УДК 621.391.15

ОБ ЭФФЕКТИВНОСТИ РАВНОМЕРНОГО ПО ВЫХОДУ КОДИРОВАНИЯ БЕРНУЛЛИЕВСКИХ ИСТОЧНИКОВ ПРИ НЕИЗВЕСТНОЙ СТАТИСТИКЕ СООБЩЕНИЙ*

В. К. Трофимов

Государственное образовательное учреждение высшего профессионального образования «Сибирский государственный университет телекоммуникаций и информатики», 630102, г. Новосибирск, ул. Кирова, 86
E-mail: trofimov@sibsutis.ru

Предложен метод построения равномерного по выходу кодирования источников без памяти с неизвестной статистикой сообщений. Проведена оценка эффективности данного метода. Установлено, что это кодирование эффективнее равномерного по входу.

Ключевые слова: энтропия, кодирование, избыточность, стоимость кодирования, источник сообщений.

Введение. Уменьшение объёма передаваемой информации — одна из проблем теории информации. Уменьшение объёма, например, эквивалентно увеличению скорости передачи, уменьшению процессорного времени, необходимого для её обработки. В настоящее время существует два направления сокращения избыточности: сжатие данных и кодирование дискретных источников. В работе [1] заложены основы теории передачи информации, с помощью которой получены различные алгоритмы устранения избыточности как при известной, так и при неизвестной статистике сообщений. Достаточно подробную библиографию по этому вопросу можно найти в [2]. Кодирование дискретных источников используется при факсимильной передаче изображений [3], в задачах поиска и хранения данных [4], в теории управления [5], при выявлении скрытой информации [6]. Решение этой проблемы значимо и при создании большемасштабных распределённых вычислительных систем [7].

В предлагаемой работе исследуется метод пословного сжатия информации блоками (словами одинаковой длины). Такое кодирование называют равномерным по выходу, оно является обобщением кодирования длин серий [8], характеризуется отсутствием бегущей ошибки синхронизации и удобно для последующего применения корректирующих кодов. Равномерное по выходу кодирование при известной статистике сообщений изучалось в [9–11], а при неизвестной — в [12–14]. В частности, в [11] доказано, что равномерное по выходу кодирование известных марковских источников с ростом памяти эффективнее, чем равномерное по входу.

Цель данной работы — показать, что при неизвестной статистике сообщений равномерное по выходу кодирование (даже бернуллиевских источников) эффективнее равномерного по входу.

Основные определения и обозначения. Пусть буквы конечного входного алфавита $A = \{a_1, \ldots, a_k\}$, $2 \le k < \infty$, порождаются источником θ независимо с вероятностями $P_\theta(a_i) = \theta_i$, $i = \overline{1,k}$, $\theta_1 + \theta_2 + \ldots + \theta_k = 1$. В этом случае считаем, что θ — бернуллиевский источник, однозначно определённый неотрицательными числами θ_i, $i = 1, k$,

*Работа выполнена при поддержке Совета по грантам Президента РФ (ведущая научная школа НШ-5176.2010.9) и Российского фонда фундаментальных исследований (грант № 09-07-00095-а).

$\theta_i \geq 0$, сумма которых равна единице. Верно и обратное утверждение: любой набор чисел θ_i, $i = \overline{1,k}$, удовлетворяющий перечисленным выше условиям, однозначно определяет бернуллиевский источник. Множество слов, взятых в произвольном алфавите, называется префиксным, если никакое слово не является началом другого. Множество слов T, взятых в алфавите A, назовём кодовым, если оно полное, конечное, префиксное. В этом случае произвольная полубесконечная последовательность букв входного алфавита, порождаемая источником, однозначно разбивается на последовательность слов из множества T. Из неравенства Крафта — Макмиллана [15] следует: самое общее из всех возможных дешифрируемых кодирований φ состоит в том, что полубесконечная последовательность букв, порождённая источником, разбивается в соответствии с кодовым множеством T на слова, переведённые с помощью отображения φ в слова выходного алфавита B, который, не уменьшая общности, можно считать двоичным. При этом множество слов в выходном алфавите $\varphi(T) = \{\varphi(u),\ u \in T\}$ является префиксным.

