УДК 517.5, 519.72

Морозов Д. І.

DOI: https://doi.org/10.18523/2617-70804202120-23

ПОЛІНОМІАЛЬНЕ ПРЕДСТАВЛЕННЯ ДВІЙКОВИХ ДЕРЕВ ЕНТРОПІЙНИХ БІНАРНИХ КОДІВ[1]

Важливою складовою потокового обміну великими об'емами інформації є алгоритми стискання інформаційного потоку, які своєю чергою поділяють на алгоритми стискання без втрат (ентропійні) — Шенона, Хафмана, арифметичне кодування, умовно стискаючі — LZW та інші бієкції інформаційного конусу, алгоритми стискання з втратами, наприклад, mp3, jpeg та низка інших.

Під час побудови алгоритму стискання з втратами важливим є дотримуватись певної формальної стратегії. Сформулювати її можна таким чином: після опису множини об'єктів, які є атомарними елементами обміну в інформаційному потоці, необхідно побудувати абстрактну схему цього опису, що дозволить визначити межу для абстрактних зрізів цієї схеми, за якою починаються допустимі втрати.

Підходи до виявлення абстрактної схеми, що породжує алгоритми стискання з допустимими втратами, можуть бути отримані з контексту предметної області. Наприклад, алгоритм стискання аудіопотоку може розділяти сигнал на прості гармоніки та залишає серед них ті, що розташовані в певному діапазоні сприйняття. Таким чином, отриманий на виході сигнал є певною абстракцією вхідного, що містить важливу інформацію згідно з контекстом слухового сприйняття аудіопотоку та представлений меншою кількістю інформації. Подібний підхід використовується в форматі mp3, який є стискаючим представленням.

На відміну від алгоритмів стискання з втратами, ентропійні алгоритми стискання не вимагають аналізу контексту, а можуть бути побудовані згідно з частотною картиною. Серед відомих алгоритмів побудови таких кодів можна згадати алгоритм Шеннона–Фано, алгоритм Хафмана та арифметичне кодування.

Знаходження інформаційної ентропії для заданого коду Шеннона є тривіальною задачею. Обернена задача, а саме пошук відповідних кодів Шеннона, що мають наперед задану ентропію та з невизначеними ймовірностями, що є від'ємними цілими степенями двійки, є достатньо складною. Вона може бути вирішена прямим перебором, але суттєвим недоліком цього підходу є його обчислювана складність. У цій статті запропоновано альтернативний підхід до пошуку таких кодів. Описана техніка поліноміального представлення двійкових дерев бінарних кодів Шеннона з імовірностями, що є від'ємними цілими степенями двійки, дає змогу будувати відповідні коди за відомим значенням інформаційної ентропії.

Ключові слова: код Шеннона, інформаційна ентропія, бінарне дерево.

Вступ

Бінарний код Шеннона [1] для алфавіту потужності n та з імовірностями, що є від'ємними цілими степенями двійки, може бути заданий двійковим кореневим деревом з n листами. В природній координатизації двійкового кореневого дерева листи, шлях від кореня до яких містить більше ребер, будуть знаходитися лівіше. Для листа кількість ребер, що зв'язують його з коренем, будемо називати рівнем цього листа. При такому представленні двійкове кореневе дерево, що задає даний код Шеннона, буде єдиним з точністю до ізоморфізму. Зауважимо, що поліноміальне представлення алгебраїчних об'єктів у багатьох випадках надає зручну техніку для розв'язання низки проблем [2–4].

Означення 1. За заданим деревом T_n кодом побудуємо відповідний поліном

$$p[T_n](x) = b_1 x + \cdots + b_l x^l,$$

де b_i дорівнює кількості стрічок довжини i у відповідному коді.

Приклад 1. Код Шеннона:

$$0, 1$$

[1]За часткової підтримки проєкту «Створення центру цифрових інновацій НАН України» Цільової програми наукових досліджень НАН України «Математичне моделювання у міждисциплінарних дослідженнях процесів і систем на основі інтелектуальних суперкомп'ютерних, грід- і хмарних технологій» на 2021–2025 рр. (державний реєстраційний номер: 0121U110979).

