УДК 004.627
DOI: 10.15587/2312-8372.2019.169897

РАЗРАБОТКА МЕТОДОВ СЖИМАЮЩЕГО КОДИРОВАНИЯ ДАННЫХ НА ОСНОВЕ ДВОИЧНЫХ БИНОМИАЛЬНЫХ ЧИСЕЛ

Кулик И. А., Бережная О. В., Шевченко М. С.

1. Введение

Применение сжимающего кодирования данных имеет целью, прежде всего, повышение производительности информационных систем, к числу которых можно отнести компьютеризированные системы управления, системы архивирования данных, распределенные базы данных и т. п. Сжатие информации позволяет добиться производительности двумя путями [1, 2]:

1) уменьшение временных затрат на передачу информации каналами связи;

2) увеличение емкости памяти для хранения информации, которая применяется в системе.

Повышение производительности системы согласно первому направлению заключается в том, что за тот же период времени увеличивается количество информации, которая передается по каналу связи. Тем самым увеличивается пропускная способность канала, но, условно говоря, «виртуально» без изменения его реальных характеристик. Таким образом, информационная система получает возможность за тот же период времени обрабатывать большее количество данных.

Согласно второму направлению рост производительности информационной системы возможен за счет:

– увеличения количества сжатых данных в памяти той же емкости;

– уменьшения объема памяти для хранения сжатых данных с тем же самым количеством информации.

Таким образом, сжатие информации представляет собой малозатратный в экономическом плане способ увеличения производительности обработки данных в различных информационных системах. В связи с этим, разработка и исследование методов и алгоритмов сжатия является актуальной задачей, имеющей не только научную, но и технико-экономическую ценность.

2. Объект исследования и его технологический аудит

Объект исследования – методы сжатия данных без потерь, предназначенные для устранения информационной избыточности сообщений и минимизации их длины, т. е. сокращения разрядности их представления.

Методы сжимающего кодирования характеризуются [1–3]:

– типом сжимаемых данных;

– моделями процессов сжатия и восстановления на основе аналитических соотношений или статистических зависимостей;

– алгоритмами (схемами устройств) сжатия и восстановления данных.

Использование сжатия данных в информационных системах позволяет:

1) снизить стоимостные затраты на их создание за счет удешевления устройств памяти и возможности применения недорогих низкоскоростных каналов связи;

2) повысить их производительность за счет увеличения скорости передачи информации по уже существующим каналам и увеличения количества хранимых данных в уже имеющейся памяти без существенных затрат на их модернизацию.

При этом особый интерес вызывают методы, над результатами сжатия которых можно проводить вычислительные операции без обратного восстановления к исходным последовательностям.

Одними из самых проблемных мест внедрения сжатия являются высокие требования к вычислительным ресурсам системы, значительные затраты при его реализации и невысокая скорость кодирования/декодирования. Для улучшения указанных характеристик сжатия данных в информационных системах предлагаются к рассмотрению методы сжатия на основе двоичных биномиальных чисел.

3. Цель и задачи исследования

Целью данной работы является минимизация времени сжатия и восстановления двоичных последовательностей при ограничении на объем аппаратно-программных затрат при практической реализации методов сжатия на основе двоичных биномиальных чисел.

Задачи исследования работы формулируются следующим образом:

1. Построение математической модели метода сжатия двоичных *n*-разрядных равновесных комбинаций с фиксированным числом $0 < k < n$ единиц на основе двоичных (n,k)-биномиальных чисел.

2. Построение математической модели обобщенного метода сжатия двоичных *n*-разрядных последовательностей с переменным числом $0 \le k \le n$ единиц на основе двоичных (n,k)-биномиальных чисел.

4. Исследование существующих решений проблемы

Огромную роль, как и прежде, в информационных системах играют методы сжатия без потерь информации, что объясняется тем, что [3, 4]:

1) сжимающее кодирование без потерь является более универсальным решением в виду того, что оно, как правило, не ориентируется на конкретный, узко специализированный тип информации, а может работать сразу со множеством типом данных;

2) значительное количество методов сжатия без потерь имеют дело с двоичным представлением информации, что дополнительно подчеркивает универсальный характер таких методов;

3) для многих информационных систем применение сжатия без потерь является безальтернативным в виду неопределенности ценностного критерия используемых данных или отсутствия психофизиологического фактора восприятия информации. К таким системам, например, можно отнести системы автоматизации научного эксперимента, системы автоматического управления, распределенные базы данных.

Перспективными являются методы и алгоритмы сжатия без потерь на основе структурных чисел [5, 6], которые генерируются структурными системами счисления. Основной идеей сжимающего преобразования информации на основе структурных систем счисления является то, что в структуре любой кодовой последовательности можно выявить соответствующее структурное число. Таким образом, поставив в соответствие исходным кодовым последовательностям их структурные числа можно существенно уменьшить информационную избыточность.

