Практична робота 1

Оцінка ентропії сигналу

Мета роботи: навчитися генерувати імітацію сигналу та оцінювати його ентропію.

1.1 Короткі теоретичні відомості

В теорії інформації інформація розглядається як міра невизначеності системи – під інформацією розуміються всі ті відомості, які зменшують ступінь невизначеності наших знань про конкретний об'єкт або систему (концепція Клода Шеннона – чим імовірнішим є повідомлення, тим менше інформації міститься в ньому).

Цей підхід, незважаючи на те, що він не враховує змістового аспекту інформації, став основою для її кількісної оцінки та оптимального кодування повідомлень.

Інформаційна ентропія (H) – це міра невизначеності або непередбачуваності інформаційного джерела. Вимірюється в бітах або більш загально в натах (натуральний логарифм).

Формула Шеннона для ентропії:

$$H(X) = -\sum_{i=1}^{M} p_i(x) \log_2(p_i(x)) \tag{1.1}$$

де $p_i(x)$ – ймовірність того, що подія x (отримання символу x) відбудеться;

M – можлива кількість станів системи X.

Ентропія показує, яка мінімальна кількість біт на символ (в середньому) потрібна для кодування даного повідомлення.

Хоча, як правило, розрахована ентропія буде дробовим числом, його округлюють до найближчого більшого цілого (тобто якщо розрахували *H(X)* = 7,3 – то реально це буде в 8 біт на символ (при використанні рівномірного коду)).

1.2 Алгоритм оцінки ентропії

Ентропію сигналу (або повідомлення) вручну обчислювати немає сенсу – для реальних сигналів або повідомлень це надзвичайно громіздка операція, тому зазвичай ентропію оцінюють програмним шляхом.

В даних методичних вказівках упор буде робитися на *алгоритмах* обчислень (або цифрової обробки сигналів та/або зображень). Таким чином, ця і подальші практичні роботи робляться за допомогою програмування у якомусь середовищі (Python, R, C++, Java, тощо).

Дана практична робота може складатися з двох частин: спочатку можна оцінити ентропію повідомлення (текстової фрази), а потім органічно перейти до прикладу із сигналом.

Для обчислення ентропії прямим способом по формулі Шеннона (1.1) необхідно виконати ряд дій.

1. Отримати вхідне повідомлення (текстову фразу): це може бути як константа, так і щось введене з клавіатури.
2. Визначення довжини повідомлення N (пробіли та знаки пунктуації вважаються окремими символами).
3. Створення масиву для підрахунку частот появи окремих символів у повідомленні та ймовірності їх появи. Обнулення його (за потреби, в залежності від мови програмування).
4. Визначення відносних частот всіх символів у повідомленні $[n_i]$.
5. Визначення ймовірності появи кожного символу в повідомленні:

$$p_i = \frac{n_i}{N} \tag{1.2}$$

6. Обчислення ентропії повідомлення за формулою (1.1).

На рис. 1.1 показаний цей алгоритм, представлений у вигляді блок-схеми.

