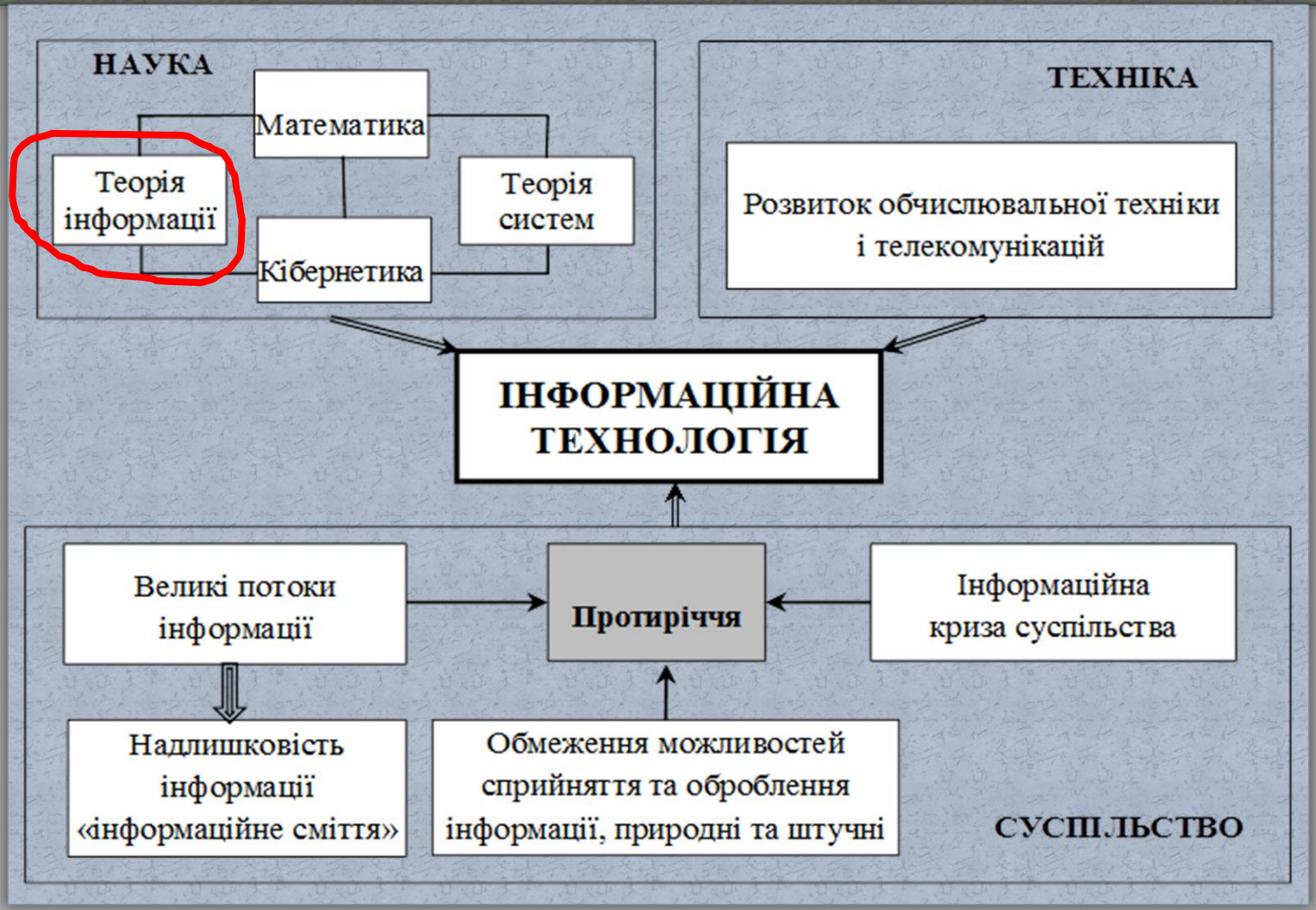
Ральф Вінтон Лайон Хартлі



Нагадування

Сутність та зміст інформаційної технології

На формування сутності та змісту ІТ на сучасному етапі її становлення впливає тріада, базис якої становить суспільство, наука і техніка