Если длины всех слов некоторого множества C равны между собой, то считается, что C состоит из блоков; в противном случае — из слов переменной длины. В зависимости от вида множества T и $\varphi(T)$ логически возможны: кодирование, отображающее блоки в слова переменной длины (обозначается bV); кодирование, отображающее слова переменной длины в блоки (Vb); кодирование, отображающее слова переменной длины в слова переменной длины (VV); кодирование, отображающее блоки в блоки (bb).

Рассмотрим первые два типа кодирований. Если u — слово во входном алфавите, то через $|u|$ обозначим число букв в этом слове и будем называть длиной слова u. Если T — множество слов в некотором алфавите, то $\|T\|$ — число слов в T. Итак, всякое кодирование φ полностью определяется тройкой $(T, \varphi, \varphi(T))$. Пусть задано кодирование φ, которое является кодированием типа σ, $\sigma = bV, Vb$. Среднее число букв выходного алфавита, приходящихся на одну букву входного при кодировании φ, назовём стоимостью кодирования φ и обозначим $C_\sigma(T, \theta, \varphi)$. Величина $C_\sigma(T, \theta, \varphi)$ определяется равенством [15, 16]

$$C_\sigma(T, \theta, \varphi) = \frac{1}{d(T, \theta)} \sum_{u \in T} P_\theta(u)|\varphi(u)|, \tag{1}$$

где $P_\theta(u)$ — вероятность порождения слова u источником θ; $d(T, \theta) = \sum\limits_{u \in T} P_\theta(u)|u|$ — средняя длина кодовых слов из T; если $T = A^n$, то $d(T, \theta) = n$. Через $H(\theta)$ обозначим энтропию источника θ. Для бернуллиевского источника θ величина $H(\theta)$ определяется выражением [1]

$$H(\theta) = -\sum_{i=1}^{k} \theta_i \log \theta_i,$$

где $\log x = \log_2 x$, $0 \log 0 = 0$.

Эффективность кодирования φ оценивается разностью между стоимостью кодирования $C_\sigma(T, \theta, \varphi)$, определяемой равенством (1), и энтропией источника $H(\theta)$. Эта разность называется избыточностью кодирования φ и обозначается $r_\sigma(T, \theta, \varphi)$. Таким образом, по определению

$$r_\sigma(T, \theta, \varphi) = C_\sigma(T, \theta, \varphi) - H(\theta). \tag{2}$$

Избыточностью универсального кодирования типа σ для заданного множества источников $\Omega \subseteq \Omega_0$ и сложностью n назовём величину

$$R_\sigma(n, \Omega) = \inf_{\varphi} \sup_{\theta \in \Omega} r_\sigma(T, \theta, \varphi), \tag{3}$$

где нижняя грань берётся по всем кодированиям φ, для которых $\|T\| \le k^n$.

При известной статистике сообщений величина $R_{bV}(n,\theta)$ изучена в [1, 2], а величина $R_{Vb}(n,\theta)$ исследовалась в [9–11, 16]. В работе [11] показано, что кодирование типа Vb для марковских источников с растущей памятью s эффективнее кодирования bV, т. е. избыточность кодирования $R_{Vb}(n,\theta)$ меньше, чем $R_{bV}(n,\theta)$. В [17] доказано, что для избыточности $R_{bV}(n,\Omega_0)$ универсального равномерного по входу кодирования бернуллиевских источников Ω_0 имеет место асимптотическое равенство

$$R_{bV}(n,\Omega_0) \sim \frac{k-1}{2}\log k\frac{\log\log\|A^n\|}{\log\|A^n\|}. \tag{4}$$

Отметим, что (4) верно и в случае, если Ω_0 заменить любым множеством $\Omega \subset \Omega_0$, которое имеет ненулевую меру Лебега. Универсальное кодирование типа Vb изучалось в [12–14].

В данной работе показано, что существует последовательность кодирований φ_n типа Vb с областью определения T_n таких, что $\|T_n\| \le k^n$, для которых избыточность $r_{Vb}(T_n,\theta,\varphi_n)$ при любом источнике θ из Ω_0 стремится к нулю. Получена верхняя оценка избыточности в зависимости от $\|T_n\|$.

