12. Ентропія джерел дискретних повідомлень. Джерело Хартлі, Бернуллі, Маркова. Надмірність джерел повідомлень

12.1. Ентропія джерел дискретних повідомлень.

Поняття джерела інформації є найважливішим поняттям теорії інформації. Як вже наголошувалося, джерело інформації може бути представлений ансамблем джерела. **Дискретним джерелом інформації** вважається джерело, представлене ансамблем, безліч значень якого є дискретною. **Безперервним джерелом інформації** вважається джерело, представлене ансамблем, безліч значень якого є безперервною.

З погляду теорії інформації джерело вважається заданим повністю, якщо є деяка імовірнісна модель, що дає опис імовірності процесу появи сигналів на виході джерела. Для опису послідовності станів джерела використовуються математичні моделі у вигляді дискретних або безперервних випадкових процесів. Основу цього опису складає імовірнісна міра ансамблю джерела. При цьому в загальному випадку вважається, що імовірнісна міра визначається станом джерела і може змінюватися під час переходу його з одного стану в інший. Для побудови дискретного джерела необхідно знати алфавіт джерела $X = \{x_1, \ldots, x_M\}$, з якого формуються букви u_i, відповідне певним станам U_i.

Хай $\bar{u} =(\ldots u_{-1}, u_0, u_1, \ldots)$ позначає послідовність букв, вироблювану джерелом, в якому кожна буква u_i вибирається з дискретного алфавіту X. Тоді повний опис імовірності джерела задається імовірністю $p(u_{j+1}, \ldots, u_{j+L})$, визначеною для всіх L послідовностей всіх початкових моментів j. При такому підході джерело описується як довільний дискретний випадковий процес.

Дискретне джерело називається **стаціонарним,** якщо опис імовірності не залежить від початку відліку часу (умови формування повідомлень дискретним джерелом не змінюються в часі), тобто якщо виконується умова:

$$p\,(u_1, u_2, \ldots, u_L) = p\,(u_{j+1}, u_{j+2}, \ldots, u_{j+L}) \qquad (12.1)$$

для всіх довжин L, цілих чисел j і послідовностей $\bar{u}_L=(u_{j+1}, u_{j+2},\ldots,u_{j+L})$.

Вираз (1) означає, що імовірність того, що джерело виробляє послідовність $\bar{u}_L=(u_1, u_2,\ldots,u_L)$ на інтервалі від 1 до L, рівна імовірності того, що виробляється точно така ж послідовність на інтервалі від j+1 до j+L. Тобто зрушення послідовності на j не змінює її імовірності.

Дискретне джерело називається **періодичним**, якщо (12.1) справедливо для всіх j, які є кратними деякого цілого числа $m > 1$. Якнайменше значення m, що задовольняє цій умові називається періодом m_T. Якщо розглядати блоки букв періодичного джерела з періодом m_T як деякі «супербукви» більшого алфавіту, то послідовність супербукв так само виявиться стаціонарною.

Стаціонарне джерело, що володіє властивістю ***ергодичності***, називається ергодичним. ***Ергодичність*** означає, що статистичні закономірності, одержані при дослідженні одного досить довгого повідомлення ($\bar{u}_L=(u_1, u_2,\ldots,u_L)$ при $L \to \infty$) з імовірністю близької до одиниці справедливі для всіх повідомлень, створюваних джерелом.

Основною інформаційною характеристикою джерел інформації є ентропія. Для дискретних джерел це ентропія на букву джерела, для безперервних –

диференціальна (відносна) ентропія.

12.2. Джерела Хартлі, Бернуллі, Маркова.

Отже, в теорії інформації ентропія джерела є його основною характеристикою. Ця величина кількісно характеризує невизначеність адресата відносно того, яке чергове повідомлення буде послане.

За відсутності перешкод в каналі зв'язку кількість інформації, що передається по каналу зв'язку, рівна апріорній ентропії адресата (або ентропії джерела). Тому **ентропія джерела** є **мірою інформативності повідомлень**, тобто характеризує кількість інформації, що міститься в повідомленнях, створюваних джерелом. Нарешті, виявляється, що **ентропія** джерела **визначає кількість двійкових знаків, необхідних для кодування створюваних джерелом повідомлень.**

Розглянемо різні джерела дискретних повідомлень і їх характеристики.

***Джерела Хартлі** – Джерела з рівноімовірними і статистично незалежними символами в повідомленнях.* При розгляді найпростішого джерела було встановлено, що його повідомлення математично моделюються дискретною послідовністю $\{x_{ij}\}$. Використовування математичної моделі для вивчення реального явища припускає обов'язковим переклад на мову математики властивостей досліджуваного явища; в даному випадку необхідно формалізувати правила складання дискретним джерелом повідомлень довжини n на своєму виході.

