IEEE Trans. Inf. Th. – 1981, v. 27, № 2. – P.199-207.

4. 4. Кнут Д. Искусство программирования для ЭВМ, сортировка и поиск [Текст] / Д. Кнут. - М.: Мир. – 1978.

Научный сотрудник, к.ф.-м.н. ***Бакулина М.П.*** *тел. 8-961-215-79-36, marina@rav.sscc.ru - Институт Вычислительной Математики и Математической Геофизики СО РАН .*

УДК 621.391

КОДИРОВАНИЕ НЕИЗВЕСТНОГО СТАЦИОНАРНОГО ИСТОЧНИКА СИМВОЛАМИ НЕРАВНОЙ ДЛИТЕЛЬНОСТИ

В.К. Трофимов, Т.В. Храмова

Предложен метод слабо универсального кодирования множества стационарных источников при условии, что символы кодового алфавита имеют неравные длительности.

Ключевые слова: кодирование, избыточность кодирования, энтропия

Изучение методов кодирования информации, с целью уменьшения занимаемого ею объема является актуальной задачей информатики. Данная работа посвящена изучению оптимальных методов сжатия информации, порожденной неизвестным дискретным стационарным источником.

Рассмотрим источник сообщений Θ, генерирующий последовательность из букв некоторого конечного алфавита

$$X = \{x_1, x_2, ..., x_k\}.$$

Последовательность должна быть закодирована и передана по каналу связи. Для решения нашей задачи оптимального кодирования, разобьем 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

$$Y = \{y_1, y_2, ..., y_m\}$$

(здесь 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* $\bar{t} = (t_1, t_2, \cdots, t_m)$, $(t_j = t(y_j),\ j = \overline{1,m})$.

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

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:

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

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

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

Эффективность метода кодирования $\varphi : X^N \mapsto Y$ определяется *избыточностью*

$$R(N, \Theta, \varphi, \bar{t}) = L(N, \Theta, \varphi, \bar{t}) - H(\Theta)/C(\bar{t}), \quad (2)$$

где $H(\Theta)$ — *энтропия источника*, определяемая законом распределения вероятностей появления букв алфавита X на выходе источника Θ, $C(\bar{t})$ — *пропускная способность* канала передачи информации, зависящая только от кодового алфавита Y.

Изучению эффективных методов кодирования посвящено множество работ. В случае известного источника, при равных

длительностях букв кодового алфавита, оптимальным является метод кодирования Хаффмена [1], а эффективный метод кодирования при неравнозначных длительностях кодовых символов предложен в работах Г. Катоны [2].

В данной работе мы решаем задачу кодирования в предположении, что статистика источника неизвестна, т.е. речь идет об *универсальном кодировании*. В работе Р.Е. Кричевского [3] дана постановка задачи универсального кодирования, как кодирования, на котором достигается наименьшее значение избыточности для наихудшего источника. Избыточность универсального кодирования $R(N,\Omega,\varphi_0,\bar{t})$ множества источников Ω определяется равенством

$$R(N,\Omega,\varphi_0,\bar{t})=\inf_{\varphi}\sup_{\Theta\in\Omega}R(N,\Theta,\varphi,\bar{t}). \quad (3)$$

Асимптотическая оценка избыточности универсального кодирования $R(N,\Omega_0,\varphi_0,\bar{t}_1)$ для множества бернуллиевских источников Ω_0 при равных длительностях кодовых символов была получена Р.Е. Кричевским [4], а для множества марковских источников порядка s Ω_s, оценка избыточности $R(N,\Omega_s,\varphi_0,\bar{t}_1)$ получена В.К. Трофимовым и Ю.М. Штарьковым [5,6]. Универсальное кодирование множества стационарных источников Ω_∞ при равных длительностях символов кодового алфавита изучалось в работе Ю.М. Штарькова и В.Ф. Бабкина [7]. В работах авторов данной статьи [8—10] получены асимптотические оценки избыточности универсального равномерного по входу кодирования множества бернуллиевских источников Ω_0 и марковских источников Ω_s.