Введём ещё ряд обозначений. Если u — произвольное слово во входном алфавите, то через $t_i(u)$, $i = \overline{1,k}$, обозначим число вхождений буквы a_i в слово u. Очевидно, что сумма всех чисел $t_i(u)$, $i = \overline{1,k}$, равна длине слова u, т. е. $t_1(u) + t_2(u) + \ldots + t_k(u) = |u|$. Вероятность слова u, порождённого источником θ, находится по формуле

$$P_\theta(u) = \prod_{i=1}^{k}\theta_i^{t_i(u)}. \tag{5}$$

На множестве источников Ω_0 определим вероятностную меру [2] с плотностью

$$\omega(\theta) = \frac{\Gamma(k/2)}{\pi^{k/2}\prod\limits_{i=1}^{k}\sqrt{\theta_i}}.$$

Здесь $\Gamma(z)$ — гамма-функция от числа z. Среднюю вероятность порождения слова u по множеству источников Ω_0 с плотностью $\omega(\theta)$ обозначим $\overline{P}(u)$. Величина $\overline{P}(u)$ определяется равенством

$$\overline{P}(u) = \int\limits_{\Omega_0}\omega(\theta)P_\theta(u)d\theta = \frac{\Gamma(k/2)}{\pi^{k/2}}\,\frac{\prod\limits_{i=1}^{k}\Gamma(t_i(u)+1/2)}{\Gamma(|u|+k/2)}. \tag{6}$$

Квазиэнтропия слова u выражается как

$$F_0(u) = -\sum_{i=1}^{k}\frac{t_i(u)}{|u|}\log\frac{t_i(u)}{|u|}. \tag{7}$$

Используя формулу Стирлинга в виде

$$\log\Gamma(z) = \log\sqrt{2\pi} + (z-1/2)\log(z-1/2) - z\log l + c(z)\log e, \tag{8}$$

где $\frac{1}{2}(\log e - 1) \le c(z)\log e \le \frac{1}{2}\log e$, из (6)–(8) получаем

$$-\log \overline{P}(u) = |u|F_0(u) + \frac{k-1}{2}\log|u| + c(u). \tag{9}$$

В (9) для постоянной $c(u)$ выполняются неравенства

$$-1/2 \le (c(u) - 1/2)\log e \le 0.$$

Построение универсального равномерного по выходу кодирования и оценка его эффективности. Построим последовательность универсальных равномерных по выходу кодирований φ_n таких, что

$$\lim_{n\to\infty} \sup_{\theta\in\Omega_0} r_{Vb}(T_n, \theta, \varphi_n) = 0.$$

Оценим скорость сходимости избыточности предложенного кодирования при заданном θ, $\theta \in \Omega_0$. Тогда справедлива

Теорема 1. Существует последовательность универсальных Vb-кодов φ_n с областью определения T_n, $\|T_n\| \le k^n$, такая, что для любого источника θ, $\theta \in \Omega_0$, выполняется неравенство

$$r_{Vb}(T_n, \theta, \varphi_n) \le \frac{k+1}{2}\frac{\log d(T_n,\theta)}{d(T_n,\theta)} + \frac{c}{d(T_n,\theta)},$$

где $d(T_n,\theta) = \sum\limits_{u\in T_n} P_\theta(u)|u|$ — средняя задержка кодового множества T_n; c — постоянная, не зависящая от θ и T_n.

Доказательство. Вся сложность определения равномерного по выходу кодирования φ_n заключается в построении последовательности кодовых множеств T_n, являющихся областью определения φ_n. Область значений $\varphi(T_n)$ — это наименьшее из множеств B^m, для которого $\|B^m\| \ge \|T_n\|$. Опишем алгоритм построения T_n, n — произвольное натуральное число. В множество T_n включаем все слова u в алфавите A, для которых одновременно выполнены следующие неравенства:

$$\begin{cases} 1/\overline{P}(u) \le k^n, & u \in T_n, \\ 1/\overline{P}(ua_j) > k^n, & u \in T_n, \text{ и, по крайней мере, для одного } a_j \in A. \end{cases} \tag{10}$$