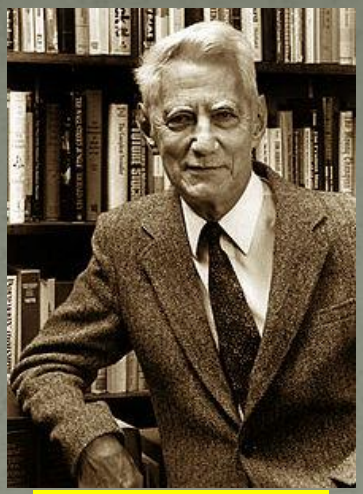


Юрій Бродський

Бродський Ю.Б., Грищук Р.В. Аналіз ролі та місця, сутності та змісту інформаційних технологій на сучасному етапі розвитку науки і техніки // Проблеми інформаційних технологій. – 2018. – №1 (023). С.36-42.

Джон фон Нейман
1903–1957





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

Клод Елвуд Ше́ннон
(1916-2001, інженер, математик, основоположник теорії інформації

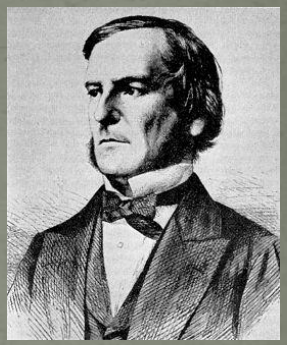


$$I = \log_2 N$$

Ральф Вінтон Лайон Хартлі
(1888 — 1970,США)



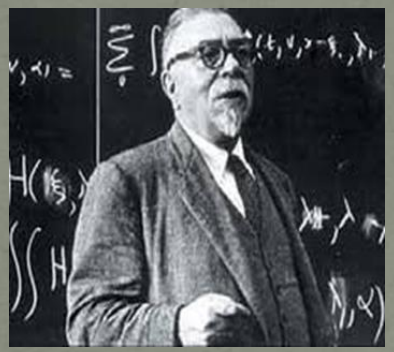
Гаррі Найквіст (1889-1976), швед



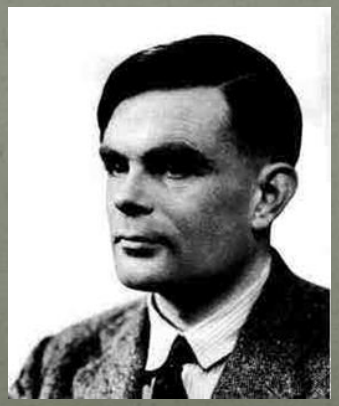
Джордж Буль (1815 — 1864, Ірландія) — англійський математик й логік.



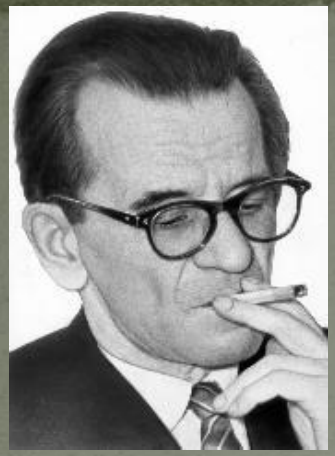
У. Р. Ешбі (1903-1972), Англія
метод «Чорного ящика»



Норберт Вінер (1894-1964) основоположник кібернетики



Т'юринг (Turing) Алан Матисон (23.6.1912, Лондон, — 7.6.1954, Уілмслоу, близ Манчестера), математик.



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) в напівзруйнованій церкві першої в СРСР лампової обчислювальної машини МЕОМ (малої електронної обчислювальної машини).



