

## ВЛИЯНИЕ ОБЪЕМА СЛОВАРЯ НА СТЕПЕНЬ СЖАТИЯ ТЕКСТА

*М. П. Шарова*

Одной из важных задач теории информации является задача неискажающего кодирования источника, в прикладных областях называемая задачей сжатия данных, например текстов на естественных языках, с сохранением возможности их однозначного восстановления (декодирования). В методах словарного сжатия алгоритм, сокращающий длину текста, обычно использует словарь, для хранения которого отводится значительный объем машинной памяти. Наряду с эффективностью сжатия эта характеристика является одной из важнейших характеристик метода. В данной работе предлагается метод, позволяющий существенно уменьшить объем хранимого в памяти словаря при сохранении эффективности сжатия. Результат работы получен аналитически и подтвержден экспериментально.

### Введение

Задача неискажающего сжатия данных, например текстов на естественных языках, является одной из центральных в теории информации. В настоящее время для кодирования текстовой информации хорошо известны и находят многочисленное практическое применение методы словарного сжатия, например обширная группа методов, базирующихся на известных алгоритмах Лемпеля — Зива [14, 15] и их модификациях [7]. Во многих методах словарного сжатия (см., например, [7, 9, 11, 12]) для кодирования текста составляется частотный словарь, представляющий собой список всех используемых в тексте различных слов с указанием частоты их появления. Тогда при кодировании сообщения каждому слову приписывается код, длина которого тем меньше, чем больше частота его появления.

Сообщение может кодироваться и побуквенно (например, с помощью известного кода Хаффмана [4]). В этом случае каждая буква сообщения кодируется отдельно от другой с учетом частоты ее встречаемости.

Достоинства словарных методов — относительно небольшая сложность их реализации, что при высокой скорости кодирования и декодирования позволяет достигать высокой степени сжатия.

Эффективность сжатия текста мы будем оценивать коэффициентом (степенью) сжатия, который определяется отношением объема текста (в битах), полученного в результате кодирования, к первоначальному объему текста. Наряду с эффективностью сжатия важнейшей характеристикой метода является также объем используемой памяти кодера и декодера. Однако в методах словарного сжатия значительный объем памяти занимает хранение слов словаря. Поэтому возникает задача построения методов, которые для хранения словаря позволяют использовать небольшой объем памяти при сохранении эффективности сжатия.

Одним из основных подходов к задаче кодирования текстов на естественных языках может служить построение кодов, базирующихся на общих законах лингвистики, в частности законе Ципфа [13], которые в отличие от универсальных методов (см. [10, 14, 15]) используют знания о статистической структуре источника сообщений. В данной работе предлагается метод, позволяющий существенно уменьшить объем хранимого в памяти словаря при сохранении эффективности сжатия текста. Метод основан на использовании смешанного пословно-побуквенного кодирования. На основе установленного в лингвистике закона Ципфа определяются оценки для коэффициента сжатия текста при использовании предлагаемого метода кодирования, позволяющие установить оптимальный объем словаря, который требуется хранить в машинной памяти.

Наряду с теоретическим результатом был проведен ряд экспериментов, в которых использовались различные русские и английские тексты с разными объемами словарей и подсчитывалась степень их сжатия. Исследования показали, что результаты экспериментов хорошо согласуются с результатами, полученными аналитически.

Остановимся кратко на содержании работы. В п. 1 приводится закон Ципфа, в п. 2 изучается влияние объема словаря на степень сжатия текста, а в п. 3 приводятся результаты экспериментов.

### 1. Закон Ципфа

Этот закон был установлен Ципфом [13] на основе анализа текстов на различных языках и состоит в следующем. Пусть имеется текст на каком-либо естественном языке,  $N$  — общее число слов в тексте,  $D$  — объем словаря этого текста. Расположим слова в словаре в порядке убывания частот их появления и пронумеруем их от 1 до  $D$ . Тогда зависимость между частотой и номером слова в словаре может быть задана формулой

$$f_i = \frac{c}{i^\gamma}, \quad (1)$$

где  $f_i$  — частота слова с номером  $i$  ( $1 \leq i \leq D$ ),  $\gamma \approx 1$  — постоянная,

а  $c > 0$  зависит только от  $N$  и  $D$ . Это соотношение и получило название закона Ципфа. В дальнейшем закон был подтвержден многими исследователями для большинства европейских языков (см., например, [1]). Кроме того, существуют математические модели, объясняющие закон Ципфа с позиций теории информации и теории случайных процессов [2, 8].