В силу полноты T_n справедливо $\sum\limits_{u\in T_n} \overline{P}(u) = 1$. Поэтому, умножив обе части первого условия из (10) на $\overline{P}(u)/k^n$ и просуммировав по всем u из T_n, получаем

$$k^n \ge \|T_n\|. \tag{11}$$

При оптимальном Vb-кодировании φ_n для любого слова u из T_n имеем

$$|\varphi_n(u)| =]\log\|T_n\|[.$$

Следовательно, избыточность $r(T_n,\theta,\varphi_n)$ кодирования φ_n для источника θ вычисляется по формуле

$$r(T_n,\theta,\varphi_n) = \frac{]\log\|T_n\|[}{d(T_n,\theta)} - H(\theta). \tag{12}$$

Из (12) с учётом (11) следует

$$r(T_n,\theta,\varphi_n)\le\frac{n\log k}{d(T_n,\theta)}-H(\theta)+\frac{1}{d(T_n,\theta)}. \tag{13}$$

Заметим, что для любого слова u из T_n в силу определения $\overline{P}(u)$ (см. выражение (6)) и второго условия из (10) существует $a_j\in A$ такое, что выполняется неравенство

$$\frac{1}{\overline{P}(ua_j)}=\Big(\frac{1}{\overline{P}(u)}\Big)\frac{t_j(u)}{|u|+1+k/2}>k^n.$$

Отсюда и из (13), учитывая $t_j(u)\ge 1$, получаем

$$r(T_n,\theta,\varphi_n)\le\frac{\sum\limits_{u\in T_n}P_\theta(u)(n\log k)}{d(T_n,\theta_n)}-H(\theta)+\frac{1}{d(T_n,\theta)}\le$$

$$\le\frac{-\sum\limits_{u\in T_n}P_\theta(u)\log\overline{P}(u)}{d(T_n,\theta)}-H(\theta)+\frac{\sum\limits_{u\in T_n}P_\theta(u)\log(|u|+k/2+1)+1}{d(T_n,\theta)}. \tag{14}$$

Из (14) в силу (9) будем иметь

$$r(T_n,\theta,\varphi_n)\le\sum_{u\in T_n}\frac{P_\theta(u)|u|F_0(u)}{d(T_n,\theta_n)}-H(\theta)+\frac{c_1}{d(T_n,\theta)}+$$

$$+\frac{k-1}{2}\,\frac{\sum\limits_{u\in T_n}P_\theta(u)\log\|u\|}{d(T_n,\theta)}+\frac{P_\theta(u)\log(|u|+k/2+1)}{d(T_n,\theta)}. \tag{15}$$

Согласно тождеству Вальда [18] для любого $j=\overline{1,k}$ справедливо равенство

$$\sum_{u\in T_n}P_\theta(u)r_j(u)=d(T_n,\theta)\theta_j. \tag{16}$$

Воспользовавшись неравенством Иенсена для функции $-x\log x$ и определением квазиэнтропии $F_0(u)$, имеем

$$\frac{1}{d(T_n,\theta)}\sum_{u\in T_n}P_\theta(u)|u|F_0(u)\le H(\theta).$$

Кроме того, применяя неравенство Иенсена для функции $\log x$, заключаем, что

$$\sum_{u\in T_n}P_\theta(u)\log|u|\le\log d(T_n,\theta).$$

Используя два последних неравенства и (15), окончательно запишем

$$r(T_n,\theta,\varphi_n)\le\frac{k+1}{2}\,\frac{\log d(T_n,\theta)}{d(T_n,\theta)}+\frac{C}{d(T_n,\varphi_n,\theta)}.$$

Теорема полностью доказана.

Замечание. Величина $d(T_n,\theta)\to\infty$ при любом $\theta\in\Omega_0$ и $n\to\infty$, поэтому из теоремы 1 вытекает, что для произвольного источника θ избыточность кодирования стремится к нулю, т. е. предложенное кодирование является оптимальным для любого источника θ из Ω_0.

Формулировка и доказательство основного утверждения. Докажем вспомогательное предложение.

Лемма. Для последовательности кодовых множеств T_n, $n = 1, 2, \ldots$, построенных при доказательстве теоремы 1, и произвольного источника $\theta \in \Omega_0$ такого, что $H(\theta) > \delta > 0$, справедливо равенство

$$d(T_n, \theta) = \frac{\log \|T_n\|}{H(\theta) + \alpha_n(\theta)},$$

где $\alpha_n(\theta) \to 0$ при $n \to \infty$.