Рассмотрим множество всех стационарных источников Ω_∞. Если с ростом длины кодируемого блока избыточность универсального кодирования (3), сходится к 0 равномерно по $\Theta\in\Omega$, то кодирование называется *сильно универсальным* на множестве Ω, а если сходимость не равномерная, то кодирование называется *слабо универсальным* на множестве Ω.

В данной работе доказано существование равномерного по входу слабо универсального кодирования стационарного источника для случая кодового алфавита с символами различных длительностей, что является обобщением результатов, полученных в работе Ю.М. Штарькова и В.Ф.Бабкина [7]. В настоящей работе доказано, что избыточность предлагаемого метода кодирования может быть сколь угодно мала.

Энтропия $H(\Theta_\infty)$ стационарного источника вычисляется по формуле [12]

$$H(\Theta_\infty)=\lim_{s\to\infty}H(\Theta_s), \quad (4)$$

где $H(\Theta_s)$ — энтропия марковского источника порядка s (источника, для которого вероятность появления любой буквы в сообщении является условной и зависит от s предыдущих).

Для избыточности марковских источников в работе [11] был получен следующий асимптотический результат:

$$R(N,\Omega_s,\bar{t})\cong\frac{(k-1)k^s}{2C(\bar{t})}\cdot\frac{\log(N-s)}{N}, \quad (5)$$

(здесь и далее $\log x=\log_2 x$).

Для стоимости предложенного в работе [11] метода кодирования φ_0^s имеет место верхняя оценка:

$$L(N,\Theta,\varphi_0^s,\bar{t})\le\frac{H(\Theta)}{C(\bar{t})}+\frac{k^s(k-1)}{2}\frac{\log(N-s)}{C(\bar{t})N}+$$
$$+\frac{k^s(k-1)}{2}\frac{T^*(k,s)+\log e(1+\alpha(N,k,s))}{C(\bar{t})N}+\frac{t^{**}}{N}, \quad (6)$$

где t^{**} — максимальная длительность кодового символа, и величины $T^*(k,s)$ и $\alpha(N,k,s)$ не зависят от длины кодируемого блока и определяются равенствами

$$T^*(k,s)=2\log k/k^s(k-1)+1/k^s+1/(k-1)-$$
$$-(1-1/k^s)\log(\pi e)-(\log(k-1))/k^s, \quad (7)$$

$$\alpha(N,k,s)=(k-1)/2(N-s) \quad (8)$$

Упомянутые результаты позволяют сформулировать и доказать следующую теорему.

Теорема. Для множества всех стационарных источников Ω_∞ существует слабо универсальное кодирование в алфавит с неравнозначными символами.

Доказательство. Докажем, что для произвольного источника $\Theta\in\Omega_\infty$ имеет место неравенство

$$\lim_{N\to\infty}R(N,\Theta,\bar{t})=\lim_{N\to\infty}\inf_{\varphi}R(N,\Theta_\infty,\varphi,\bar{t})=0. \quad (9)$$

Каждый стационарный источник можно рассматривать как предел последовательности марковских источников Θ_s, которые задаются условными вероятностями $p_\Theta^s\left(x_{i_{s+1}}\middle|x_{i_1}x_{i_2}\ldots x_{i_s}\right)$.

Для каждого фиксированного s существует универсальное кодирование [11] для которого имеет место оценка (6). Преобразуем правую часть (6):

$$L\left(N,\Theta,\varphi_0^s,\bar{t}\ \right)\le \quad \frac{H(\Theta_s)-H(\Theta_\infty)}{C(\bar{t})}+\frac{H(\Theta_\infty)}{C(\bar{t})}+$$

$$+\frac{k^s(k-1)}{2}\cdot\frac{\log(N-s)}{C(\bar{t})N}+$$

$$+\frac{k^s(k-1)}{2}\cdot\frac{T^*(k,s)+\log e\left(1+\alpha(N,k,s)\right)}{C(\bar{t})N}+\frac{t^{**}}{N}.$$

