УДК 621.391

О КОМБИНАТОРНОМ ПОДХОДЕ К СЖАТИЮ ИНФОРМАЦИИ

А.А. Борисенко, профессор, Т.А. Протасова
Сумский государственный университет

Задача сжатия сообщений существовала на всем протяжении времени развития средств передачи и хранения информации. Главные решаемые ее вопросы – это повышение скорости передачи информации и уменьшение емкости памяти, которые на сегодня для цифровой техники являются актуальными. Причем часто задача сжатия ставится таким образом, что исключаются потери информации и повышается надежность ее передачи.

Разработка таких кодов, как код Морзе, Шеннона-Фано, Хаффмана и др., была направлена на достижение одной из этих целей [1]. Методы решения задач сжатия, кроме передачи информации, имеют применение и в других областях науки и техники – в задачах поиска, защиты данных от несанкционированного доступа, дискретной оптимизации, теории сложности.

Многообразие существующих методов сжатия информации поставило задачу оценки их эффективности, что способствовало созданию современной теории информации. Недостатком последней является то, что она направлена в основном на исследование источника информации, тогда как для полного анализа процесса сжатия требуется исследование и приемника информации.

В данной работе анализируются вопросы сжатия информации для конечных комбинаторных источников, т.е. источников, в которых с помощью алфавита $A = \{a_1, a_2, \ldots, a_m\}$ порождаются с равной вероятностью слова (сообщения) конечной длины $X = x_1, x_2, \ldots, x_n$, $x_i \in A$, $i = 1, 2, \ldots, n$, принадлежащие к подмножеству $R \subseteq A^n$ разрешенных слов X_j, $j = 1, 2, \ldots, M$, формируемых конечным предикатом P.

Конечным n-мерным предикатом P над конечным множеством A^n слов X, порождаемых алфавитом A, называется логическая функция $f = f(x_1, x_2, \ldots, x_n)$, принимающая значение 0 или 1 и задающая подмножество R разрешенных $X_{j,} j = 1, 2, \ldots, M$ и $A^n \setminus R$ запрещенных X_γ, $y = M + 1, \ldots, m^n$ слов.

Число слов, которое может потенциально генерировать источник информации, определяется мощностью исходного множества A^n и равно $N = m^n$.

Определение 1 Равновероятный источник информации, в котором все значения конечного предиката P равны 1 и соответственно $M = N$, называется **простейшим** источником информации.

В общем случае источник информации может генерировать $M < N$ разрешенных слов. В этом случае $N - M$ слов относятся к запрещенным. Таким образом, множество A^n предикатом P разбивается на подмножество разрешенных R и подмножеств $A^n \setminus R$ запрещенных слов. Соответственно простейший источник информации является частным случаем равновероятного источника, имеющего запрещенные слова. Информационная нагрузка в нем, приходящаяся на одно слово, измеряется числом содержащихся в нем знаков:

$$J^* = \log_m N = \log_m m^n = n. \qquad (1)$$

При необходимости перехода к измерению информации в битах следует воспользоваться соотношением

$$\log_2 N = \log_2 m \log_m N = n \log_2 m, \qquad (2)$$

тогда

$$J^* = \log_m N = \log_2 N / \log_2 m. \qquad (3)$$

Особенностью комбинаторных источников информации является наличие для них алгоритмов, порождающих разрешенные слова. Эти алгоритмы или формируют на каждом шаге разрешенное слово, или непосредственно по виду уже сформированного слова принимают решение о возможности передачи его приемнику.

Информационная нагрузка на слово комбинаторного источника

$$J = \log_m M \qquad (4)$$

меньше нагрузки на слово простейшего источника на величину

$$I = \log_m N - \log_m M = J^* - J, \qquad (5)$$

представляющую избыточность, содержащуюся в словах длины n простейшего источника. Наибольшей величины $I_{\max} = \log_m N - 1$ она достигает при $M = m$ и равна нулю при $M = N$.

Если приемник априорно содержит информацию I о том или ином сообщении, то необходимая величина дополнительно передаваемой ему информации для восстановления исходной информации J^* будет равна J.

При этом коэффициент сжатия

$$\mu = \frac{\log_m M}{\log_m N} \le 1 \qquad (6)$$

изменяется от 0 при $M = 1$ до 1 при $M = N$.

