Оптимальное универсальное кодирование для объединения различных множеств источников символами неравной длительности¹

В.К. Трофимов, Т.В. Храмова

Предложен метод универсального кодирования для множества источников, состоящего из объединения счётного числа множеств. Показано, что предложенное кодирование объединения марковских источников с конечной памятью является асимптотически оптимальным. Кроме того, установлено, что предложенное кодирование — слабо универсальное для множества стационарных источников.

Ключевые слова: кодирование, избыточность, энтропия, хранение и обработка информации, источник сообщений.

1. Введение

Проблема сжатия (кодирования) информации [1] относится к фундаментальным в области инфокоммуникаций. В книге В. Г. Хорошевского [2] отмечено, что решение этих проблем значимо и при создании большемасштабных распределённых вычислительных систем. В таких системах метода сжатия информации, как правило, используют параллельные информационно-вычислительные технологии.

В настоящей работе исследуется вопрос о кодировании информации, порождённой неизвестным источником из объединения некоторого числа источников буквами выходного (кодированного) алфавита, имеющими различные длительности. Для букв, имеющих равные длительности, эта задача впервые рассмотрена в [3]. Задача кодирования неравнозначными буквами последовательности, порождённой известным источником, впервые рассмотрена в [1]. В дальнейшем вопросы кодирования сообщений, порождённых известным источником, неравнозначными символами рассматривались в [4–6].

Кодирование сообщений, порождённых неизвестным источником, буквами алфавита с равнозначными длительностями символов впервые рассмотрено в [7]. Кодирование сообщений, порождённых неизвестным источником, получило название *универсального кодирования*. Точная постановка задачи универсального кодирования сделана в [8]. Универсальное кодирование равнозначными символами для различных множеств источников и для различных типов кодирований интенсивно изучалось как у нас в стране, так и за рубежом. Кроме упомянутых работ отметим [9–14]. Наиболее полная библиография по этому вопросу содержится в [9, 15, 16]. В частности, в [8] предложен метод универсального кодирования бернуллиевских источников и доказана его асимптотическая оптимальность. В [13] предложен метод универсального кодирования для множества источников с конечной памятью, а в [11] доказана его асимптотическая оптимальность. В [17, 18] авторами предложен метод универсального кодирования бернуллиевских и марковских источников неравнозначными симво-

¹ Работа выполнена при поддержке Российского фонда фундаментальных исследований (грант № 12-07-00188).

лами и доказана их асимптотическая оптимальность. Как и в случае равнозначности длительности букв кодового алфавита, платой за незнание источника является множитель, пропорциональный $\log N$ (здесь и далее $\log x = \log_2 x$, $0 \log 0 = 0$, N — длина кодируемого блока).

В настоящей работе предложен метод кодирования средними для объединения различных множеств источников. Показано, что предложенное кодирование является слабо универсальным для множеств стационарных источников.

2. Основные определения. Постановка задачи

Изучение методов кодирования информации с целью уменьшения занимаемого ею объёма является актуальной задачей информатики. Данная работа посвящена изучению оптимальных методов сжатия информации, порождённой неизвестным дискретным стационарным источником, буквами выходного алфавита, которые имеют различные стоимости.

Рассмотрим источник сообщений Θ , генерирующий последовательность из букв некоторого конечного алфавита $X = \{x_1, x_2, \dots, x_k\}$.

Последовательность должна быть закодирована и передана по каналу связи. Для решения нашей задачи оптимального кодирования разобьём исходящую последовательность букв источника на блоки (слова) фиксированной длины N . Процедура кодирования источника заключается в том, что каждому блоку $w \in X^N$ ставится в соответствие некоторое кодовое слово $\varphi(w) \in Y^*$ из букв кодового алфавита $Y = \{y_1, y_2, \dots, y_m\}$ (как обычно Y^* обозначает множество всевозможных последовательностей из элементов множества Y). Кодирование, при котором блокам источника ставится в соответствие кодовое слов нефиксированной длины, называется, *равномерным по входу*. Разные буквы кодового алфавита имеют разную *длительность* или, другими словами, *стоимость передачи*: $t_j = t(y_j)$, $j = \overline{1, m}$. Таким образом, каждому кодовому алфавиту можно поставить в соответствие *вектор длительностей букв* $\bar{t}_Y = (t_1, t_2, \dots, t_m)$, ($t_j = t(y_j)$, $j = \overline{1, m}$).