Що стосується простого джерела, то правила складання його повідомлень гранично прості: будь-який символ алфавіту в повідомленні може бути використаний незалежно від інших елементів і досконало рівноправний зі всіма іншими символами алфавіту.

Статистичний аналіз символів в таких послідовностях здійснювати не потрібно, оскільки у рівноімовірних статистично незалежних символів умовна імовірність дорівнює безумовним і визначається як $p(x_i/x_j)=p(x_i)=\frac{1}{M}$, тобто залежать лише від величини потужності алфавіту M. Такі джерела називатимемо джерелами нульового типа або **джерелами Хартлі**, а величини їх інформаційних мір відзначати підрядковим індексом «о». **Ентропія джерела H_0 виходить максимальною величиною**

$$H_0=H_{max}=\log M, \tag{12.2}$$

або

$$\mathbf{H'_0=H'_{max}=N* H_{max}}, \tag{12.3}$$

де H´ - продуктивність джерела.

Для **прикладу** здійснемо математичний опис декількох простих дискретних джерел:

- двійкового джерела:

$$M=2; x_1="1"; x_2="0"; n=5; p(x_1)=p(x_2)=\frac{1}{2};$$

$$H_0=\log_2 2=1\left[\frac{\text{дв.ед}}{\text{симв.}}\right];$$

$$I_0=5H_0=5[\text{дв.ед}];$$

- трійкового джерела:

$$M = 3; x_1 = "1"; x_2 = "0"; x_3 = "-1";$$

$$n = 5; p(x_1) = p(x_2) = p(x_3) = \frac{1}{3};$$

$$H_0 = \log_2 3 = 1{,}58\left[\frac{\text{дв.ед.}}{\text{симв.}}\right]; I_0 = 5H_0 = 7{,}9[\text{дв.ед}].$$

Таким чином, для математичного опису простих джерел достатньо вказати кількість символів в їх алфавіті M і довжину повідомлення n.

Джерела Бернуллі - Джерела з нерівноімовірними і статистично незалежними символами в повідомленнях. Нескладно уявити собі інші, ніж у джерел Хартлі, правила формування повідомлень з елементів алфавіту.

Припустимо, наприклад, що про джерело відоме наступне: алфавіт джерела має потужність M, але його здатність генерувати різні символи алфавіту різна: елемент x_1 він може генерувати необмежено, елементів x_2 джерело здатне створити за час своєї роботи декілька менше, ніж елементів x_1, елементів x_3 - ще менше і т.д. Очевидно, що в повідомленнях дуже великої довжини у такого джерела імовірність появи різних букв різна, навіть якщо вибір елементів алфавіту для формування повідомлення здійснюється статистично незалежно.

Математичною моделлю таких повідомлень буде дискретна випадкова послідовність, елементи в якій нерівноімовірні, але вибір даного елементу здійснюється в незалежних від попередніх виборів умовах. Такі послідовності в математиці відомі як послідовності Бернуллі, а джерело, що формує такі послідовності називається **<u>дискретним джерелом без пам'яті</u>**.

Його повний математичний опис створюється на основі одночасно статистичного аналізу безумовної імовірності появи символів алфавіту в повідомленнях $p(x_i)$. Аналогічно дискретні джерела, символи алфавіту яких в повідомленнях з'являються статистично незалежно, але з різною імовірністю, називатимемо джерелами першого типа або джерелами Бернуллі, а величини їх інформаційних заходів відзначати підрядковим індексом «1» (наприклад, I_1, H_1).

Для математичного опису джерел Бернуллі необхідно задати їх алфавіт у вигляді ансамблю імовірності

$$X(x_1, x_2, x_3, \ldots, x_M) = \{a_i\}, i = 1, M;$$
$$P[p(x_1), p(x_2), \ldots, p(x_M)] = \{p(a_i)\}, i = 1, M$$

і вказати довжину повідомлень, сформованих на їх виходах. Зрозуміло, що кількість N_1 різних повідомлень, яка може бути створене джерелом Бернуллі на виході при відомих його параметрах M і n, вже не буде рівна, як у джерел Хартлі M^n, а буде декілька менше.

Ентропія джерела Бернуллі обчислюється по формулі:

$$H_1 = \sum_{i=1}^{M} P(X_i) \log P(X_i)$$

Джерела Маркова – Джерела з нерівноімовірними і статистично залежними символами в повідомленнях. До цього типу дискретних джерел відноситимемо такі джерела, повідомлення на виході яких формується при не рівноімовірній і залежній вибірці елементів алфавіту джерела. Дискретні випадкові послідовності з такими властивостями вивчав російський математик А.А.Марков, його ім'ям і названі джерела з вказаними правилами формування повідомлень.