Так как число передаваемых сообщений комбинаторным источником M < N, то приемник информации должен априорно располагать информацией

$$J = MI = M(\log_m N - \log_m M)\,. \qquad (7)$$

Продифференцировав J по M и приравняв результат нулю, определяем, что своего максимального значения J достигает при $M = \frac{N}{e}$:

$$J = J_{max} = \frac{N}{e}\log_2 e, \qquad (8)$$

а минимального $J = 0$ в случае $M = N$. Для случая $M = 1$

$$J = \log_m N.$$

Сжатое слово содержит $n' < n$ букв и поэтому требует дешифрации на входе приемника, осуществляемой путем сравнения дешифрируемого слова с хранящимися в структуре дешифратора их образами.

Требуемая в этом случае емкость памяти дешифратора

$$\theta = M\log_m M. \tag{9}$$

С учетом памяти шифратора, хранящего информацию J, необходимую для восстановления сжатого слова с n' буквами до слова с n буквами, информационная емкость приемника

$$G = I + J = M\log_m N. \tag{10}$$

Из выражений (9, 10) путем вычисления отношения θ к G определяется коэффициент сжатия μ и затем кратность сжатия μ^{-1}.

На основе проведенного выше анализа систему сжатия дискретных сообщений представим в виде схемы, приведенной на рис.1, состоящей из комбинаторного источника информации (КИИ), генерирующего под воздействием предиката P только разрешенные слова; кодирующего устройства (КУ), устраняющего избыточную информацию из разрешенных слов; канала связи (КС), по которому передаются сжатые слова; приемника информации (ПИ), содержащего дешифратор (ДШ) для идентификации входных слов, и шифратора (Ш), вводящего недостающую избыточную информацию во входные слова, т.е. восстанавливающего переданное сжатое сообщение.

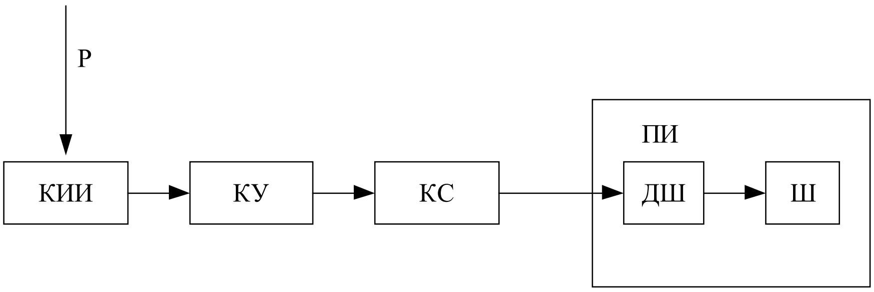

Рисунок 1- Обобщенная структура системы сжатия дискретных сообщений

Проиллюстрируем работу системы сжатия на конкретном примере.

Пример 1 Пусть дано

$$m = 2,\ \ n = 4,\ \ M = 4,$$

$$P = \bar{x}_1\bar{x}_2\bar{x}_3x_4 + \bar{x}_1x_2\bar{x}_3\bar{x}_4 + x_1\bar{x}_2x_3\bar{x}_4 + x_1x_2x_3x_4,$$

тогда

$$N = 2^4 = 16;\ J = n' = \log_2 4 = 2;\ I = \log_2 16 - \log_2 4 = 2;$$

$$\mu = \frac{J}{\log_2 N} = \frac{2}{\log_2 16} = 0,5;\ \mu^{-1} = 2;$$

$$J = 4(\log_2 16 - \log_2 4) = 8;\ J_{\max} = \frac{16}{2,87}\log_2 2,87 \approx 10;$$

$$\theta = 4\log_2 4 = 8;\ \ G = 4\log_2 16 = 16.$$

Величины J, I, M, N, k, n, n' и предикат P характеризуют источник; $J, J_{\max}, G, \theta$ – приемник; μ, μ^{-1} - эффективность сжатия.

На рис. 2 представлено дерево, отображающее разрешенные слова множества P, задаваемое предикатом P в соответствии с табл. 1.

