

УДК 621.391

В.С. Чернега, доцент, канд. техн. наук

Севастопольский национальный технический университет

Студгородок, г. Севастополь, Украина, 99053

E-mail: v_chernega@rambler.ru

ОЦЕНКА СНИЖЕНИЯ КОЭФФИЦИЕНТА СЖАТИЯ ТЕКСТОВЫХ СООБЩЕНИЙ ВО ВРЕМЯ ПЕРЕХОДНОГО ПЕРИОДА РАБОТЫ КОМПРЕССОРА

Получены граничные и точные оценки коэффициента сжатия текстовых сообщений в процессе переходного периода для словарных методов компрессии данных семейства LZW. Показано, что путем предварительного заполнения таблицы кодирования возможно снижение среднего коэффициента сжатия текстовых сообщений и уменьшение объема выходного сообщения

Коэффициент сжатия текстового сообщения $K_{сж}$ определяется отношением количества битов $N_{сжб}$, затраченных на представление сжатого сообщения к общей длине исходного сообщения $L_{исх}$ в битах [1]

$$K_{сж} = N_{сжб} / L_{исх}. \quad (1)$$

Предположим, что в процессе сжатия входного сообщения было закодировано n_1 односимвольных строк, n_2 двухсимвольных и n_k k -символьных строк. Общее количество закодированных строк равно

$$N_{стр} = \sum_{i=1}^K n_i.$$

Здесь K – строка максимальной длины, закодированная в сообщении.

Тогда общая длина исходного сообщения определяется выражением

$$L_{исх} = n_1 l_1 + n_2 l_2 + \dots + n_k l_k = n_6 [n_1 + 2n_2 + 3n_3 + \dots + Kn_k], \quad (2)$$

где l_i – длина строки в символах; n_6 – количество битов, затрачиваемых на один символ исходного сообщения.

Подставив (2) в (1) и с учетом того, что объем сжатого сообщения в битах $N_{сжб}$ равен произведению количества закодированных строк $N_{стр}$ на количество битов, затрачиваемых на отображение строки $N_{бс}$, получим

$$K_{сж} = N_{бс} \cdot N_{стр} / \{n_6 [n_1 + 2n_2 + 3n_3 + \dots + Kn_k]\} = \chi N_{стр} / [n_1 + 2n_2 + 3n_3 + \dots + Kn_k], \quad (3)$$

где $\chi = N_{бс} / n_6$ – коэффициент расширения, зависящий от размера кодовой таблицы. На практике χ чаще всего принимает значения 1,25; 1,375 и 1,5 соответственно для 10-, 11- и 12-разрядной таблиц кодирования.

Выражение (3) можно записать в более компактном виде:

$$K_{сж} = \frac{\chi N_{cmp}}{\sum_{i=1}^K in_i}. \quad (4)$$

Из (4) видно, что если на начальном этапе в процессе кодирования были закодированы только односимвольные строки, т.е. $K=1$, то вместо сжатия происходит расширение выходного потока, объем которого по сравнению с входным увеличивается в χ раз. В этом и состоит основной недостаток *LZW*-способа сжатия.

Предположим, что в таблицу кодирования до начала процедуры сжатия занесены все *ASCII*-символы. Определим верхнюю границу коэффициента сжатия в зависимости от длины входного сообщения N_c . Для этого рассмотрим худший случай сжатия, когда каждый вновь поступивший на вход кодера символ приводит к несопоставимой строке, т.е. двухсимвольной, которой еще нет в кодовой таблице. При этом следует учесть то обстоятельство, что образование k -символьной строки (k -граммы) осуществляется путем объединения $(k-1)$ -символьной строки с последующим одиночным символом. Поэтому для заполнения кодовой таблицы всеми возможными k -граммами понадобится $(k-1)m^k + 1$ входных символов, при условии, что каждая вновь образованная из входного потока k -грамма отсутствует в кодовой таблице. Здесь m – количество символов алфавита.

С учетом вышесказанного заполнение таблицы всеми возможными двухсимвольными комбинациями (биграммами) произойдет после поступления на вход кодера $N_c = m^2 + 1$ символов. В этом случае общее число закодированных строк N_{cmp} равно количеству односимвольных строк n_1 . В связи с этим до поступления на вход кодера $m^2 + 1$ символов на основании (4) коэффициент сжатия $K_{сж}$ остается равным χ .