Дискретне джерело **Маркова** називається **<u>дискретним джерелом з пам'яттю</u>** - це джерело, яке створює послідовності u_L, у яких імовірність формування чергового знаку залежить від того, які знаки були вибрані до цього.

Причини, що приводять до появи неравноймовірності елементів алфавіту в повідомленні, зрозумілі з розгляду джерел Бернуллі. Зупинимося стисло на причинах появи взаємної залежності елементів в повідомленні.

У російській мові, письмовий текст якого є характерним прикладом джерела Маркова, залежність елементів в словах обумовлена обмеженнями, що накладаються російською граматикою на правила формування повідомлень. Так, в граматиці російської мови виключене написання букви Ъ після букви Ь і не допускається написання, наприклад, після букви Ч букв Ы, Я, Ю і т.п.

Обмеження такого вигляду можуть інтерпретуватися як двовимірна зв'язність в послідовності Маркова (або ланцюги, як іноді називають послідовності). Обмеження можуть розповсюджуватися на три і більш символів повідомлень дискретного джерела. У цих випадках математичними моделями повідомлень будуть двохзв'язні, трьохзв'язні і більшої зв'язності ланцюга Маркова. Природно, що при цьому ускладнюються методи визначення інформаційних характеристик повідомлень в порівнянні з джерелами першого і другого типів.

Для математичного опису джерел Маркова необхідно окрім вказівки параметрів М і n повністю задати ансамбль імовірності, як і при описі джерел Бернуллі, і матрицю переходів кожного попереднього стану джерела в подальше:

$$\begin{vmatrix} p(x_1/x_1)p(x_2/x_1)\ldots p(x_M/x_1) \\ p(x_1/x_2)p(x_2/x_2)\ldots p(x_M/x_2) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ p(x_1/x_M)p(x_2/x_M)\ldots p(x_M/x_M) \end{vmatrix} \{x_i\}i=1,M$$

$$\{p(x_i)\} = |p(x_j/x_i)| i-1,M;\ j=1,M.$$

Якщо імовірність переходу $p(x_j/x_i)$ залежить лише від того, яка подія спостерігалася в один з попередніх моментів часу, то послідовність символів і джерело повідомлень, яке вона описує, називатимемо двозв'язковими ланцюгами Маркова, а якщо вона залежить від декількох попередніх подій, то послідовність і джерело називатимемо багатозв'язковими ланцюгами Маркова.

Ентропія двовимірного (двозв'язкового) джерела Маркова обчислюється по формулі:

$$H_2 = \sum_{i=1}^{M}\sum_{j=1}^{M} P(X_i)P({}^{Y_j}\!/\!{}_{X_i})\log P({}^{Y_j}\!/\!{}_{X_i})$$

Дискретні джерела Маркова є найзагальнішими моделями джерел дискретних повідомлень, одновимірні (Бернуллі) і нуль мірні (Хартлі) моделі є їх окремими випадками.

Приклад. Візьмемо повідомлення у вигляді буквеного тексту. Відповідно, елементарними повідомленнями, наступними один за одним з інтервалом дискретності T_0, будуть букви російського алфавіту.

Число М букв, включаючи пропуск між словами “ – ” (число елементарних повідомлень), рівно 32. Джерело, вважатимемо, створює N букв в секунду.

Наближення до реальної статистики буквеного тексту почнемо з найгрубішого нульового наближення. У нульовому наближенні букви вважаються рівноімовірними P(“ – ”)=P(а)=...=P(я)=1/32 і статистично незалежними (імовірність появи даної букви не залежить від того, якими були попередні букви). У цьому наближенні ентропія джерела H_0 виходить максимальною величиною

$$\mathbf{H_0=H_{max}=logM=5} \text{ дв. од. на букву,}$$

або

$$\mathbf{H'_0=H'_{max}=N*5} \text{ дв. од./сек.}$$

У першому наближенні (індекс 1) враховується, що букви мають різну умовну імовірність появи P(“ – ”)=0,145, P(а)=0,087 і т. д., але допущення про незалежність сусідніх букв зберігається. Підрахунок ентропії H_1 в цьому випадку дає

$$\mathbf{H_1=4{,}35} \text{ дв. од. на букву,}$$

або

$$\mathbf{H'_1=4{,}35*N} \text{ дв. од./сек}$$

У російській мові, як і в інших мовах, імовірність появи даної букви в тексті залежить від того, які букви стоять на попередніх місцях. Фахівці з математичної лінгвістики підрахували імовірність появи чергової букви за умови, що попередня буква відома, скориставшись як типовий текст (що розглядається як ергодичний процес) російської мови романом “Війна і мир” Л. Н. Товстого. В результаті ентропія, що доводиться на одну букву, виходить рівною

$$H_2=\sum_{i=1}^{32}\sum_{j=1}^{32}P(X_i)P(Y_j/X_i)\log P(Y_j/X_i)=3{,}52 \text{ дв ед. на бит}$$

12.3. Надмірність джерел повідомлень

Більшість реальних джерел повідомлень володіє надмірністю, яка визначається двома чинниками: відмінністю закону розподілу імовірності появи символів від рівномірного і наявністю зв'язків між ними. Ентропія таких джерел менше за максимальну - кожен символ в середньому несе меншу кількість інформації, ніж він міг би нести.