В частности, длительности кодовых символов могут быть одинаковы, в этом случае соответствующий алфавиту вектор обозначим $\bar{t}_1 = \underbrace{(1, 1, \dots, 1)}_m$.

Самым популярным примером кода с неравными длительностями является код Морзе, актуальность которого не теряется и в наше время — на основе кода Морзе созданы широко используемые штрих-коды.

Длительностью кодового слова будем считать величину, равную сумме длительностей входящих в слово букв:

$$l(\varphi(w), \bar{t}_Y) = \sum_{y \in \varphi(w)} t(y).$$

Стоимость кодирования определяется как отношение средней длительности кодового слова к средней длительности слова источника и в рассматриваемом случае принимает вид:

$$L(N, \Theta, \varphi, \bar{t}_Y) = \frac{1}{N} \sum_{w \in X^N} p_\Theta(w) l(\varphi(w), \bar{t}_Y),$$

где $p_\Theta(w)$ — вероятность слова w , порождённого источником Θ . Эффективность кодирования φ определяется его избыточностью $R(N, \Theta, \varphi, \bar{t}_Y)$, определяемой равенством

$$R(N, \Theta, \varphi, \bar{t}_Y) = L(N, \Theta, \varphi, \bar{t}_Y) - H(\Theta)/C(\bar{t}_Y).$$

Здесь $H(\Theta)$ — энтропия источника Θ , определяемая обычным образом [19, 20], $C(\bar{t}_Y)$ — пропускная способность канала без шума [1, 6], вычисляемая по формуле

$$C(\bar{t}_Y) = \log(\omega_0(\bar{t}_Y)),$$

где $\omega_0(\bar{t}_Y)$ — наибольший положительный корень уравнения $\omega^{-t_1} + \omega^{-t_2} + \dots + \omega^{-t_m} = 1$.

Избыточность универсального кодирования $R(N, \Omega, \bar{t}_Y)$ множества источников Ω определяется равенством

$$R(N, \Omega, \bar{t}_Y) = \inf_{\varphi} \sup_{\Theta \in \Omega} R(N, \Theta, \varphi, \bar{t}_Y).$$

3. Адаптивное кодирование для объединения различных множеств источников

Сформулируем и докажем утверждение, которое позволит строить коды, адаптивные по множеству классов источников.

Лемма. Если множество источников $W = \bigcup_{k=0}^{\infty} W_k$ и для каждого из множеств W_k известно оптимальное кодирование $\varphi^{(k)}$, $k=1,2,\dots$, то существует такое кодирование φ , что для любого источника $\Theta \in W_k$ выполняется неравенство

$$R(N, \Theta, \varphi, \bar{t}_Y) \leq R(N, \Theta, \bar{t}_Y) + \frac{2 \log k}{N} + \frac{T^*}{N},$$

где T^* — постоянная, не зависящая от N .

Доказательство. Кодирование $\varphi^{(k)}$ дешифруемо, поэтому для чисел $l(\varphi^{(k)}(w), \bar{t}_Y)$ выполняется неравенство Макмиллана

$$\sum_{w \in X^N} 2^{-l(\varphi^{(k)}(w), \bar{t}_Y) \log \omega_0(\bar{t}_Y)} \leq 1, \quad k = 1, 2, \dots$$

Рассмотрим кодирование φ , которое каждому слову w , $w \in X^N$, ставит в соответствие слово $\varphi(w)$ длительности $l(\varphi(w), \bar{t}_Y)$:

$$l(\varphi(w), \bar{t}_Y) = \left\lceil -\log \lambda \sum_{k=0}^{\infty} \frac{1}{(k+1)^2} 2^{-l(\varphi^{(k)}(w), \bar{t}_Y) \log \omega_0(\bar{t}_Y)} \right\rceil,$$

здесь λ — нормирующий множитель.

Из определения $l(\varphi(w), \bar{t}_Y)$ имеем $\sum_{w \in X^N} 2^{-l(\varphi(w), \bar{t}_Y) \log \omega_0(\bar{t}_Y)} \leq 1$, значит, как следует