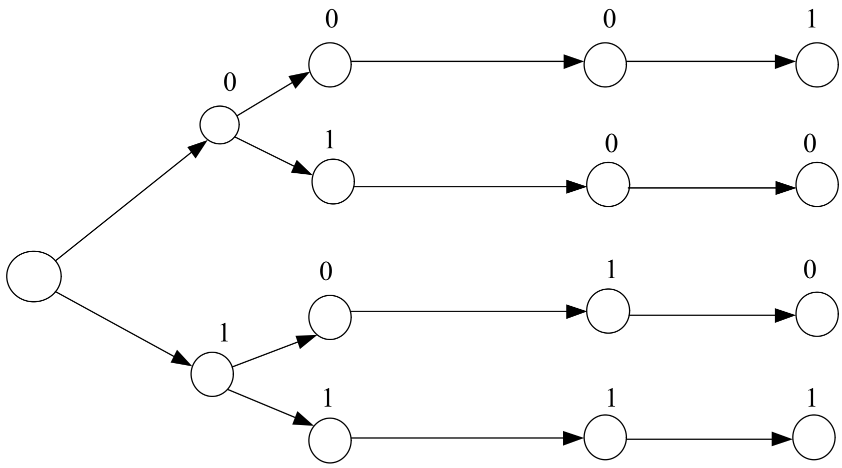

Рисунок 2 - Дерево разрешенных слов

Из рисунка видно, что для передачи информации достаточно использовать первые 2 яруса дерева, так как только они содержат разветвления из вершин. Остальная часть дерева не имеет разветвлений, и, следовательно, соответствующая избыточная информация хранится в шифраторе приемника.

Таблица 1 - Разрешенные слова

Ном.	x_1	x_2	x_3	x_4	P	P'	№	x_1	x_2	x_3	x_4	P	P'
0	0	0	0	0	0	0	8	1	0	0	0	0	0
1	0	0	0	1	1	1	9	1	0	0	1	0	0
2	0	0	1	0	0	0	10	1	0	1	0	1	0
3	0	0	1	1	0	0	11	1	0	1	1	0	0
4	0	1	0	0	1	1	12	1	1	0	0	0	0
5	0	1	0	1	0	0	13	1	1	0	1	0	0
6	0	1	1	0	0	0	14	1	1	1	0	0	0
7	0	1	1	1	0	0	15	1	1	1	1	1	1

Эта информация, отображаемая ветвями дерева третьего и четвертого ярусов, присоединяется шифратором к начальной префиксной части слова после ее дешифрации приемником. В результате происходит восстановление переданного простейшим источником информации слова.

В рассмотренном примере 1 величина М сознательно была взята равной $m^{n'}$, так как тогда сообщения могли быть переданы с помощью n' разрядов, не несущих в себе избыточной информации. В этом случае эффект сжатия достигает своего возможного максимума.

В случае $m^{n'-1} < M < m^{n'}$ для безошибочной передачи одного сообщения из M необходимо также использовать n' разрядов. Однако в этом случае в них появляется избыточная информация

$$i = n' - \log_m M. \qquad (11)$$

Теорема 1 Если $m^{n'} > M > m^{n'-1}$, то величина избыточной информации i определяется из неравенства 0< i< 1.

Доказательство. Из условия $M < m^{n'}$ следует, что $\log_m M < n'$, а из условия $M > m^{n'-1}$, что $\log_m M > n' - 1$, то есть

$$n' > \log_m M > n' - 1\,. \qquad (12)$$

Так как из (11) следует, что $n' = \log_m M + i$, то, подставив это выражение в неравенство (12), получим, что

$$\log_m M + i > \log_m M \ \text{ и } \ \log_m M > \log_m M + i - 1.$$

Отсюда следует, что $i > 0$ и $1 > i$, т.е. $0 < i < 1$. Теорема доказана.

Следствие. Если $M \neq 2n'$, то при сжатии полное устранение избыточной информации невозможно.

Наличие избыточной информации приводит к увеличению коэффициента сжатия до величины $\frac{n'}{n}$.

Для его снижения необходимо уменьшить значение m алфавита А в пределе до двух.

В таблице 1 представлен предикат

$$P' = \bar{x}_1\bar{x}_2\bar{x}_3x_4 + \bar{x}_1x_2\bar{x}_3\bar{x}_4 + x_1x_2x_3x_4\,,$$

число единичных значений которого не является целой степенью 2.