Разделим обе части равенства (1) на  $N$ . Тогда закон Ципфа можно записать в виде

$$p_i = \frac{k}{i^\gamma},$$

где  $p_i$  — вероятность слова с номером  $i$ , а

$$k = c/N = 1 / \sum_{i=1}^D \frac{1}{i^\gamma}.$$

Так как для слов естественного языка  $\gamma \approx 1$ , то в дальнейшем изложении будем пользоваться формулой закона Ципфа вида  $p_i = k/i$ .

## 2. Влияние объема словаря на степень сжатия текста

Пусть имеется некоторый текст с объемом словаря  $D$ . Наша задача — выяснить, как влияет изменение объема словаря на степень сжатия текста и найти оптимальный объем словаря, который требуется хранить в памяти при использовании метода словарного сжатия. Напомним, что под объемом словаря мы понимаем число различных слов, содержащихся в словаре, а коэффициентом (степенью) сжатия называем отношение объема текста (в битах), полученного в результате кодирования, к первоначальному объему текста.

Рассмотрение данной задачи начнем с описания метода кодирования, который обозначим через  $A$ . Все слова в словаре расположим по убыванию частот их появления и применим смешанное пословно-побуквенное кодирование:  $j - 1$  слов словаря ( $1 < j \leq D$ ) будем кодировать на основе частоты их появления в тексте (с помощью кода Шеннона [5]), а остальные слова — побуквенно с помощью кода Хаффмана [4]. Для различения слов первой части (включающей  $j - 1$  слов словаря) и слов второй части (включающей  $D - j + 1$  слов) введем специальное кодовое слово, отличное от имеющихся кодовых слов, которое назовем «флагом». Для определенности будем считать, что «флаг» появляется перед словами второй части словаря. Длину этого кодового слова возьмем равной  $\left\lceil -\log \sum_{i=j}^D p_i \right\rceil$ , где

$p_i$  — вероятность  $i$ -го слова. Отметим, что при таком способе кодирования используются и хранятся в памяти частоты лишь первых  $j - 1$  слов словаря.

Пусть  $l_1, \dots, l_D$  — длины слов словаря объема  $D$  (в битах). Введем ряд вспомогательных величин, которые будем использовать в дальнейшем. Определим величины  $l' = \sum_{i=j}^D l_i p_i$  и  $l = \sum_{i=1}^D l_i p_i$  как значения средней длины слов, относящихся ко второй части словаря и ко всему словарю соответственно. Через  $h'$  обозначим среднее количество информации (или энтропию) в одной букве для слов второй части словаря (в битах), через  $h$  — среднее количество информации в одной букве для слов всего словаря, а через  $F_A(j)$  — коэффициент сжатия текста методом  $A$ , использующим словарь из  $j$  слов. Следует отметить, что величины  $h'$  и  $l'$  зависят от  $j$ , причем в реальных текстах значение  $l'$  для редких слов больше, чем значение  $l$  для всех слов словаря. Кроме того, для реальных текстов на естественных языках объем словаря  $D$  не превышает 50000 слов при средней длине слова 30–60 бит, т. е.  $l > 2 \cdot \ln D$ . Определим

$$\delta(j) = \frac{(h' + 1)l'}{8} \ln 2, \quad a = \frac{1}{l \ln 2}. \quad (2)$$

Следующая теорема устанавливает верхнюю оценку коэффициента сжатия  $F_A(j - 1)$ .