Доказательство. Совершенно очевидно, что

$$d(T_n, \theta) \geq \min_{u \in T_n} |u|.$$

Величина $\min\limits_{u \in T_n} |u| \to \infty$ при $n \to \infty$. Отсюда с учётом того, что функция $\dfrac{\log x}{x}$ убывает при $x \to \infty$, и из теоремы 1 вытекает неравенство

$$r(T_n, \theta, \varphi_n) \leq \frac{k+1}{2} \frac{\log \min\limits_{u \in T_n} |u|}{\min\limits_{u \in T_n} |u|} + \frac{C}{\min\limits_{u \in T_n} |u|}.$$

Правая часть последнего неравенства не зависит от θ, а это означает, что $r(T_n, \theta, \varphi_n) \to 0$ равномерно по θ при $n \to \infty$.

Используя определение $r(T_n, \theta, \varphi_n)$ и теорему о связи предела и бесконечно малых величин, имеем

$$\frac{\log |T_n|}{d(T_n, \theta)} - H(\theta) = \alpha(n)$$

($\alpha(n)$ бесконечно малая при $n \to \infty$). Отсюда получаем утверждение леммы

$$d(T_n, \theta) = \frac{\log |T_n|}{H(\theta) + \alpha(n)}.$$

Перейдём к формулировке и доказательству основного результата работы.

Теорема 2. Существует последовательность кодирований φ_n, $n = 1, 2, \ldots$, такая, что избыточность $r_{Vb}(T_n, \theta, \varphi_n)$ кодирования φ_n при заданной мощности кодового множества $\|T_n\| \leq k^n$ для произвольного источника $\theta \in \Omega_0$ удовлетворяет неравенству

$$r_{Vb}(n, \theta) \leq \frac{k+1}{2} H(\theta) \frac{\log \log \|T_n\|}{\log \|T_n\|} (1 + 0(1)),$$

где $0(1) \to 0$ равномерно по θ при $n \to \infty$.

Доказательство. Из теоремы 1 и доказанной леммы имеем

$$r(T_n, \varphi_n, \theta) \leq \frac{\dfrac{k+1}{2} \log \dfrac{\log \|T_n\|}{H(\theta) + \alpha(n)} + C}{\dfrac{\log \|T_n\|}{H(\theta) + \alpha(n)}}, \tag{17}$$

или после преобразований

$$r(T_n,\varphi_n,\theta)\le\frac{k+1}{2}\,[H(\theta)+\alpha(n)]\,\frac{\log\log\|T_n\|-\log(H(\theta)+\alpha(n))}{\log\|T_n\|}+\frac{(H(\theta)+\alpha(n))C}{\log\|T_n\|}. \quad (18)$$

Так как $H(\theta)\le\log k$, то функция $(H(\theta)+\alpha(n))\log[H(\theta)+\alpha(n)]$ по абсолютной величине не превосходит $(1+\log k)\log(1+\log k)$, поэтому из соотношения (18) вытекает неравенство

$$r_{Vb}(T_n,\varphi_n,\theta)\le\frac{k+1}{2}\,H(\theta)\frac{\log\log\|T_n\|}{\log\|T_n\|}\,(1+0(1)), \quad (19)$$

где $0(1)\to 0$ равномерно по θ.

Теорема доказана.

Следствие. При выполнении неравенства $\dfrac{H(\theta)}{\log k}<\dfrac{k-1}{k+1}$ равномерное по выходу кодирование эффективнее равномерного по входу.

Доказательство. При фиксированном источнике $\theta\in\Omega_0$ избыточность равномерного по входу кодирования, как это было доказано в теореме 2, стремится к нулю со скоростью

$$\frac{k+1}{2}\,H(\theta)\frac{\log\log\|T_n\|}{\log\|T_n\|}, \quad (20)$$

а равномерное по входу кодирование согласно (4) имеет скорость стремления к нулю, равную величине

$$\frac{k-1}{2}\log k\frac{\log\log\|T_n\|}{\log\|T_n\|}. \quad (21)$$

При фиксированном n, если $H(\theta)\to 0$, величина (20) также стремится к нулю, при этом величина (21) постоянная. Следовательно, для источников, у которых энтропия мала, равномерное по выходу кодирование всегда лучше равномерного по входу. Из сравнения (20) и (21) видно, что при выполнении неравенства $\dfrac{H(\theta)}{\log k}<\dfrac{k-1}{k+1}$ предложенное кодирование эффективнее bV-кодирования.