Он задает множество из $M = 3$ допустимых решений, для которых

$$i = 2 - \log_2 3 \cong 0,3\,.$$

Теоретически коэффициент сжатия μ в этом случае должен равняться отношению

$\frac{\log_2 3}{\log_2 4} = \frac{1,7}{4} = 0,425$, а реально это отношение равно $\frac{2}{4} = 0,5$.

Дерево кода для предиката P' приведено на рис. 3.

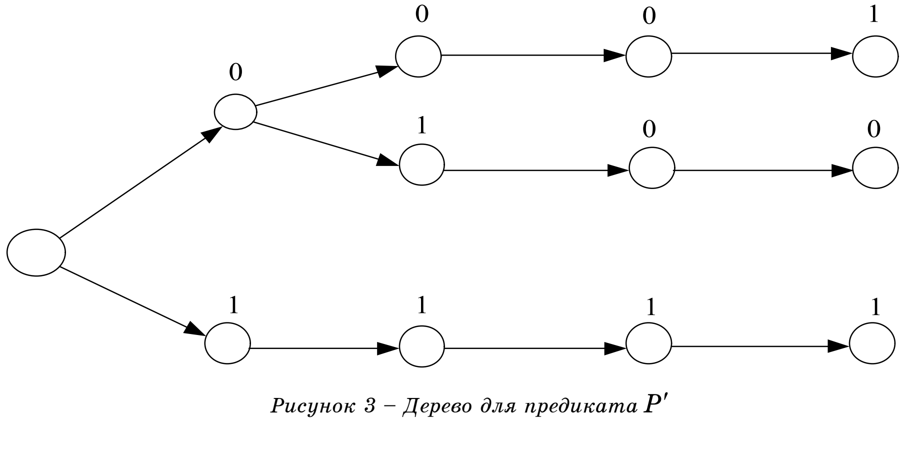

Рисунок 3 – Дерево для предиката P'

Этот код является префиксным и содержит слова 00, 01, 1, которые несут информационную нагрузку. Указанные буквы по каналу связи поступают на вход дешифратора приемника, который после их дешифрации возбуждает соответствующие входы шифратора, и тот восполняет недостающие в переданных словах буквы, восстанавливая таким образом слова, поступающие от простейшего источника информации.

Рассмотренная процедура характеризует важный метод сжатия, который, однако, не является единственным. На практике часто кодирующее устройство решает задачу преобразования генерируемого простейшим источником сообщения в другое, более короткое. Исходное сообщение длины n ставит в соответствие каждому разрешенному слову другое меньшей длины n' и в общем случае с другим числом m' букв в алфавите А. К такому способу сжатия, например, относится нумерационное кодирование [2]. В результате на приемном конце происходят дешифрация слов длины n' и выбор с его помощью из шифратора сообщений длины n.

Емкость шифратора для этого случая

$$G = N \log_m N . \tag{14}$$

Таким образом, проведенный анализ показывает, что для сжатия информации имеет значение характеристики не только источника информации, но и приемника. Причем эффект сжатия является результатом двустороннего взаимодействия источника и приемника информации.

SUMMARY

The paper has material on principles of information combinatory compression. For its analysis it is used combinatory information sources, each of which is of a finite predicate. The estimation of compression efficiency for codes, been formed different predicates, is carried out for the purpose of searching optimal values.

СПИСОК ЛИТЕРАТУРЫ

1. Березюк И.Т. и др. Кодирование информации (двоичные коды). – Харьков: Вища школа, 1978. - 252 с.
2. Амелькин В.А. Методы нумерационного кодирования. – Новосибирск: Наука, 1986.

Поступила в редакцию 11 мая 2006 г.

УДК 621.038

МЕТОД СЖАТИЯ ДВОИЧНЫХ КОДОВ НА ОСНОВЕ БИНОМИАЛЬНЫХ ЧИСЕЛ

В. Б. Чередниченко, ст. преподаватель
Харьковский национальный университет внутренних дел

В статье излагаются процедуры сжатия двоичных кодов, в которых используются методы двоичного биномиального счета. Приведены алгоритмы преобразований, которые отличаются простотой, надежностью и адаптивностью. Они иллюстрируются примерами и таблицами.