Начиная с $(m^2 + 2)$ -го входного символа происходит формирование трехсимвольных комбинаций (триграмм) и заполнение ими кодовой таблицы. В худшем случае, когда поступление на вход компрессора очередного символа приводит к образованию несопоставимой триграммы, то до поступления следующих $2m^3 + 1$ входных символов коэффициент сжатия (4) сохраняется неизменным, т.е.

$$K_{сж} = \frac{\chi N_{cmp}}{2n_2} = 0,5\chi. \quad (5)$$

Проводя аналогичные рассуждения, можно получить, что начиная с $(2m^3 + 2)$ -го до $(3m^4 + 1)$ -го входного символа $K_{сж} = 0,3\chi$, а с $(3m^4 + 2)$ -го по $(4m^5 + 1)$ -го символа $K_{сж} = 0,25\chi$. Учитывая, что входное сжимаемое сообщение редко превышает несколько гигабайт, то при $m = 256$ можно пола-

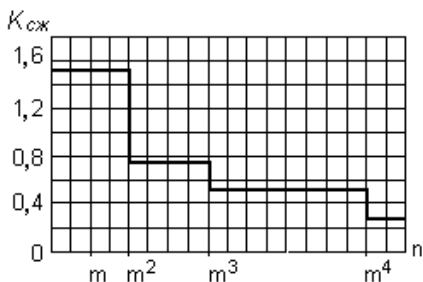


Рисунок 1 – Зависимость граничного коэффициента сжатия текстового сообщения от его длины для LZW-способа

гать, что минимальный коэффициент сжатия для LZW-способа будет не хуже $0,25\chi$.

Графическая зависимость граничного коэффициента сжатия $K_{сж}$ от длины входного сообщения n при $\chi=1,5$ имеет вид ломанной кривой, изображенной на рисунке 1.

Найдем выражение для точного значения коэффициента сжатия текстового сообщения для начальных условий, сформулированных при получении граничных оценок. Предположим, что входное сжимаемое сообщение имеет неограниченную длину, а формируемые k -граммы равновероятны, вероятности которых равны $1/(m^k)$. Разделив числитель и знаменатель выражения (3) на N_{cmp} , получим

$$K_{сж} = \frac{\chi}{P_{1cc} + 2P_{2cc} + 3P_{3cc} + \dots + KP_{kcc}} = \frac{\chi}{\sum_{i=1}^K (iP_{icc})}, \quad (6)$$

где $P_{icc} = \lim_{N_{cmp} \rightarrow \infty} \frac{n_i}{N_{cmp}}$ – вероятность сопоставления i -символьной строки,

т.е. вероятность того, что очередная i -символьная входная строка, сформированная из последовательности символов входного потока, находится в таблице кодирования.

Так как процедура сопоставления строк представляет собой полную группу событий, сумма вероятностей сопоставления строк удовлетворяет условию

$$\sum_{i=1}^K P_{icc} = 1.$$

В связи с тем, что очередная k -символьная строка (k -грамма) для всех $k \geq 2$ образуется путем объединения предыдущей $(k-1)$ -символьной комбинации с последующим одиночным символом входной строки, то занесение в процессе сжатия очередной k -граммы в таблицу кодирования увеличивает текущую оценку P_{kcc} на величину $1/(k-1)m^k$ на каждый входной символ.

Пусть на начальном этапе сжатия сообщения в кодовой таблице располагаются только одиночные ASCII-символы. В этом случае $P_{1cc}=1$, а $P_{icc}=0$ для всех $i=2,3,\dots,K$. Предположим, что вероятность появления очередного символа во входной строке не зависит от предыдущего символа. Тогда зане-

сение очередной биграммы в процессе кодирования в таблицу увеличивает P_{2cc} на величину $1/m^2$ и соответственно на такую же величину уменьшает P_{1cc} . С учетом того, что на начальном этапе процедуры компрессии $P_{1cc}=1$, а $P_{2cc}=0$, коэффициент сжатия (6) после поступления на вход кодера n -го символа для $n \leq m^2 + 1$ символов рассчитывается по формуле

$$K_{сж} = \frac{\lambda}{[P_{1cc} - (n/m^2)] + 2[P_{2cc} + (n/m^2)]} = \frac{\lambda}{1 + (n/m^2)}. \quad (7)$$

Из (7) следует, что при $n = m^2$ точная оценка коэффициента сжатия совпадает с граничной оценкой, полученной по формуле (5).

После заполнения таблицы всеми возможными двухсимвольными комбинациями начинается формирование триграмм и заполнение ими таблицы кодирования. В этом случае $P_{2cc}=1$, а $P_{1cc}=0$ для всех $i \neq 2$. Этот процесс, как отмечалось выше при определении граничной оценки, начнется не позднее, чем после поступления на компрессор (m^2+1) -входных символов.