Віктор Михайлович Глушков
(1923-1982), вітчизняний вчений, академік, один із основоположників вітчизняної обчислювальної техніки та інформатики.
У 1996 г. міжнародним комп'ютерним товариством IEEE (International Electrical and Electronic Engineers) Computer Society було присвоєно академіку В.М.Глушкову звання «Computer Pioneer» за його внесок у розвиток автоматики та обчислювальної техніки.



Юрій Бродський

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

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:

- оцінка кількості інформації;
- математичний опис сигналів інформації;
- перетворення сигналів інформації в процесах модуляції, демодуляції, фільтрації, аналого-цифрового та цифро-аналогового перетворень;
- кодування та декодування сигналів;
- розробка технологій передачі інформації по каналах зв'язку.

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

Постулати теорії інформації :

– джерело повідомлення здійснює вибір повідомлення з деякої множини з певною ймовірністю;

– повідомлення можуть передаватися по каналу зв’язку в закодованому виді. Кодовані повідомлення утворюють множину, що є взаємно однозначним відображенням множини повідомлень. Правило декодування відоме декодеру (записане в його програмі).

– повідомлення вважається прийнятим правильно, якщо в результаті декодування воно може бути в точності відновленим. При цьому не враховується, скільки часу пройшло з моменту передачі повідомлення до моменту закінчення декодування, та яка складність операцій кодування і декодування.

Основні поняття

Інформація – продукт взаємодії даних і адекватних їм методів

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

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

ПОВІДОМЛЕННЯ – дані, що підлягають передачі.

СИГНАЛ – це матеріальний носій повідомлення.

СИГНАЛИ ⟶ ПОВІДОМЛЕННЯ ⟶ ДАНІ ⟶ ІНФОРМАЦІЯ

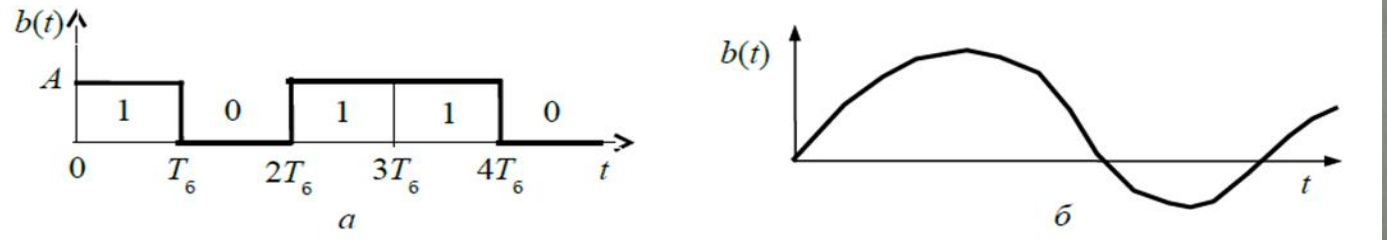


Рисунок 1 – Первинні сигнали: *а* – цифровий сигнал; *б* – неперервний сигнал

Інформація існує у вигляді:

- документів, креслень, рисунків, текстів;
- світових 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 і та ін.

