Тема2. Інформаційні характеристики джерел дискретних повідомлень

2.1. Інформація, повідомлення, сигнал.

Інформація – це будь-які відомості, що передаються будь-яким чином між джерелом інформації і одержувачем інформації. Інформація це категорія, яка більше відноситься до філософії. Для системи передавання більш корисною є категорія повідомлення – спосіб представлення інформації. Однакова інформація може бути представлена у вигляді:

- ✔мовного повідомлення;
- ✔візуального повідомлення;
- ✔текстового повідомлення.

Перераховані види повідомлень є вихідними для будь-якої існуючої системи передачі або системи електрозв'язку. Однак ці види повідомлень не пристосовані для передавання системами електрозв'язку, тому вони перетворюються у особливий вид повідомлення – електричний сигнал.

Дискретне джерело інформації – це таке джерело, яке може виробити (згенерувати) за скінчений відрізок часу тільки скінчену множину повідомлень. Кожному такому повідомленню можна співставити відповідне число, та передавати ці числа замість повідомлень.

Дискретне джерело інформації є достатньо адекватною інформаційною моделлю дискретних систем, а також неперервних систем, інформаційні сигнали про стан яких піддають аналого - цифровому перетворенню; таке перетворення виконується в більшості сучасних автоматизованих систем управління.

Первинні характеристики дискретного джерела інформації – це алфавіт, сукупність ймовірностей появи символів алфавіту на виході дискретного джерела та тривалості символів. $X = \{x_1, x_2, x_3, \square, x_M\}$

Алфавіт – множина символів, які можуть з’явитися на виході дискретного джерела; M – алфавіт джерела, тобто кількість різноманітних символів алфавіту.

Якщо всі ймовірності, які визначають виникнення символів на виході джерела, не залежать від часу, джерело називають ***стаціонарним.*** Ми будемо розглядати тільки стаціонарні джерела та для скорочення замість “стаціонарне джерело” будемо всюди використовувати “джерело”.

Для опису джерел, які не мають пам’яті, достатньо мати значення безумовних імовірностей $p(x_i)$ виникнення символів x_i, $i = 1, 2, 3, \ldots, M$ на його виході.

Більшість реальних джерел інформації є джерелами ***з пам'яттю***. Розподіл ймовірностей виникнення чергового символу на виході дискретного джерела з пам'яттю залежить від того, які символи були попередніми. Таке джерело інформації називають ***марковським***, оскільки процес появи символів на його виході адекватний ланцюгам Маркова; останні в свою чергу отримали таку назву на честь російського математика Маркова Андрія Андрійовича (1856 – 1922), який заклав основи розділу теорії випадкових процесів.

Будемо говорити, що ***глибина пам'яті*** марковського дискретного джерела інформації дорівнює h, ($h \geq 0$), якщо ймовірність появи чергового символу залежить тільки від h попередніх символів на виході цього джерела.

Сигнал – це фізичний процес або явище, що переносить інформацію о будь-якій події або о стані об’єкта спостереження. Якщо фізичним процесом є електричний струм, електрична напруга або напруженість поля електромагнітної хвилі, то сигнал називається електричним. Отже електричні сигнали є переносниками інформації (повідомлень) у радіотехнічних системах.

2.2. Кількісна міра інформації

Кількість інформації – одне із основних понять теорії інформації, яка розглядає технічні аспекти інформаційних проблем, тобто вона дає відповіді на запитання такого типу: якою повинна бути ємність запам’ятовуючого пристрою для запису даних про стан деякої системи, якими повинні бути характеристики каналу зв’язку для передачі певного повідомлення тощо.

Кількість інформації $I(a_k)$, дв. од., у повідомленні a_k з імовірністю його появи $P(a_k)$, обчислюється як

$$I(a_k) = -\log_2 P(a_k)$$

Ентропія джерела $H(A)$, це – середня кількість інформації в одному повідомленні. Для M незалежних повідомлень обчислюється як математичне сподівання

$$H(A) = -\sum_{k=1}^{M} P(a_k)\log_2 P(a_k)$$

Ентропія, як і інформація, завжди додатна й досягає максимального значення

$$H_{\max}(A) = \log_2 M.$$

Надмірність (надлишковість) джерела характеризує зменшення ентропії джерела в порівнянні з максимальним значенням внаслідок того, що деякі з повідомлень несуть малу (а то і нульову) кількість інформації. Кількісно надмірність оцінюється коефіцієнтом $K_{\text{н}}$, що називається коефіцієнтом надмірності і визначається за формулою

$$K_{\text{Н}} = \frac{H_{\max}(A) - H(A)}{H_{\max}(A)}$$

Продуктивність джерела, $R_{\text{дж}}$, біт/с – це середня кількість інформації, що видається джерелом за одиницю часу

$$R_{\text{дж}} = H(A)/\, T_{\text{сер}}$$

де $T_{\text{сер}}$ – середня тривалість одного повідомлення джерела.