Следствие доказано.

Заключение. В данной работе предложен метод неравномерного по входу, но равномерного по выходу кодирования информации, порождаемой неизвестным источником без памяти. Доказано, что при таком кодировании среднее число букв выходного алфавита, приходящихся на одну букву входного алфавита, с ростом числа кодовых слов стремится к энтропии. Установлено, что предложенное кодирование лучше равномерного по входу.

СПИСОК ЛИТЕРАТУРЫ

1. **Шеннон К.** Математическая теория связи. Работы по теории информации и кибернетике. М: Изд-во иностр. лит., 1963. С. 243–332.
2. **Krichevsky R. E., Trofimov V. K.** The performance of universal encoding // IEEE Trans. Inform. Theory. 1981. **27**, N 2. P. 199–207.
3. **Usubuchi T., Omachi T., Iinuma K.** Adaptive predictive coding for newspaper facsimile // Proc. IEEE. 1980. **68**, N 7. P. 807–813.

4. **Бабкин В. Ф., Куделова К., Луценко В. Н. и др.** Опыт применения бортовой информационно-вычислительной системы для обработки данных и управления экспериментом «Интершок» // Космические исследования. 1986. **24**, № 2. С. 210–216.
5. **Петров Б. Н., Добрушин Р. Л., Пинскер М. С. и др.** О некоторых взаимосвязях теории информации и теории управления // Проблемы управления и теории информации. 1976. **5**, № 1. С. 31–38.
6. **Жилкин М. Ю., Меленцова Н. А., Рябко Б. Я.** Метод выявления скрытой информации, базирующейся на сжатии данных // Вычислительные технологии. 2007. **12**, вып. 4. С. 26–31.
7. **Хорошевский В. Г.** Архитектура вычислительных систем. М.: МГТУ им. Н. Э. Баумана, 2005. 512 с.
8. **Блох Э. Л.** О передаче бинарной последовательности равномерным кодом // Проблемы передачи информации. 1960. Вып. 5. С. 12–22.
9. **Jelinek F., Schneider K.** On variable-length-to-block coding // IEEE Trans. Inform. Theory. 1972. **18**, N 6. P. 765–774.
10. **Трофимов В. К.** Эффективное кодирование блоками слов различной длины, порожденных известным марковским источником // Обработка информации в системах связи. Л.: ЛЭИС, 1985. Том 29. С. 9–15.
11. **Ziv J.** Variable-to-fixed length codes are better than fixed-to-variable length codes for Marcov sources // IEEE Trans. Inform. Theory. 1990. **36**, N 4. P. 861–863.
12. **Трофимов В. К.** Универсальное равномерное по выходу кодирование бернуллиевских источников // Методы дискретного анализа в теории кодов и схем. Новосибирск: ИМ СО АН СССР, 1976. Вып. 29. С. 87–99.
13. **Lawrence J. C.** A new universal coding scheme for the binary memoryless source // IEEE Trans. Inform. Theory. 1977. **23**, N 4. P. 446–472.
14. **Штарьков Ю. М.** Равномерное по выходу универсальное кодирование дискретных источников без памяти // Проблемы передачи информации. 1991. **27**, № 1. С. 3–13.
15. **Галлагер Р.** Теория информации и надежная связь. М.: Сов. радио, 1974. 720 с.
16. **Ходак Г. Л.** Оценки избыточности при пословном кодировании сообщений, порождаемых бернуллиевским источником // Проблемы передачи информации. 1972. **8**, № 2. С. 21–32.
17. **Кричевский Р. Е.** Связь между избыточностью кодирования и достоверностью сведений об источнике // Проблемы передачи информации. 1968. **4**, № 3. С. 48–57.
18. **Боровков А. А.** Курс теории вероятностей. М.: Наука, 1972. 287 с.

Поступила в редакцию 30 июня 2010 г.