из алгоритма описанного в [18], можно построить кодирование с длительностями кодовых слов $\tilde{l}(\varphi(w), \bar{t}_Y)$, отличающиеся от $l(\varphi(w), \bar{t}_Y)$ на равномерно ограниченную постоянную T^* , т.е. для любого слова w , $w \in X^N$ выполняется неравенство

$$l(\varphi(w), \bar{t}_Y) \leq \tilde{l}(\varphi(w), \bar{t}_Y) + T^*.$$

Покажем, что кодирование φ является искомым для источника Θ :

$$\begin{aligned} R(N, \Theta, \varphi, \bar{t}_Y) &= \frac{1}{N} \sum_{w \in X^N} p_\Theta(w) \cdot l(\varphi(w), \bar{t}_Y) - \frac{H(\Theta)}{C(\bar{t}_Y)} \leq \\ &\leq \frac{1}{N} \sum_{w \in X^N} p_\Theta(w) \cdot \left(-\log \lambda \sum_{k=0}^{\infty} \frac{1}{(k+1)^2} 2^{-l(\varphi^{(k)}(w), \bar{t}_Y) C(\bar{t}_Y)} \right) - \frac{H(\Theta)}{C(\bar{t}_Y)} + \frac{T^* + C_1}{N} = \\ &= \frac{1}{N} \sum_{w \in X^N} p_\Theta(w) \cdot l(\varphi^{(k_0)}(w_0), \bar{t}_Y) - \frac{H(\Theta)}{C(\bar{t}_Y)} + \\ &+ \frac{1}{N} \sum_{w \in X^N} p_\Theta(w) \cdot \left(-\log \left(1 + \sum_{k \neq k_0} \frac{(k_0+1)^2}{(k+1)^2} \cdot 2^{-l(\varphi^{(k)}(w), \bar{t}_Y) + l(\varphi^{(k_0)}(w_0), \bar{t}_Y)} \right) \right) + \\ &+ \frac{T^* + C_1 + \log \lambda + 2 \log(k_0+1)}{N}. \end{aligned}$$

Так как при любых $w \in X^N$ выполняется неравенство

$$-\log \left(1 + \sum_{k \neq k_0} \frac{(k_0+1)^2}{(k+1)^2} \cdot 2^{-l(\varphi^{(k)}(w), \bar{t}_Y) + l(\varphi^{(k_0)}(w_0), \bar{t}_Y)} \right) \leq 0,$$

то из предыдущего получаем:

$$R(N, \Theta, \varphi, \bar{t}_Y) \leq R(N, \Theta, \varphi^{(k_0)}, \bar{t}_Y) + \frac{2 \log(k_0+1)}{N} + \frac{T^*}{N}. \quad (1)$$

Из универсальной оптимальности $\varphi^{(k_0)}$ для множества источников W_{k_0} из (1) имеем

$$R(N, \Theta, \varphi, \bar{t}_Y) \leq R(N, \Theta, \varphi^{(k_0)}, \bar{t}_Y) + \frac{2 \log(k_0+1)}{N} + \frac{T^{***}}{N}, \quad (2)$$

где T^{***} не зависит ни от N , ни от k , ни от k_0 . Из (2) и определения $R(N, W_{k_0}, \bar{t}_Y)$ следует утверждение леммы. $\square$

Следует отметить, что предложенное кодирование φ не требует перестройки в зависимости от выбранного источника: оно само автоматически подстраивается по выбранному источнику.

4. Универсальное кодирование для объединения множеств марковских источников

Пусть $\Omega = \bigcup_k \Omega_k$, где Ω_k — марковские источники с конечной памятью $s(k)$, $k=1, 2, \dots$.

Имеет место следующее утверждение.

Теорема. Для множества марковских источников $\Omega = \bigcup_k \Omega_k$ существует такое кодирование φ , что для любого источника $\Theta \in \Omega_{k_0}$, k_0 — произвольное, выполняется асимптотическое неравенство

$$R(N, \Theta, \varphi_Y) \leq \frac{k^{s(k_0)}(k-1)}{2N} \cdot \log N + \frac{2 \log k_0 + C}{N}.$$

Доказательство. В работе [17] построен оптимальный метод кодирования марковских источников с конечной памятью. Этот метод кодирования позволяет получить для марковских источников связанности s оценку вида:

$$R(N, \Omega_s, \varphi_Y) \leq \frac{k^s(k-1)}{2NC(\bar{t}_Y)} \cdot \log N + \frac{T}{N}, \quad (3)$$

которую асимптотически нельзя улучшить. Пусть φ^s — кодирование, оптимальное для множества марковских источников связанности s . В [18] такое кодирование построено. Каждому слову u оно ставит в соответствие слово $\varphi^s(u)$, имеющее длительность $|\varphi^s(u)|$:

$$|\varphi^s(u)| \leq -\log \bar{p}^s(u) + T^*,$$

где $\bar{p}^s(u)$ — средняя вероятность слова u по множеству источников Ω_s , при условии, что на Ω_s задано КТ-распределение [15]. Для множества источников Ω определена средняя вероятность слова u по формуле

$$\bar{p}(u) = \frac{6}{\pi^2} \sum_{k=1}^{\infty} \frac{1}{k^2} \bar{p}_{s(k)}(u). \quad (4)$$