**Теорема.** Пусть имеется текст с объемом словаря  $D$ , в котором распределение слов по частоте подчинено закону Ципфа. Тогда для метода  $A$ , использующего при кодировании словарь из  $j - 1$  слов ( $1 < j \leq D$ ), коэффициент сжатия текста  $F_A(j - 1)$  удовлетворяет неравенству

$$F_A(j - 1) < a \left( \frac{\ln^2 j}{2 \ln(D + 1)} + \ln(2 \ln(D + 1)) \right) + C \frac{\ln \ln(D + 1)}{\ln(D + 1)} + a \left( \frac{\ln(D + 1) - \ln j + c_1}{\ln(D + 1)} \right) \delta(j), \quad (3)$$

где  $\delta(j)$  и  $a$  определяются формулой (2),  $c_1 = 0,577 \dots$  — постоянная Эйлера, а  $C > 0$  — не зависящая от  $j$  константа.

**Доказательство.** Оценим объем текста (в битах), полученный в результате кодирования по методу  $A$ . Очевидно, что требуемый объем складывается из трех частей: объема текста, полученного при пословном кодировании  $j - 1$  слов словаря, объема текста, полученного при побуквенном кодировании оставшихся слов словаря, и объема, занимаемого «флагом». Согласно закону Ципфа

$$p_i = \frac{k}{i},$$

где  $p_i$  — вероятность  $i$ -го слова, а  $k = 1 / \sum_{i=1}^D \frac{1}{i}$ . Для дальнейших преобразований воспользуемся следующими известными оценками конечных сумм (см., например, [3]):

$$\ln j < \sum_{i=1}^{j-1} \frac{1}{i} < \ln j + c_1, \quad (4)$$

$$\frac{\ln^2 j}{2} + c'_2 < \sum_{i=1}^{j-1} \frac{\ln i}{i} < \frac{\ln^2 j}{2} + c_2, \quad (5)$$

где  $c_1 = 0,577\dots$  — постоянная Эйлера,  $c'_2 = -0,3$ ,  $c_2 = -0,105\dots$  (численные значения констант можно получить из ряда Эйлера — Маклорена (см., например, [3])). Так как  $k \sum_{i=1}^D \frac{1}{i} = 1$ , то

$$\frac{1}{\ln(D+1) + c_1} < k < \frac{1}{\ln(D+1)}. \quad (6)$$

Сначала оценим сверху объем текста  $B_1(j)$ , полученный при пословном кодировании  $j-1$  слов словаря. Эти слова кодируются с помощью кода Шеннона, длины кодовых слов которого равны  $l_i = \lceil -\log p_i \rceil$ , где  $p_i$  — вероятность  $i$ -го слова. Так как  $-\log p_i \leq \lceil -\log p_i \rceil < -\log p_i + 1$ , то

$$B_1(j) < N \sum_{i=1}^{j-1} p_i (-\log p_i + 1).$$

Применяя известное неравенство  $\ln(1+x) < x$ , справедливое при любом  $x > -1$ , и неравенства (4)–(6), получаем

$$\begin{aligned} B_1(j) &< N \sum_{i=1}^{j-1} \frac{k}{i} (\log i - \log k + 1) = \frac{Nk}{\ln 2} \left( \sum_{i=1}^{j-1} \frac{\ln i}{i} + (\ln 2 - \ln k) \sum_{i=1}^{j-1} \frac{1}{i} \right) \\ &< \frac{N}{\ln 2 \cdot \ln(D+1)} \left( \frac{\ln^2 j}{2} + c_2 + (\ln 2 + \ln(\ln(D+1) + c_1))(\ln j + c_1) \right) \\ &< \frac{N}{\ln 2 \cdot \ln(D+1)} \left( \frac{\ln^2 j}{2} + c_2 + (\ln \ln(D+1) + \alpha)(\ln j + c_1) \right), \quad (7) \end{aligned}$$

где  $\alpha = \ln 2 + \frac{c_1}{\ln(D+1)}$ .