2 стрічки довжини 1
Ентропія:

$$E[T_2] = 1 \cdot \left(\frac{1}{2}\right)^1 + 1 \cdot \left(\frac{1}{2}\right)^1 = 1.$$

Поліном:

$$p[T_2](x) = 2x.$$

Приклад 2. Код Шеннона:

$$00, 01, 1$$

2 стрічки довжини 2, 1 — довжини 1
Ентропія:

$$E[T_3] = 2 \cdot \left(\frac{1}{2}\right)^2 + 2 \cdot \left(\frac{1}{2}\right)^2 + 1 \cdot \left(\frac{1}{2}\right)^1 = \frac{3}{2}.$$

Поліном:

$$p[T_3](x) = 2x^2 + x.$$

Основна частина

Теорема 1.

$$p[T_n](x) = 2x + (2x^2 - x)k(x).$$

Доведення. Дійсно, кожне дерево коду, що має n листів, можна отримати з дерева, що має $n-1$ листів додаванням двох ребер до певного листа. В результаті зникне один лист рівня t та додадуться два листа рівня $t+1$. В поліноміальному представленні ця операція має такий вигляд:

$$p[T_n](x) = p[T_{n-1}](x) + 2x^{t+1} - x^t =$$

$$= p[T_{n-1}](x) + (2x - 1) \cdot x^t.$$

Для кожного n дерево T_n можна індуктивно отримати з T_2 за допомогою описаної вище операції.

Для дерева T_2 маємо:

$$p[T_2](x) = 2x.$$

Отже, для будь-якого n має місце рівність:

$$p[T_n](x) = 2x + (2x^2 - x)k(x).$$

Лема 2.

$$p[T_n]\left(\frac{1}{2}\right) = 1.$$

Доведення. Згідно з теоремою 1 маємо:

$$p[T_n]\left(\frac{1}{2}\right) = 2 \cdot \frac{1}{2} + 0 \cdot k\left(\frac{1}{2}\right) = 1.$$

Лема 3.

$$p[T_n](1) = n.$$

Доведення. Дійсно, $p[T_n](1)$ дорівнює сумі коефіцієнтів цього полінома, а отже, згідно з означенням 1, дорівнює кількості листів дерева T_n.

Означення 2.

$$e[T_n](x) = x\frac{\partial p[T_n](x)}{\partial x}.$$

Лема 4.

$$e[T_n]\left(\frac{1}{2}\right) = E[T_n].$$

Доведення. Дійсно, для коду, що представлений поліномом $p[T_n](x) = b_1 x + \cdots + b_l x^l$, інформаційна ентропія дорівнює:

$$E[T_n] = b_1 \cdot \frac{1}{2} + ... + b_l \cdot l \cdot \left(\frac{1}{2}\right)^l =$$

$$= \frac{1}{2} \cdot \left(b_1 + ... + b_l \cdot l \cdot \left(\frac{1}{2}\right)^{l-1}\right).$$

А отже має місце рівність:

$$E[T_n] = e[T_n]\left(\frac{1}{2}\right).$$

Означення 3. Означимо породжуючий поліном для коду Шеннона, що заданий деревом T_n, таким чином:

$$k[T_n](x) = \frac{p[T_n] - 2x}{2x^2 - x}.$$

Згідно з теоремою 1 означення 3 є коректним.