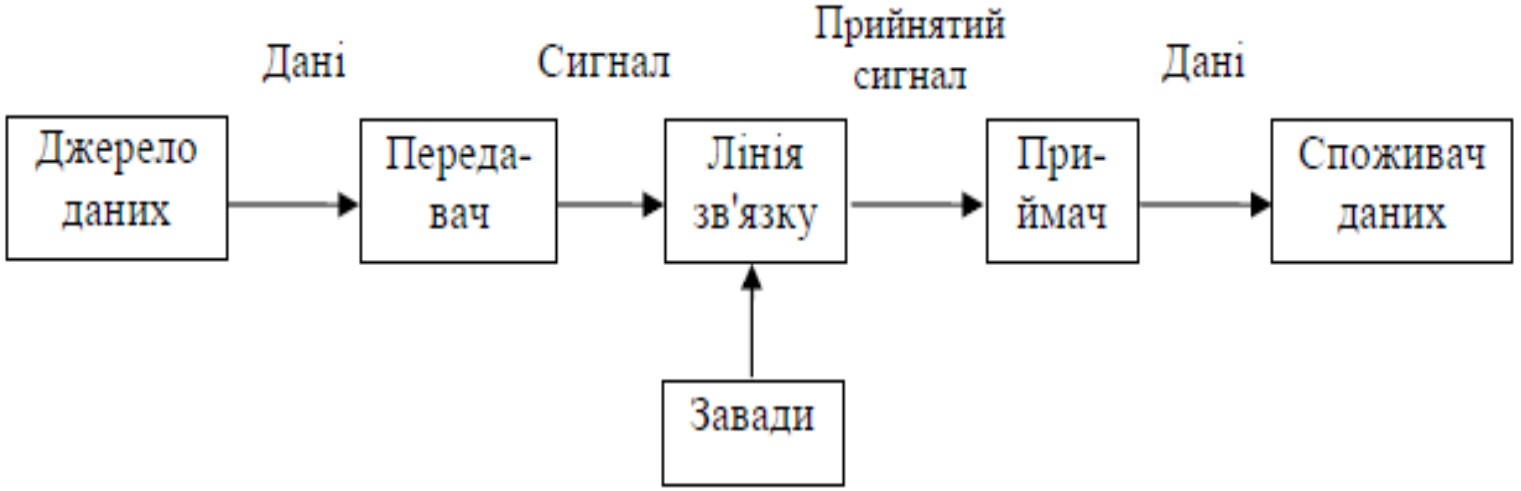
Найбільш важливими властивостями інформації є:

- об'єктивність та суб'єктивність;
- повнота;
- достовірність;
- адекватність;
- доступність;
- цінність;
- своєчасність;
- актуальність.

Інформацію можна:

- створювати;
- передавати;
- сприймати;
- використовувати;
- запам'ятовувати;
- приймати;
- копіювати;
- формалізувати;
- розповсюджувати;
- перетворювати;
- комбінувати;
- обробляти;
- ділити на частини;
- спрощувати;
- збирати;
- зберігати;
- шукати;
- вимірювати;
- знищувати;

Загальна модель зв'язку по К. Шеннону приведена на рис. В.1.



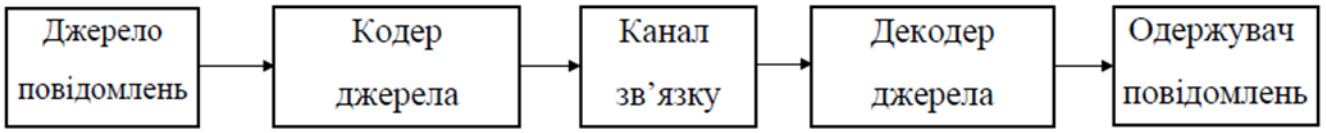


Рисунок 2 – Система електричного зв'язку

ВИСНОВОК:

- теорія інформації, як наука, вивчає:
  - кількісну міру інформації,
  - інформаційні характеристики джерел повідомлень і каналів зв’язку,
  - кодування в дискретних та неперервних каналах,
  - коди, що виявляють та виправляють помилки,
  - ефективність кодування та передачі інформації.

1. Кількісна міра інформації

В реальних системах для кількісної оцінки інформації використовують статистичний підхід, в рамках якого інформація розглядається як сукупність відомостей, повідомлень про поведінку деякої системи, яка випадково може знаходитись в одному з можливих станів. Така система має деяку ступінь невизначеності (ентропію) і фактичний стан її до отримання повідомлення залишається невідомим. Повідомлення про фактичний стан системи, яку ми розглядаємо, і є інформація про неї.

Якщо стан системи визначений і не може змінюватись або всі можливі зміни станів відомі до появи повідомлення про них, то смислу передавати повідомлення про систему немає. Воно не представляє інтересу, не дає нічого нового, тобто не несе інформацію (наприклад, "січень - перший місяць року"). Очевидно й те, що повідомлення про невідомий фактичний стан системи буде передавати тим більше інформації, чим більша степінь невизначеності системи (більша ентропія системи), тобто чим більша кількість станів (різноманіття), які вона може приймати.