Занесение очередной триграммы в процессе кодирования в таблицу увеличивает P_{3cc} на величину $1/(2m^3)$ на каждый входной символ и соответственно на такую же величину уменьшает P_{2cc} . С учетом того, что на начальном этапе процедуры компрессии $P_{2cc}=1$, а $P_{3cc}=0$, коэффициент сжатия после поступления на вход кодера n -го символа для $(m^2+1) < n \leq (2m^3+1)$ рассчитывается по формуле

$$K_{сж} = \frac{\lambda}{2[P_{2cc} - (n/2m^3)] + 3[P_{3cc} + (n/2m^3)]} = \frac{\lambda}{2 + (n/2m^3)}.$$

Проводя аналогичные рассуждения можно на основании (7) получить общую формулу зависимости оценки текущего коэффициента сжатия на этапе последовательного заполнения таблицы кодирования k -граммами от номера текущего символа входного потока. Она принимает вид

$$K_{сж-k} = \frac{\lambda}{(k-1) + \frac{n}{(k-1)m^k}} \quad \text{для} \quad m^{k-1} \leq n < (k-1)m^k.$$

График изменения текущего коэффициента сжатия текстового сообщения для *LZW*-способа на этапе переходного периода во время заполнения таблицы кодирования *k*-граммами, при условии, что сначала заполняется таблица кодирования только биграммами, затем триграммами и т.д., изображен на рисунке 2. На этом же рисунке штриховой ломаной показан график граничного коэффициента сжатия, полученный ранее.

Следует заметить, что сжатие входного сообщения в текст-ориентированном компрессоре на начальном этапе кодирования произойдет лишь в том случае, если в таблице кодирования до начала сжатия будут занесены *i*-символьные строки, которые встречаются во входном потоке. Поэтому для априорной оценки $K_{сж}$ необходимо учесть вероятности появления *i* – символьных строк на входе кодера и вероятности нахождения совпадающих с ними строк в кодовой таблице. С учетом этого обстоятельства априорная оценка коэффициента сжатия текст-ориентированного компрессора (6) определяется по формуле

$$K_{сж} = \frac{\gamma}{\sum_{i=1}^K (iP_{iCC} P_{iCT})}. \quad (8)$$

Здесь P_{iCT} – вероятность появления *i*-символьной строки во входном потоке сжимаемого сообщения; P_{iCC} – вероятность нахождения совпадающей *i*-символьной строки в текст-ориентированной таблице кодирования строк.

С учетом того, что входной поток состоит из последовательности одиночных символов, а в таблице кодирования всегда содержится таблица *ASCII*-символов, выражение (8) принимает вид

$$K_{сж} = \frac{\gamma}{1 + \sum_{i=2}^K i P_{iCC} P_{iCT}}. \quad (9)$$

Таким образом, размещая наиболее вероятные *k*-граммы в таблицы кодирования и декодирования словарных устройств сжатия данных можно уменьшить переходный период компрессора, в течение которого объем выходного потока превышает входной, и обеспечить с первых шагов работы компрессора сжатие сообщений. Вариант реализации такого устройства сжатия описан в [2]. Зная вероятности появления

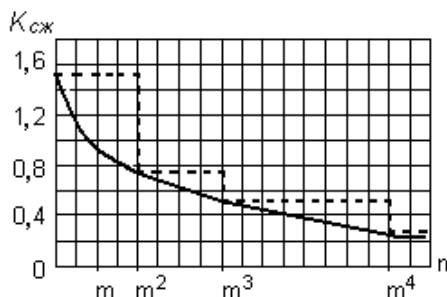


Рисунок 2 – Зависимость граничного коэффициента сжатия текстового сообщения от его длины для *LZW*-способа

отдельных k -грамм в текстовых сообщениях, можно получить априорную оценку коэффициента сжатия как в течение переходного периода, так и средний коэффициент сжатия.

Для вычисления конкретных значений коэффициентов сжатия по (9) необходимы знания плотностей распределения вероятностей $PiCП$ для текстовых сообщений различных стилей и тематической направленности. В дальнейшей работе предполагается провести серию экспериментальных исследований текстовых сообщений с целью получения таких распределений.

Библиографический список

1. Чернега В.С. Сжатие информации в компьютерных сетях / В.С. Чернега. — Севастополь: Изд-во СевГТУ, 1997. — 214 с.
2. Пат. 64907А. Україна, МКП G11 C19/00. Адаптивний пристрій для стиснення текстових повідомлень / В. Чернега, В. Шишкевич (Україна). — №2003010412; Заявл. 16.01.2003; Опубл. 15.03.2004, Бюл. № 3.

Поступила в редакцію 27.02.2006 г.