Очевидно, что если объединение состоит из бесконечного числа источников, то $\sum_{u \in X^N} \bar{p}(u) = 1$, в противном случае, $\sum_{u \in X^N} \bar{p}(u) \leq 1$. Построим разделимое кодирование, как это сделано в [18], используя распределение (4), тогда кодирование φ имеет для слова u длительность

$$|\varphi(u)| \leq -\log \bar{p}(u) + T^{**}. \quad (5)$$

В (5) T^{**} — равномерно ограниченная постоянная. Оценим для построенного кодирования $R(N, \Theta, \varphi, \bar{t}_Y)$. Пусть $\Theta \in \Omega_{s(k_0)}$ по определению имеем:

$$R(N, \Theta, \varphi, \bar{t}_Y) = \left(\frac{1}{N} \sum_{u \in X^N} p_{\Theta}(u) \cdot (-\log \bar{p}_{s(k_0)}(u)) - \frac{H(\Theta)}{\log \omega_0(\bar{t}_Y)} \right) + \frac{T^{**}}{N}. \quad (6)$$

Согласно (4), (5) из (6) следует

$$\begin{aligned} R(N, \Theta, \varphi, \bar{t}_Y) &= \frac{1}{N} \sum_{u \in X^N} p_{\Theta}(u) \cdot (-\log \bar{p}_{s(k_0)}(u)) - \frac{H(\Theta)}{\log \omega_0(\bar{t}_Y)} + \\ &+ \left(-\frac{1}{N} \sum_{u \in X^N} p_{\Theta}(u) \cdot \log \left(1 + \sum_{k \neq k_0} \frac{k_0^2}{k^2} \frac{\bar{p}_{s(k)}(u)}{\bar{p}_{s(k_0)}(u)} \right) \right) + \frac{T^{***}}{N}. \end{aligned} \quad (7)$$

Так как

$$\frac{1}{N} \sum_{u \in X^N} p_{\Theta}(u) \cdot (-\log \bar{p}_{s(k_0)}(u)) - \frac{H(\Theta)}{\log \omega_0(\bar{t}_Y)} = R(N, \Theta, \varphi_{s(k_0)}, \bar{t}_Y) + \frac{T}{N} < R(N, \Omega_{s(k_0)}, \bar{t}_Y) + \frac{T}{N}$$

(T — равномерно ограниченная постоянная) и

$$-\frac{1}{N} \sum_{u \in X^N} p_{\Theta}(u) \cdot \log \left(1 + \sum_{k \neq k_0} \frac{k_0^2}{k^2} \frac{\bar{p}_{s(k)}(u)}{\bar{p}_{s(k_0)}(u)} \right) \leq 0,$$

то из (7) получаем: $R(N, \Theta, \varphi, \bar{t}_Y) \leq R(N, \Omega_{s(k_0)}, \bar{t}_Y) + \frac{T}{N}$.

Отсюда и из (3) имеем $R(N, \Theta, \varphi, \bar{t}_Y) \leq \frac{k^s(k-1)}{2NC(\bar{t}_Y)} + \frac{T}{N}$. Теорема доказана. $\square$

5. Заключение

Предложенное выше кодирование целесообразно применять в том случае, когда избыточность кодирования одного класса источников существенно отличается от избыточности другого класса источников.

Например, известно, что кодируемый источник может быть либо бернуллиевским, либо марковским связанности $s = 20$. Если кодировать все источники как марковские источники с памятью $s = 20$, то в [19] показано, что избыточность такого кодирования будет асимптотически равна $\frac{k^{20}(k-1)}{2NC(\bar{t}_Y)} \cdot \log N$, а кодирование, предложенное в настоящей работе, имеет эту

избыточность только для марковских источников с памятью $s = 20$, для бернуллиевских источников это же самое кодирование будет давать избыточность $\frac{k-1}{2NC(\bar{t}_Y)} \cdot \log \frac{N+L}{L+1}$, что

в k^{20} раз меньше.