Теорема 5. *$k[T_n](x) = a_0 + \cdots + a_t x^t, a_i \in \mathbb{Z}^+$ є породжуючим поліноном для певного бінарного дерева коду Шеннона тоді, і тільки тоді, коли $0 \leq a_0 \leq 2$, та послідовність $a_0, \frac{a_1}{2^1} \ldots, \frac{a_t}{2^t}$ є спадною.*

Доведення. Згідно з розкладом з теореми 1 та означення 3 маємо:

$$p[T_n](x) = 2x + (2x^2 - x) \cdot k[T_n](x).$$

Для існування відповідного дерева коду Шеннона необхідно і достатньо, щоб коефіцієнти результуючого поліному $p[T_n](x)$ були невід'ємними цілими числами. Ця умова відповідає таким обмеженням на коефіцієнти полінома $k[T_n](x)$:

$$a_0 \leqslant 2$$

$$a_{i+1} \leqslant 2a_i$$

$$0 \leqslant a_t$$

що рівносильно тому, що послідовність $a_0, \frac{a_1}{2^1} \ldots, \frac{a_t}{2^t}$ є спадною.

Лема 6.

$$k[T_n](1) = n - 2.$$

Доведення. Згідно з означенням 3:

$$k[T_n](1) = p[T_n](1) - 2.$$

Та, згідно з лемою 3, маємо потрібне твердження:

$$k[T_n](1) = n - 2.$$

Означення 4.

$$\widetilde{E}[T_n] = 2(E[T_n] - 1).$$

Теорема 7.

$$k[T_n]\left(\frac{1}{2}\right) = \widetilde{E}[T_n].$$

Доведення. Мають місце такі співвідношення:

$$(2x + x(2x-1)k(x))' = 2 + (x(2x-1))'k(x) +$$

$$+x(2x-1)k'(x) = 2+(4x-1)k(x)+(2x^2-x)k'(x).$$

Отже, згідно з означенням 3:

$$e[T_n](x) = x \cdot \frac{\partial p[T_n](x)}{\partial x} =$$

$$= x\cdot\left(2 + (4x-1)k[T_n](x) + (2x^2 - x) \cdot \frac{\partial k[T_n](x)}{\partial x}\right).$$

Тому, згідно з лемою 4:

$$E[T_n] = e[T_n]\left(\frac{1}{2}\right) = \frac{1}{2} \cdot \left(2 + k[T_n]\left(\frac{1}{2}\right)\right).$$

Остаточно маємо:

$$k[T_n]\left(\frac{1}{2}\right) = 2(E[T_n] - 1) = \widetilde{E}[T_n].$$

Приклад 3. Нехай ентропія коду Шеннона з деревом T_n дорівнює $\frac{3}{2}$.

$$\widetilde{E}[T_n] = 2(\frac{3}{2} - 1) = 1.$$

Нехай

$$k[T_n](x) = a_0 + a_1x + ...a_tx^t.$$

Отже, за теоремою 7

$$a_0 + \frac{a_1}{2} + ...\frac{a_t}{2^t} = 1$$

і єдиний розв'язок в умовах теореми 5 — $a_0 = 1, a_i = 0, i > 0$. Звідси

$$k[T_n](x) = 1$$

$$p[T_n](x) = 2x + (2x^2 - x) = 2x^2 + x.$$

Єдиним двійковим кодом Шеннона з ентропією $\frac{3}{2}$ буде такий:

$$00, 01, 1$$

2 стрічки довжини 2, 1 — довжини 1.

Приклад 4. Нехай ентропія коду Шеннона з деревом T_n дорівнює 2.

$$\widetilde{E}[T_n] = 2(2 - 1) = 2.$$

Нехай

$$k[T_n](x) = a_0 + a_1x + \cdots + a_tx^t.$$

Отже

$$a_0 + \frac{a_1}{2} + \cdots + \frac{a_t}{2^t} = 2$$

і єдиний розв'язок для кожного n в умовах теореми для T_n — це

$$a_i = 1, i \leqslant (n-5)$$

$$a_{n-4} = 2.$$

Звідси

$$k[T_n](x) = 1 + x + \cdots + 2x^{n-2}$$

$$p[T_n](x) = 2x + (2x^2 - x)(1 + x + \cdots + 2x^{n-2}).$$

Наприклад

$$p[T_4](x) = 2x + (2x^2 - x)2 = 4x^2,$$

тому єдиним двійковим кодом Шеннона з ентропією 2 та деревом T_4 буде такий:

$$00, 01, 10, 11,$$

що містить 4 стрічки довжини 2.