Теперь оценим сверху объем памяти  $B_2(j)$ , занимаемый «флагом». Применяя неравенства (5), (6) и учитывая, что  $-x \ln x < 1$ , получаем

$$\begin{aligned} B_2(j) &< N \sum_{i=j}^D p_i \left( -\log \sum_{i=j}^D p_i + 1 \right) = \frac{Nk}{\ln 2} \sum_{i=j}^D \frac{1}{i} \left( \ln 2 - \ln k - \ln \sum_{i=j}^D \frac{1}{i} \right) \\ &< \frac{Nk}{\ln 2} \left( \sum_{i=j}^D \frac{1}{i} (\ln 2 - \ln k) + 1 \right) \\ &< \frac{N}{\ln 2 \cdot \ln(D+1)} ((\ln(D+1) + c_1 - \ln j)(\ln 2 + \ln(\ln(D+1) + c_1)) + 1) \\ &< \frac{N}{\ln 2 \cdot \ln(D+1)} ((\ln(D+1) + c_1 - \ln j)(\ln \ln(D+1) + \alpha) + 1). \quad (8) \end{aligned}$$

Наконец, оценим сверху объем  $B_3(j)$  текста, полученный при побуквенном кодировании слов второй части словаря. Так как  $h'$  и  $l'$  — значения среднего количества информации (число бит на букву) и средней длины слова, относящиеся ко второй части словаря, то, считая, что каждая буква кодируется 8 битами, имеем, что  $(h'l')/8$  — среднее количество информации в одном слове. Поскольку избыточность кода Хаффмана не превосходит 1, из (4) и (6) следует, что

$$B_3(j) < \frac{N(h'+1)l'}{8} \sum_{i=j}^D p_i < \frac{N(h'+1)l'}{8 \ln(D+1)} (\ln(D+1) - \ln j + c_1). \quad (9)$$

Используя (7)–(9), получаем, что общий объем памяти  $B(j)$  (в битах), требующийся при кодировании текста, удовлетворяет соотношению

$$\begin{aligned} B(j) &= B_1(j) + B_2(j) + B_3(j) \\ &< \frac{N}{\ln 2 \cdot \ln(D+1)} \left( \frac{\ln^2 j}{2} + c_2 + (\ln \ln(D+1) + \alpha) \right. \\ &\quad \left. \times (\ln j + c_1) + (\ln(D+1) - \ln j + c_1)(\ln \ln(D+1) + \alpha + \delta(j)) + 1 \right), \end{aligned}$$

где  $\delta(j) = \ln 2 \frac{(h'+1)l'}{8}$ . Так как первоначальный объем  $B$  текста удовлетворяет соотношению

$$B = N \sum_{i=1}^D l_i p_i = Nl,$$

окончательно получаем, что коэффициент сжатия текста

$$\begin{aligned}
 F_A(j-1) &= \frac{B(j)}{B} < \frac{1}{l \cdot \ln 2 \cdot \ln(D+1)} \left( \frac{\ln^2 j}{2} + c_2 + (\ln \ln(D+1) + \alpha) \right. \\
 &\quad \times (\ln j + c_1) + (\ln(D+1) - \ln j + c_1)(\ln \ln(D+1) + \alpha + \delta(j)) + 1 \Big) \\
 &= \frac{1}{l \cdot \ln 2} \left( \frac{\ln^2 j}{2 \ln(D+1)} + \alpha + \delta(j) + \ln \ln(D+1) - \frac{\delta(j) \ln j}{\ln(D+1)} \right. \\
 &\quad \left. + \frac{2c_1 \ln \ln(D+1)}{\ln(D+1)} + \frac{c_2 + c_1(\delta(j) + 2\alpha) + 1}{\ln(D+1)} \right). \quad (10)
 \end{aligned}$$

Отсюда следует утверждение теоремы. Теорема доказана.

Используя неравенства (4)–(6), аналогично неравенству (10) получаем нижнюю оценку для коэффициента сжатия  $F_A(j-1)$  ( $1 < j \leq D$ ):

$$\begin{aligned}
 F_A(j-1) &> \frac{a}{\ln(D+1) + c_1} \left( \frac{\ln^2 j}{2} + c'_2 + \ln \ln(D+1) \ln j + (\ln(D+1) \right. \\
 &\quad \left. - \ln j - c_1)(\ln \ln(D+1) - \ln(\ln(D+1) - \ln j + c_1) + \delta'(j)) \right), \quad (11)
 \end{aligned}$$

где  $\delta'(j) = \frac{h'l'}{8} \ln 2$ .