Согласно (4), $\lim\limits_{s\to\infty}\left(H(\Theta_s)-H(\Theta_\infty)\right)=0$, следовательно, с ростом N, поведение правой части (6) определяется слагаемым

$$\left(k^s(k-1)\log N\right)\Big/2N,$$

которое стремится к нулю при выборе

$$s=O\left(\log\log N/\log k\right).$$

Действительно, пусть $s=\left(c\log\log N\right)\Big/\log k$, $c=const$. Тогда

$$\lim_{N\to\infty}\frac{k^s(k-1)\log N}{2N}=\lim_{N\to\infty}\frac{k^{\frac{c\log\log N}{\log k}}(k-1)\log N}{2N}=$$

$$=\lim_{N\to\infty}\frac{k-1}{2}\cdot\frac{\log^{c+1}N}{N}=0.$$

Заключение

Полученный в данной работе результат утверждает существование *слабо универсального* кодирования при кодировании произвольного стационарного источника неравнозначными символами, т.е. сходимость избыточности предлагаемого метода кодирования не является равномерной.

СПИСОК ЛИТЕРАТУРЫ

1. Хаффман, Д.А. Метод построения кодов с минимальной избыточностью [Текст] / Д.А. Хаффман. - Кибернетический сборник. – 1961, Вып.3. – С.79 - 87.
2. Katona, G. General theory of noiseless channels. // G. Katona. – UDINE, Courses and lectures – 1970, №.31. – P.69.
3. Кричевский, Р.Е., Длина блока, необходимая для получения заданной избыточности [Текст] / Р.Е. Кричевский. – Доклад АНСССР. – 1966, Т171, №1.11. – С.37-40.
4. Кричевский, Р.Е. Связь между избыточностью кодирования и достоверностью сведений об источнике [Текст] / Р.Е. Кричевский. – Проблемы передачи информации. – 1968, Т.4, № 3. – С.48-57.
5. Трофимов, В. К. Избыточность универсального кодирования произвольных марковских источников [Текст] / В.К. Трофимов. - Проблемы передачи информации. – 1974, Т. 10, № 4. - С.16-24.
6. Штарьков, Ю. М. Кодирование сообщений конечной длины на выходе источника с неизвестной статистикой [Текст] / Ю.М. Штарьков. - Материалы V конф. по теории кодирования и передачи информации, Москва-Горький. — 1972, Ч. 1. - С. 147-152.
7. Штарьков, Ю.М. Кодирование длин серий в условиях априорной неизвестности [Текст] / Ю.М. Штарьков, В.Ф. Бабкин. - Тематический выпуск «Аппаратура для космических исследований» ИКИ АН СССР. –1973. – С. 3-9.
8. Трофимов, В.К. Сжатие неравнозначными символами информации, порожденной неизвестным источником без памяти [Текст] / В.К. Трофимов, Т.В. Храмова. - Автометрия. - 2012, Т.48, №1. - С.30- 44.
9. Трофимов, В.К. Сжатие информации порожденной неизвестным источником [Текст] / В.К. Трофимов, Т.В. Храмова. - Электросвязь. - 2012, №4. - С.41-44.
10. Trofimov, V. K. Compression of information generated by an unknown memoryless source by nonequivalent symbols / V. K. Trofimov, T. V. Khramova. - Optoelectronics, Instrumentation and Data Processing. New York. - February 2012, V.48, Is. 1. – P. 24-36.
11. Трофимов, В.К. Универсальное кодирование марковских источников неравнозначными символами [Текст] / В.К. Трофимов, Т.В. Храмова. - Дискретный анализ и исследование операций. - Май—июнь 2013, Т. 20, № 3. — С. 71–83.
12. Фано, Р. Передача информации. Статистическая теория связи [Текст] / Р. Фано. - М.: Мир, 1965..

д.т.н., профессор каф. ВМ СибГУТИ ***Трофимов В.К.***- *trofimov@sibsutis.ru .к.т.н., доцент каф. ВМ СибГУТИ* ***Храмова Т.В.*** *tvkhramova@gmail.com.*