Для ілюстрації сказаного розглянемо приклад: (про 4 стани системи «Інформатика і системологія» [3]).

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:

$$I_1 = H - H_1' = 4, \qquad (1.1)$$

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:

$$I_2 = H - H_2' = 4 - 3 = 1 \text{ (од.)} \qquad (1.2)$$

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…

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.:

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

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:

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

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 $I=\log_2 N$

? Однак...

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

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

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:

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

де знак мінус перед сумою враховує від'ємне значення логарифмічної функції, оскільки $0 \le p_i \le 1$, 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 $p_1 = p_2 = \ldots = p = \frac{1}{N}$. 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. 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 ($p_1 \neq p_2 \neq p_3 \neq \ldots \neq p_N$), 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.

Таким чином:

кількість інформації за Хартлі, тобто у припущенні повної незалежності та рівноймовірності появи окремих символів повідомлення, *визначає максимально можливу кількість інформації в повідомленні заданої довжини (n)*

$$I_n = n \cdot \log_2 N \quad \text{(біт)}$$

Якщо поділити на кількість символів у повідомленні n отримаємо середню кількість інформації, що припадає на один символ:

$$I_1 = \log_2 N = -\log_2 P_N, \frac{біт}{символ}, \quad N = \frac{1}{P_N}$$

P_N - імовірність появи одного символу.

Якщо імовірності появи символів відрізняються $P_1 \neq P_2 \neq \ldots \neq P_N$ кількість інформації в повідомленні зменшується (формула Шеннона) :

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

Якщо поділити на кількість символів у повідомленні n отримаємо середню кількість інформації, що припадає на один символ (ентропію):

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

Іншим фактором, що зменшує ентропію (відповідно і кількість інформації) в повідомленні довжиною n – наявність статистичної залежності між символами (кореляції): кореляційні зв'язки між символами та неоднакова ймовірність їх появи приводить до зменшення кількості інформації в реальних повідомленнях.

Кількісно такі втрати характеризуються коефіцієнтом надмірності.

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.

наприклад

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

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 $\overline{H}$ джерела інформації – це кількість інформації, що виробляється джерелом за одиницю часу:

$$\overline{H} = H / \tau \ , \qquad (1.4)$$

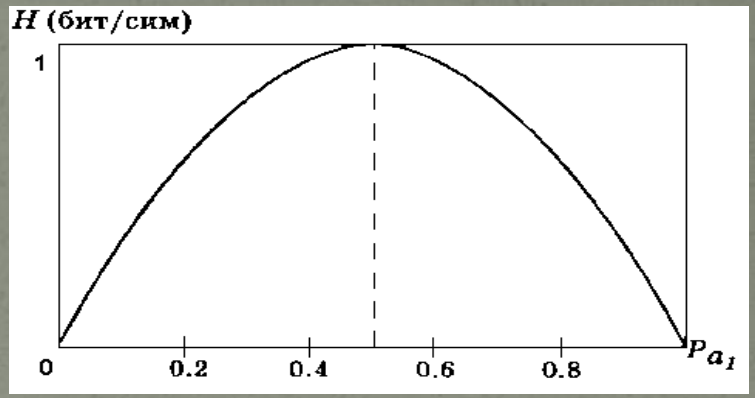
де $\tau = \sum_{i=1}^{M} p(x_i)\tau_i$ – середня тривалість символу, τ_i – тривалість символу x_i.

Висновки:

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

- Ентропія дорівнює нулю, якщо повідомлення відоме завчасу.

- Ентропія максимальна, якщо всі знаки (елементи) алфавіту рівноймовірні



- ***Ентропія, продуктивність та надмірність*** – інтегральні інформаційні характеристики дискретного джерела інформації

Ю. Бродський (кафедра КБ та ЗІ)