**Утверждение.** Пусть

$$\tilde{\delta} = \max_j \delta(j), \quad \sqrt{\frac{\ln \ln(D+1)}{2}} < \tilde{\delta} \leq \frac{\ln \ln(D+1)}{\ln 2 + c_1}, \quad l > 2 \ln D.$$

Тогда для метода  $A$ , использующего при кодировании словарь из  $j-1 = \lfloor (D+1)e^{-m} \rfloor - 1$  слов, где

$$m = \frac{\ln \ln(D+1)}{\tilde{\delta}} - c_1, \quad (12)$$

справедливо неравенство

$$F_A(j-1) - F_A(D) < C' \frac{\ln \ln(D+1)}{\ln(D+1)},$$

где  $C' > 0$  — не зависящая от  $j$  константа.

**Доказательство.** Оценим коэффициент сжатия  $F_A(D)$  при  $j-1 = D$  (весь словарь кодируется по Шеннону):

$$\begin{aligned}
 F_A(D) &> -\frac{1}{l} \left( \sum_{i=1}^D p_i \log p_i \right) = \frac{k}{l \cdot \ln 2} \left( \sum_{i=1}^D \frac{\ln i}{i} - \ln k \sum_{i=1}^D \frac{1}{i} \right) \\
 &> \frac{a}{\ln(D+1) + c_1} \left( \frac{\ln^2(D+1)}{2} + c'_2 + \ln \ln(D+1) \ln(D+1) \right) \quad (13)
 \end{aligned}$$

и

$$F_A(D) < \frac{1}{l} \left( \sum_{i=1}^D p_i (-\log p_i + 1) \right) < \frac{a}{\ln(D+1)} \\ \times \left( \frac{\ln^2(D+1)}{2} + c_2 + \ln(\ln(D+1) + c_1)(\ln(D+1) + c_1) + \ln 2 \right). \quad (14)$$

Пусть  $\tilde{\delta} = \max_j \delta(j)$ , где  $\delta(j)$  определяется формулой (2). Из доказанной теоремы следует, что при  $j - 1 = \lfloor (D+1)e^{-m} \rfloor - 1$ , где  $m \in [\ln 2, \ln(D+1))$ , справедливо неравенство

$$F_A \left( \left\lfloor \frac{D+1}{e^m} \right\rfloor - 1 \right) < \frac{a \ln(D+1)}{2} + \frac{m^2 a}{2 \ln(D+1)} + a \ln \ln(D+1) + \alpha a \\ + \frac{2ac_1 \ln \ln(D+1)}{\ln(D+1)} - ma + \frac{\tilde{\delta} a(m+c_1)}{\ln(D+1)} + \frac{b}{\ln(D+1)}, \quad (15)$$

где  $b = a(c_2 + 2\alpha c_1 + 1)$ .

Рассмотрим правую часть неравенства (15). При  $\sqrt{\frac{\ln \ln(D+1)}{2}} < \tilde{\delta} \leq \frac{\ln \ln(D+1)}{\ln 2 + c_1}$  положим

$$m = \frac{\ln \ln(D+1)}{\tilde{\delta}} - c_1.$$

Тогда  $m < 2\tilde{\delta} - c_1$  и при данных  $\tilde{\delta}$  и  $m$  выполняются соотношения

$$\frac{\tilde{\delta} a(m+c_1)}{\ln(D+1)} > \frac{m^2 a}{2 \ln(D+1)}, \\ \frac{\tilde{\delta} a(m+c_1)}{\ln(D+1)} = \frac{a \ln \ln(D+1)}{\ln(D+1)}.$$

Учитывая (13) и (15), получаем

$$F_A \left( \left\lfloor \frac{D+1}{e^m} \right\rfloor - 1 \right) - F_A(D) < C' \frac{\ln \ln(D+1)}{\ln(D+1)},$$

где  $C' > 0$  — константа. Утверждение доказано.