Для $n = 5$ маємо:

$$p[T_5](x) = 2x + (2x^2 - x)(1 + 2x) = 4x^3 + x,$$

тому єдиним двійковим кодом Шеннона з ентропією 2 та деревом T_5 буде такий:

$$000, 001, 010, 011, 1,$$

що містить 4 стрічки довжини 3 та одну стрічку довжини 1.

Висновки

Описана техніка поліноміального представлення двійкових дерев бінарних кодів Шеннона з ймовірностями, що є від'ємними цілими степенями двійки, дозволяє будувати відповідні коди за відомим значенням інформаційної ентропії.

Список літератури

1. Shannon C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* 1948. Vol. 27, no. 3. Pp. 379–423.
2. Морозов Д. І. Ізометричність поліномів над кільцем цілих 2-адичних чисел. *Наукові записки НаУКМА.* 2011. Т. 113: Фізико-математичні науки. С. 13–15.
3. Морозов Д. І. Спряженість автоморфізмів, що задаються лінійними функціями в групі скінченностанових автоморфізмів кореневого сферично-однорідного дерева. *Вісник Київського ун-ту. Серія: Фізико-математичні науки.* 2008. Вип. 1. С. 40–43.
4. Blahut R. E. Theory and practice of error control codes. Reading, MA: Addison-Wesley Pub. Co., 1983.

References

1. Claude E. Shannon, "A mathematical theory of communication", Bell Syst. Tech. J. **27** (3), 379–423 (1948).
2. Д. І. Морозов, «Ізометричність поліномів над кільцем цілих 2-адичних чисел», Наукові записки НаУКМА. Серія: Фізико-математичні науки. **113**, 13–15 (2011).
3. Д. І. Морозов, «Спряженість автоморфізмів, що задаються лінійними функціями в групі скінченностанових автоморфізмів кореневого сферично-однорідного дерева», Вісник Київського ун-ту. Серія: Фізико-математичні науки. **1**, 40–43 (2008).
4. R. E. Blahut, *Theory and practice of error control codes* (Addison-Wesley Pub. Co.: Reading, MA, 1983).

D. Morozov

POLYNOMIAL REPRESENTATION OF BINARY TREES OF ENTROPY BINARY CODES

An important component of streaming large amounts of information are algorithms for compressing information flow. Which in turn are divided into lossless compression algorithms (entropic) - Shannon, Huffman, arithmetic coding, conditional compression - LZW, and other information cone injections and lossy compression algorithms - such as mp3, jpeg and others.

It is important to follow a formal strategy when building a lossy compression algorithm. It can be formulated as follows. After describing the set of objects that are atomic elements of exchange in the information flow, it is necessary to build an abstract scheme of this description, which will determine the boundary for abstract sections of this scheme, which begins the allowable losses.

Approaches to the detection of an abstract scheme that generates compression algorithms with allowable losses can be obtained from the context of the subject area. For example, an audio stream compression algorithm can divide a signal into simple harmonics and leave among them those that are within a certain range of perception. Thus, the output signal is a certain abstraction of the input, which contains important information in accordance with the context of auditory perception of the audio stream and is represented by less information. A similar approach is used in the mp3 format, which is a compressed representation.

Unlike lossy compression algorithms, entropic compression algorithms do not require context analysis, but can be built according to the frequency picture. Among the known algorithms for constructing such codes are the Shannon-Fano algorithm, the Huffman algorithm and arithmetic coding.

Finding the information entropy for a given Shannon code is a trivial task. The inverse problem, namely finding the appropriate Shannon codes that have a predetermined entropy and with probabilities that are negative integer powers of two, is quite complex. It can be solved by direct search, but a significant disadvantage of this approach is its computational complexity. This article offers an alternative technique for finding such codes.

Keywords: Shannon code, information entropy, binary tree.

Матеріал надійшов 18.10.2021