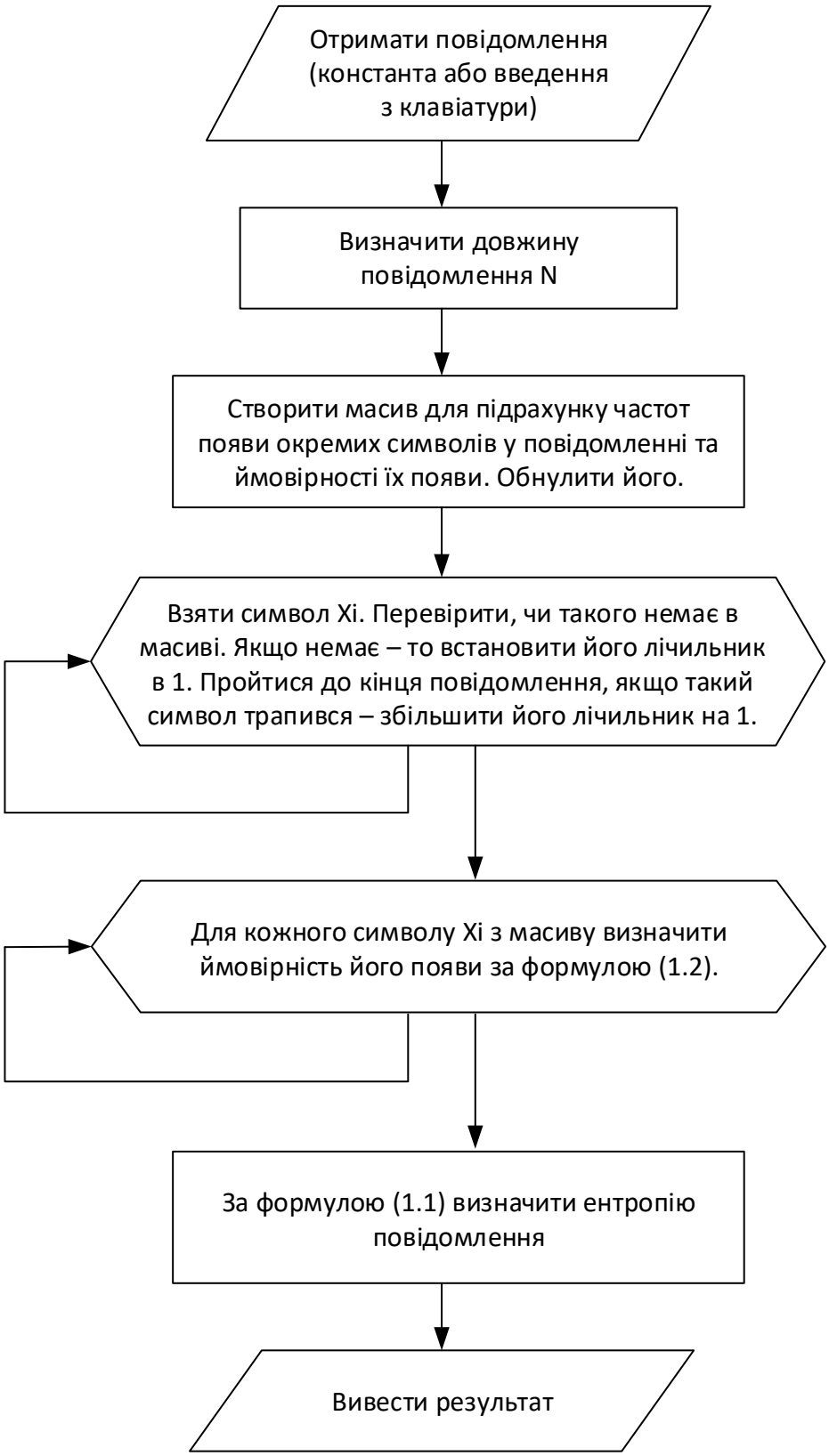

Рис. 1.1 – Блок-схема алгоритму обчислення ентропії повідомлення.

При заміні повідомлення (тексту) на сигнал в масиві будуть зберігатися не символи, а миттєві значення сигналу (відліки) у певні моменти часу. Тоді задача

змінюється по суті лише на першому кроці – замість введення текстової фрази можна протабулювати довільну (але обмежену та неперервну) функцію на певному інтервалі $[x_{min},\ x_{max}]$ N разів. Очевидно, що при цьому крок буде становити

$$\Delta x = \frac{x_{max} - x_{min}}{N}$$

Обчислюється значення функції в усіх точках Δx_i з точністю, наприклад, 2 знака після коми. Потім для отриманої сукупності відліків можна рахувати ентропію.

Важливо: після того, як порахована ентропія для випадку округлення значення відліків до двох знаків після коми, можна це все повторити, але тепер округлювати з точністю до 4 або 5 знаків після коми. В такому випадку значення ентропії повинно бути більшим. Чому?

В якості функції (імітації сигналу) рекомендується брати якусь суму синосоїдальних функцій (ніби це ряд Фур'є). Наприклад:

$$s(t) = 0{,}2\ \sin(\omega t) + 0{,}5\cos(\omega t) + 1{,}3\sin(2\omega t) + 1{,}6\cos(2\omega t) + \cdots$$

Зрозуміло, що коефіцієнти в цьому ряду можуть бути довільними, не обов'язково зростаючими або спадаючими. Можна навіть написати функцію, яка буде генерувати ці коефіцієнти випадковими. Складових гармонік (синусів і косинусів з частотами $k\omega$) може бути скільки завгодно, але краще менше трьох або чотирьох не робити (бо буде надто простий сигнал з малою ентропією).

1.3 Результат практичної роботи

1) Програмний код для обчислення ентропії повідомлення (фрази).

2) Результат тестування цього коду: значення ентропії для кількох різних текстових фраз різної довжини і складності.

3) Програмний код функції для генерування (імітації) сигналу з метою її табуляції на заданому інтервалі.

4) Програмний код для обчислення ентропії сигналу.

5) Результати тестування цього коду: значення ентропії для випадків округлення значень відліків до 2 та більше знаків після коми; результат обчислення ентропії для різних значень коефіцієнтів.

6) Висновки. Чи обов'язково ентропія довших повідомлень буде більшою за ентропію коротких повідомлень? Чи обов'язково ентропія для сигналів, визначених точніше (тобто при обчисленні значень відліків ми округлюємо до більшого числа знаків після коми) буде більшою? Чи залежить якось ентропія від значень коефіцієнтів у ряду Фур'є? Що ще може впливати на значення ентропії? (Наприклад, можна поекспериментувати із значеннями частот тощо)

1.4 Контрольні запитання

1) Що таке ентропія в теорії інформації?

2) Чи відрізняється якось за суттю та ентропія, що обчислюється для текстової фрази, від тієї ентропії, що обчислюється для сигналу?

3) Що показує ентропія?

4) В чому вимірюється ентропія (в теорії інформації)?

5) Від чого може залежати значення ентропії?

6) В якому випадку значення ентропії буде максимально можливим? Мінімально можливим?