Следует отметить, что для реальных текстов на естественных языках среднее количество информации в одной букве  $h \approx h'$ , причем  $h > 3$  и различно для разных языков (так, например, для слов русского языка  $h \approx 4,35$ , для слов английского языка  $h \approx 4,03$ ). Кроме того, как отмечено выше,  $l' > l > 2 \ln D$ , а  $\tilde{\delta} = \max_j \delta(j)$ , где  $\delta(j)$  определяется формулой (2). Поэтому для реальных текстов неравенство



$\sqrt{\frac{\ln \ln(D+1)}{2}} < \tilde{\delta} \leq \frac{\ln \ln(D+1)}{\ln 2 + c_1}$  выполняется при любых  $D$ . Так как  $m$ , определяемое формулой (11), удовлетворяет неравенству  $m \geq \ln 2$  и, следовательно,  $j \leq \lfloor (D+1)/2 \rfloor$ , то значение  $\tilde{\delta}$  можно вычислить по формуле  $\tilde{\delta} = \frac{(h'+1)l'}{8} \ln 2$ , где значения  $h'$  и  $l'$  определяются для  $\lfloor (D+1)/2 \rfloor$  слов второй части словаря. Заметим также, что величина  $\delta(j) = \frac{(h'+1)l'}{8} \ln 2$  при  $h > 3$ ,  $l > 2 \ln D$  и  $j < D$  удовлетворяет неравенствам  $\delta(j) > 0,69 \ln j$  и  $\delta(j) < \max_j \delta(j)$ . Эти неравенства показывают, что выражения, стоящие в правых частях неравенств (10) и (3) и зависящие от  $j$ , являются убывающими функциями.

Из доказанного утверждения следует, что в методе  $A$  при изменении объема словаря от  $j-1 = D$  до  $j-1 = \lfloor (D+1)e^{-m} \rfloor - 1$ , где  $m$  определяется формулой (11), коэффициент сжатия увеличивается на величину  $\beta(D) < C' \frac{\ln \ln(D+1)}{\ln(D+1)}$ , т. е.  $\beta(D) \rightarrow 0$  при  $D \rightarrow \infty$ . Таким образом, величина  $\lfloor (D+1)e^{-m} \rfloor - 1$  является оптимальным объемом словаря, который требуется хранить в памяти при сохранении эффективности сжатия.

### 3. Экспериментальные результаты

Наряду с полученным теоретическим результатом было проведено несколько экспериментов, где в качестве тестов использовались три различных текста (на русском и английском языках) с разными объемами словарей (см. таблицу):

- 1) document 1 — *документация* (на русском языке);
- 2) document 2 — *документация* (на английском языке);
- 3) works — *сочинения Шекспира (комедии)* на английском языке.

Отметим, что для проведения тестовых расчетов использовались обычные тексты на естественных языках, представляющие собой последовательность слов и разделителей (знаков препинания и пробелов). При этом кодирование текста проводилось с использованием двух различных и независимых словарей для слов и разделителей соответственно (этот подход часто используется на практике (см., например, [6])). Кодирование слов, встречающихся в тексте, проводилось по методу  $A$ , аналитические результаты коэффициентов сжатия получены по формулам (10), (11) при  $j-1 = \lfloor (D+1)e^{-m} \rfloor - 1$  и по формулам (13), (14) при  $j-1 = D$  (число  $m$  вычислялось по формуле (12), где значения  $\tilde{\delta}$ , как отмечено выше, определялись по формуле (2) для  $\lfloor (D+1)/2 \rfloor$  слов второй части словаря) и приведены в таблице в виде оценок. Отметим также, что при подсчете числа слов и составлении частотного словаря учитывалось число графически различных слов, а не лексем, что обусловлено использованием в методе  $A$  кодирования по Шеннону. В таблице

приведены результаты сжатия текстов по методу  $A$  при  $j - 1 = D$  и  $j - 1 = \lfloor (D + 1)e^{-m} \rfloor - 1$ , полученные аналитически и экспериментально (экспериментальные результаты приведены в скобках).