За визначенням***, надмірністю дискретного М-ічного джерела називається величина***

$$\rho=\frac{H_{max}(X)-H(X)}{H_{max}(X)}=1-\frac{H(X)}{H_{max}(X)}=1-\frac{H(X)}{\log M} \qquad (12.4)$$

Вона характеризує ступінь використовування інформаційної ємності алфавіту джерела: якщо $\rho = 0$, ємність алфавіту використовується повністю, якщо $\rho > 0$, то, у принципі існує інший, більш стислий спосіб представлення повідомлень джерела. Декілька огрублюючи суть справи, можна трактувати ρ як відносну частку букв, необов'язкових для розуміння значення повідомлень.

Велика надмірність повідомлень, з одного боку, ускладнює інформаційний обмін, вимагаючи зайвих витрат енергії і часу на передачу повідомлень. З іншого боку, повідомлення, що володіють малою інформаційною надмірністю, виявляються вельми чутливими до дії перешкод, а це, у свою чергу, утрудняє забезпечення достовірності їх передачі по реальних каналах телекомунікаційних систем.

Саме тому **всі інформаційні перетворення повідомлень і сигналів розділяються на два основні класи**: одні мають на меті зменшити первинну (природну) надмірність повідомлень, щоб підвищити ефективність їх передачі, інші направлені на те, щоб внести додаткову (штучну) надмірність для підвищення достовірності (перешкодостійкості) передаваних повідомлень.

Якщо в результаті кодування надмірність повідомлень зменшується – таке кодування називається ефективним (інакше – кодуванням для джерела), якщо надмірність зростає, кодування називається перешкодостійким (інакше – кодування для каналу). Якщо в результаті кодування надмірність повідомлень зберігається без змін, то кодування називається примітивним.

Надмірність повідомлень джерела приводитиме до надмірності кодових комбінацій, що зрештою приведе до зниження швидкості передачі інформації в каналі. Припустимо, необхідно закодувати 32 букви російського алфавіту. При двійковому коді ($m_k = 2$) їх можна кодувати послідовностями з п'яти двійкових символів, оскільки таких послідовностей теж 32. Якщо вважати алфавіт рівноімовірним, то кожна буква і кожна кодова комбінація переносять 5 біт інформації, а кожен кодовий символ - 1 біт.

Але насправді через нерівноімовірність і залежність кожна буква несе в середньому меншу кількість інформації. Так, ентропія джерела при обліку статистичних зв'язків всього лише між трьома буквами рівна трьом бітам і на частку кожного кодового символу доводиться в середньому 0,6 біт. При цьому швидкість передачі інформації буде менше, ніж пропускна спроможність каналу: символ коду при одній і тій же тривалості $\tau_к$ міг би доставляти в секунду $I/\tau_к$ біт інформації, а доставляє $0{,}6/\tau_к$ битий. Отже, канал використовується неефективно: швидкість передачі інформації в ньому менше його пропускної спроможності. Проте в теорії інформації доводиться, що відповідним вибором способу кодування при будь-якій надмірності джерела повідомлень можна забезпечити швидкість передачі інформації по каналу без перешкод, скільки завгодно близьку до його пропускної спроможності. Подібне кодування називається **ефективним**, а самі коди - **ефективними**.

Для оптимального узгодження джерела з каналом зв'язку без перешкод необхідно, щоб середнє число двійкових символів на букву джерела було не менше ентропії H(A).

Це очевидно, якщо символи джерела незалежні і рівноімовірні. Такі джерела не мають надмірності, а їх ентропія максимальна і рівна $\log M_a$. Якщо M_a цілий ступінь двійки ($M_a = 2n$), то застосувавши рівномірний код на всі поєднання, можна кожен символ джерела закодувати n-значной кодовою комбінацією. Оскільки використовуються всі 2n комбінацій, які так само, як і символи джерела, є рівноімовірними і незалежними, то $n = \log M_a = H(A)$, тобто кількість символів коду на символ джерела рівна ентропії. Наприклад, якщо джерело видає чотири рівноімовірні букви ($H_0 = 2$ біт), то ефективний код має вигляд: a_1 - 00; a_2 - 01; a_3 - 10; a_4 - 11. Число символів коду, що доводяться на одну букву джерела, рівне ентропії і ні при якому іншому способі кодування не може бути меншим. В цьому випадку можна говорити про повне узгодження джерела і каналу.