Литература

1. Шеннон К. Математическая теория связи. Работы по теории информации и кибернетике. — 1969. — Ил., М. — С.243 - 332.
2. Хорошевский В.Г. Архитектура вычислительных систем: Учеб. пособие. - 2-е изд., перераб. и доп. - М.: Изд-во МГТУ им. Н.Э. Баумана, 2008. - 520 с.
3. Рябко Б. Я. Дважды универсальное кодирование.// Проблемы передачи информации. — 1984. — Т. 20, № 3. — С. 24-28.
4. Csiszar I. Simple proof of some theorems on noiseless channels// Information and control. — 1969. — V.14. — P.285 - 298.
5. Чусар И. О каналах без шума. // Пробл. передачи информ. — 1970. — Т.6. № 4. — С.3-15.
6. Katona G. General theory of noiseless channels. // UDINE, Courses and lectures — 1970, № 31. — P.69.
7. Фитингоф Б. М. Оптимальное кодирование при неизвестной и меняющейся статистике сообщений.// Пробл. передачи инф. — 1966 — т.2, № 2. с. 3-11.
8. Кричевский Р. Е. Связь между избыточностью кодирования и достоверностью сведений об источнике // Проблемы передачи информации. — 1968, Т.4, № 3. — С.48-57.

9. *Потанов В. Н.* Обзор методов неискажающего кодирования дискретных источников.// Дискр. анализ и иссл. операций. Сер.1 – 1999 – Т.6, № 4. – С. 49-91.
10. *Рябко Б. Я.* Кодирование источника с неизвестными, но упорядоченными вероятностями.// Проблемы передачи информации. — 1979. — Т. 15, № 2. — С. 71-77.
11. *Трофимов В. К.* Избыточность универсального кодирования произвольных марковских источников // Пробл. передачи инф. – 1974. – Т.10, № 4. – С.16-24.
12. *Devisson L. D.* Universal noiseless coding // IEEE Trans. Inform. Theory. – 1973. – V.19, № 6. – P.783-795.
13. *Штарьков Ю. М.* Кодирование сообщений конечной длины на выходе источника с неизвестной статистикой// Материалы V конференции по теории кодирования и передачи информации. – Москва–Горький. – 1972. – ч. 1. – С. 147–152.
14. *Elias P.* Universal codeword sets and representation of the integers // IEEE Trans. Inform. Theory. – 1975. – V.21, № 2. – P.194-203.
15. *Krichevsky R. E., Trofimov V. K.* The performance of universal encoding // IEEE Transactions on Information Theory. – 1981. – V. 27, № 2. – P. 199-207.
16. *Verdu S.* Fifty Years of Shannon Theory.// IEEE Transactions on Information Theory. – 1998. – V. 44, № 6. – P.2057-2078.
17. *Трофимов В. К., Храмова Т.В.* Сжатие неравнозначными символами информации, порождённой неизвестным источником без памяти. // Автометрия. – Новосибирск. – 2012. — Т.48, №1. – С.30— 44.
18. *Трофимов В. К., Храмова Т. В.* Универсальное кодирование марковских источников неравнозначными символами.// Дискретный анализ и исследование операций. – Новосибирск. – май—июнь 2013. – Т.20, № 3. — С. 71–83.
19. *Галлагер Р. Г.* Теория информации и надёжная связь. М.:Советское радио. – 1974. – С. 720.
20. *Фано Р.* Передача информации. Статистическая теория связи. – М.: «Мир» – 1965. – С. 440.

Статья поступила в редакцию 01.09.2014

Трофимов Виктор Куприянович

д.т.н., профессор, заведующий кафедрой высшей математики ФГОБУ ВПО СибГУТИ, декан факультета информатики и вычислительной техники ФГОБУ ВПО СибГУТИ (630102, г. Новосибирск, ул. Кирова, д. 86), тел.: (383) 269-82-70, e-mail: trofimov@sibsutis.ru

Храмова Татьяна Викторовна

к.т.н., доцент кафедры высшей математики ФГОБУ ВПО СибГУТИ (630102, г. Новосибирск, ул. Кирова, д. 86), тел.: +79139294100, e-mail: tvkhramova@gmail.com

Optimal universal coding sets for combining various sets of sources by symbols of unequal duration

V.K. Trofimov, T.V. Khramova

The method of universal coding for sets of sources, consisting of the union of a countable number of sets is proposed. It is shown that the proposed encoding of Markov's sources combination with finite memory is asymptotically optimal. Furthermore, it was found that the proposed coding is weakly universal for stationary sources.

Keywords: coding, redundancy, entropy, storage and processing of information, the source of messages.