Результаты сжатия текстов на английском и русском языках по методу  $A$ , полученные аналитически и экспериментально

$N$	$D$ слов	$m$	Коэффициент сжатия ( $r$ , %)	
			$i - 1 = D$	$j - 1 = \lfloor (D + 1)e^{-m} \rfloor - 1$
1	2483	0,73	$19,3\% < r < 19,7\%$ (19,4%)	$22,4\% < r < 22,9\%$ (22,8%)
2	3721	0,81	$23,2\% < r < 23,8\%$ (23,6%)	$25,8\% < r < 26,3\%$ (26,0%)
3	4372	0,86	$23,7\% < r < 24,2\%$ (23,9%)	$26,1\% < r < 26,5\%$ (26,2%)

Из таблицы видно, что для рассматриваемых текстов  $m > 0,7$ , т. е. оптимальный объем словаря, который требуется хранить в памяти при использовании метода  $A$ , для данных текстов не превосходит  $\lfloor (D + 1)e^{-0,7} \rfloor$  и при изменении объема словаря от  $D$  до  $\lfloor (D + 1)e^{-m} \rfloor - 1$  коэффициент сжатия увеличивается незначительно. При этом результаты экспериментов хорошо согласуются с оценками, полученными аналитически.

Автор выражает благодарность В. Н. Потапову за ценные замечания, способствовавшие улучшению статьи.

## ЛИТЕРАТУРА

1. Пиотровский Р. Г. Текст, машина, человек. Л.: Наука, 1975.
2. Рябко Б. Я. Кодирование источника с неизвестными, но упорядоченными вероятностями // Проблемы передачи информации. 1979. Т. 15, вып. 2. С. 71–77.
3. Фихтенгольц Г. М. Курс дифференциального и интегрального исчисления. М.: Наука, 1966. Т. 2.
4. Хаффман Д. А. Метод построения кодов с минимальной избыточностью // Кибернетический сборник. М.: Мир, 1961. Вып. 3. С. 79–87.
5. Шеннон К. Работы по теории информации и кибернетике. М.: Изд-во иностр. лит., 1963.
6. Bell T. C., Cleary J. H., Witten I. H. Text compression. Englewood Cliffs, NJ: Prentice-Hall, 1990.
7. Bell T. C., Witten I. H., Cleary J. H. Modeling for text compression // ACM Comput. Surv. 1989. V. 21, N 4. P. 557–591.

8. **Mandelbrot B.** On the theory of word frequencies and on related Markovian models of discourse // The Structure of Language and its Mathematical Aspects. Providence, RI: Amer. Math. Soc., 1961. P. 190–219. (Proc. of Symposia on Applied Mathematics; V. 12).
9. **Moffat A. M.** Word based text compression // Software: Practice and Experience. 1989. V. 19, N 2. P. 185–198.
10. **Rissanen J., Langdon G. G.** Universal modeling and coding // IEEE Trans. Inform. Theory. 1981. V. 27, N 1. P. 12–23.
11. **Schwartz E. S.** A dictionary for minimal redundancy encoding // J. Assoc. Comput. Mach. 1963. V. 10, N 4. P. 413–439.
12. **Witten I. H., Bell T. C., Nevill C. G.** Models for compression in full-text retrieval systems // Proc. IEEE Data Compression Conference. Snowbird, Utah. 1991. P. 23–32.
13. **Zipf G. K.** Human behavior and the principle of least effort. Cambridge: Addison Wesley, 1949.
14. **Ziv J., Lempel A.** A universal algorithm for sequential data compression // IEEE Trans. Inform. Theory. 1977. V. 33, N 3. P. 337–343.
15. **Ziv J., Lempel A.** Compression of individual sequences via variable-rate coding // IEEE Trans. Inform. Theory. 1978. V. 24, N 5. P. 530–536.

Адрес автора:

Институт математики  
им. С. Л. Соболева СО РАН,  
пр. Академика Коптюга, 4,  
630090 Новосибирск, Россия.  
E-mail: ir@net.ict.nsc.ru

Статья поступила

14 сентября 1998 г.