2.3 Математична модель двох джерел дискретних повідомлень та її статистичні характеристики.

Математична модель двох джерел. За термінологією два стаціонарні джерела *A* та *B* мають об'єднаний ансамбль *AB* і видають дискретні повідомлення a_k та b_k.

***Статистичні характеристики двох джерел повідомлень такі*:**

– апріорна (безумовна) імовірність повідомлень a_k та b_k – $P(a_k)$ та $P(b_k)$;

– апостеріорна (умовна) імовірність повідомлення a_k джерела A, якщо має місце повідомлення b_k джерела B – $P(a_k/b_k)$;

– спільна ймовірність повідомлень a_k та b_k –

$$P(a_k, b_k) = P(a_k)P(b_k/a_k) = \\ = P(b_k)P(a_k/b_k);$$

– $T_{зн}$ – тривалістю видачі повідомлення a_k чи b_k джерелом.

Типовим прикладом двох джерел повідомлень є повідомлення від якогось джерела *A* на вході каналу зв'язку та ті ж самі повідомлення на виході каналу. Ми спостерігаємо повідомлення на виході каналу (джерело *B*), а інформацію маємо отримати про джерело *A*.

Кількість інформації в повідомленні a_k чи b_k об' єднаного ансамблю AB двох джерел повідомлень A і B характеризується *умовною* та *спільною* інформацією.

Умовна кількість інформації в повідомленні a_k чи b_k об'єднаного ансамблю AB двох джерел дискретних повідомлень A і B $i(a_k/b_k)$, визначається:

$$i(a_k/b_k) = -\log_2 P(a_k/b_k)$$
$$i(b_k/a_k) = -\log_2 P(b_k/a_k) \qquad [\text{дв.од.}]$$

Спільна кількість інформації в повідомленнях a_k та b_k об’єднаного ансамблю *AB* двох джерел повідомлень A та B – $i(a_k, b_k)$ чи $i(b_k, a_k)$, визначається:

$$i(a_k, b_k) = -\log_2 P(a_k, b_k)$$
$$i(b_k. a_k) = -\log_2 P(b_k. a_k) \qquad \text{[дв.од.]}$$

Середня кількість інформації об’єднаного ансамблю AB двох джерел повідомлень A та B характеризується ***умовною, спільною*** та ***взаємною*** ентропіями.

Умовна ентропія – середня кількість інформації, яку переносить одне повідомлення джерела при залежних повідомленнях. **Умовна ентропія** двох джерел повідомлень A і B чи B і A, визначається:

$$H(A/B) = M\left(i(a_k/b_k)\right)$$
$$H(B/A) = M\left(i(b_k/a_k)\right) \qquad \text{[дв.од./зн. (біт/зн.)]}$$

Спільна ентропія – середня кількість інформації, яку переносять повідомлення джерел *A* та *B* з врахуванням статистичної залежності між ними.

Спільна ентропія об'єднаного ансамблю AB двох джерел повідомлень A та B, визначається:

$$H_{\text{сп}}(AB) = H(A) + H_{\text{ум}}(B/A) =$$
$$= H(B) + H_{\text{ум}}(A/B) \qquad [\text{дв.од./зн. (біт/зн.)}]$$

Взаємна ентропія об'єднаного ансамблю *AB* двох джерел повідомлень показує середню кількість інформації, яку можна отримати про джерело *A*, спостерігаючи джерело *B*.

Взаємна ентропія об'єднаного ансамблю *AB* двох джерел повідомлень A та B, визначається:

$$H_{\text{вз}}(A, B) = H(A) - H_{\text{ум}}(A/B) =$$
$$= H(B) - H_{\text{ум}}(B/A) \qquad [\text{дв.од./зн. (біт/зн.)}]$$

Продуктивність об’єднаного ансамблю AB двох джерел повідомлень A та B, визначається:

$$R_{\text{дж}}(A,B) = H_{\text{сп}}(A,B)/T_{ср} \qquad [\text{дв.од./с (біт/с)}]$$

Швидкість передачі інформації між двома джерелами повідомлень A та B, визначається:

$$R(A \rightarrow B) = H_{\text{вз}}(A, B)/T_{ср}\,[\text{дв.од./с (біт/с)}]$$

2.4 Ентропія джерела залежних повідомлень.

Користуючись поняттям умовної ентропії, можна отримати вираз для обчислення ентропії $H_{\text{п}}(X)$ джерела з пам'яттю, яке має алфавіт X. Якщо глибина пам'яті такого джерела дорівнює h, а потужність алфавіту M, то можна вважати, що перед генерацією чергового символу джерело знаходиться в одному з $Q=M^h$ станів, де під станом розуміємо одну з можливих послідовностей попередніх символів довжиною h на його виході.

Тоді частинна умовна ентропія $H(X/S)$ при умові, що джерело перебуває в s-му стані

$$H(X/S) = -\sum_{i=1}^{M} p(x_i/S) \cdot log_2 p(x_i/S),$$

де $p(x_i/s)$ – умовна ймовірність появи символу x_i, якщо джерело перебуває в s-му стані.

Усереднюючи $H(X/S)$ по усіх станах, отримаємо вираз для ентропії марковського джерела:

$$H_П(X) = \sum_{s=1}^{Q} p(s) \cdot H(X/s) =$$

$$= -\sum_{s=1}^{Q} p(s) \cdot \sum_{i=1}^{M} p(x_i/s) \cdot log_2 p(x_i/s),$$

$p(s)$ – ймовірність перебування джерела в s-му стані.

Для джерела з глибиною пам’яті $h = 2$ стан визначається парою символів (x_i, x_j), а ентропія:

$$H_{П2}(X) = -\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} p(x_i, x_j, x_k) \cdot \log_2 p(x_k / x_i, x_j).$$

Аналогічно можна отримати вирази для ентропій марковських джерел при більш глибоких статистичних зв’язках.

Отже, *за наявності зв'язку між елементарними повідомленнями ентропія джерела знижується,* причому в тим більшому ступені, чим сильніше зв'язок між елементами повідомлення.

Таким чином, можна зробити наступні висновки щодо ступеня інформативності джерел повідомлень:

1. *Ентропія джерела і кількість інформації тим більше, чим більше розмір алфавіту джерела*.

2. *Ентропія джерела залежить від статистичних властивостей повідомлень. Ентропія максимальна, якщо повідомлення джерела рівноімовірні і статистично незалежні*.

3. *Ентропія джерела, що виробляє нерівноімовірні повідомлення, завжди менше за максимально досяжну*.

4. *За наявності статистичних зв'язків між елементарними повідомленнями* (*пам'яті джерела*) *його ентропія зменшується*.

Як приклад розглянемо джерело з алфавітом, що складається з літер A={*а, б, в,....., ю, я*}. Вважатимемо для простоти, що розмір алфавіту джерела $= 2^5 = 32$.

Якби всі літери алфавіту мали однакову імовірність і були статистично незалежні, то середня ентропія, що доводиться на один символ, склала б

$$H_{max}(A) = \log_2 32 = 5 \text{ біт/літеру.}$$

Якщо тепер врахувати лише різну імовірність літер в тексті (а неважко перевірити, що так воно і є), розрахункова ентропія складе

$$H(A) = 4{,}39 \text{ біт/літеру.}$$

З урахуванням кореляції (статистичного зв'язку) між двома і трьома сусідніми літерами (після літери *“П”* частіше зустрічається *“А”* і майже ніколи – *“Ю”* і *“Ц”*) ентропія зменшиться, відповідно, до

$$H(A) = 3{,}52 \text{ біт/літеру і } H(A) = 3{,}05 \text{ біт/літеру.}$$

Якщо врахувати кореляцію між вісьма і більш символами, ентропія зменшиться до

$$H(A) = 2{,}0 \text{ біт/літеру}$$

і далі залишається без змін.

Реальні джерела з одним і тим же розміром алфавіту можуть мати абсолютно різну ентропію. Наприклад надмірність літературного російського тексту складе

$$K = 1 - (\text{2 біт/літеру})/(\text{5 біт/літеру}) = 0{,}6 .$$

Іншими словами, при передаванні тексту по каналу зв'язку кожні шість букв з десяти переданих не несуть ніякої інформації і можуть без жодних втрат просто не передаватися.

Виникає питання: чи можна не займати носій інформації або канал зв'язку передачею символів, які практично не несуть інформації, або ж можливе таке перетворення вихідного повідомлення, при якому інформація скорочувалася б в мінімально необхідне для цього число символів?

Таке можливе і необхідне. Сьогодні системи передачі інформації і зв'язку просто не змогли би працювати, якби в них не використовувалось такого роду кодування. Не було б цифрового стільникового. Не працювали б системи цифрового супутникового телебачення, дуже неефективною була б робота Internet, не було б можливості подивитися відеофільм або послухати музику з лазерного диска. Все це забезпечується ефективним або економним кодуванням інформації в даних системах.