Особое место среди структурных систем счисления занимают двоичные биномиальные системы с параметрами n и k, которые генерируют двоичные (n,k)-биномиальные числа [6, 7]. Числовая функция двоичной (n,k)-биномиальной, которая определяет десятичный количественный эквивалент $F_j = \mathrm{dec}\, X_j$ двоичного биномиального числа X_j, имеет вид [6, 7]:

$$F_j = \mathrm{dec}\ X_j = \sum_{i=1}^{r} x_i C_{n-i}^{k-q_i},$$

где $X_j = x_1 x_2 \ldots x_i \ldots x_r$, $r < n$, $X_j \in X$, $j = 1, 2, \ldots, C_n^k$;

q_i – сумма единичных цифр x_i от первого разряда до (i–1)-го включительно:

$$q_i = \sum_{t=1}^{i-1} x_t ,\ q_i \le k.$$

Генерируемые (n,k)-биномиальные числа X_j должны удовлетворять следующим кодообразующим ограничениям [6, 8]:

$$\begin{cases} l = n - k, \\ x_r = 0, \end{cases} \quad \text{и} \quad \begin{cases} q = k, \\ x_r = 1, \end{cases}$$

где q и l – числа единиц и нулей в двоичном биномиальном числе X_j.

В работах [9, 10] приводятся способы кодирования источников информации с использованием комбинаторной системы счисления, но структурные числа при кодировании не используются, а осуществляется непосредственный переход от сжимаемой последовательности к двоичному номеру. Такой переход характеризуется сложностью вычислений. Хотя в работе [9] предпринимается попытка рассмотрения биномиальных кодов, но без системного подхода при отсутствии кодообразующих ограничений.

В работах [11, 12] рассматриваются методы сжимающего кодирования на основе нумерационных функций, но, с одной стороны, данные методы обладают существенной вычислительной сложностью, а, с другой, числа при рассмотрении этих методов не рассматриваются как таковые. Кроме того, сжимаемые последовательности характеризуются сложными комбинаторными ограничениями, что ограничивает использование предлагаемых методов.

В работе [13] приводятся тезисно модель сжатия на основе двоичных биномиальных чисел, но только для одного типа двоичных последовательностей – равновесных комбинаций, которые имеют весьма ограниченное распространение.

Теоретической и практической основой данной работы является тот факт, что в основе любых сжимаемых двоичных последовательностей можно выявить соответствующие им структурные числа, генерируемые структурными системами счисления [6]. Такой подход к разработке методов сжимающего кодирования характеризуется универсальностью решения задач, а также получением сжатых образов, обладающих числовыми характеристиками. Это позволяет в случае необходимости проводить их вычислительную обработку без их восстановления.

В приведенной работе в качестве сжимаемых последовательностей рассматриваются двоичные равновесные комбинации, которые можно получить из любой двоичной последовательности путем простого подсчета числа содержащихся в ней двоичных единиц.

5. Методы исследования

Реализация биномиального отображения $\varphi^{-1}: Y \to X$ в рамках построения математической модели перечисления для исходных двоичных *n*-разрядных последовательностей вида $Y_j \in Y[n,k]$ представляет собой сжатие:

$$f_b : Y[n,k] \to X[n,k], \tag{1}$$

равновесных комбинаций Y_j на основе двоичных (n,k)-биномиальных чисел $X_j \in X[n,k]$ [14]. В свою очередь, реализация биномиального отображения $\varphi: X \to Y$ в рамках построения математической модели генерирования для $Y_j \in Y[n,k]$ означает восстановление:

$$f_b^{-1} : X[n,k] \to Y[n,k], \tag{2}$$

исходных равновесных комбинаций Y_j на основе двоичных (n,k)-биномиальных чисел $X_j \in X[n,k]$.

Нижеследующая теорема 1, которую приведем без доказательства, приводит свойства отображения f_b и способ его практической реализации. Условимся операцию декатенации далее обозначать символом вида «/».

Теорема 1. Всякой двоичной последовательности $Y_j = y_1 y_2 ... y_i ... y_n$, $Y_j \in Y[n,k]$, $j = \overline{1, C_n^k}$, составленной из *n* разрядов y_i, сумма значений которых равна k, можно поставить в соответствие единственное двоичное (n,k)-биномиальное число $X_j = x_1 x_2 ... x_i ... x_r$, $X_j \in X[n,k]$, $r < n$, с помощью функции $X_j = f_b(Y_j)$ вида:

$$X_j = x_1 x_2 \ldots x_i \ldots x_r = \begin{cases} y_1 y_2 \ldots y_i \ldots y_{n-1} 0 / 00 \ldots 0, \\ y_1 y_2 \ldots y_i \ldots y_{n-1} 1 / 11 \ldots 1. \end{cases} \qquad (3)$$

Таким образом, отображение $f_b : Y[n,k] \to X[n,k]$, которое задается теоремой 1, будем называть методом сжатия на основе двоичных (n,k)-биномиальных чисел или биномиальным сжатием.

Моделирование процесса сжатия f_b двоичных равновесных комбинаций $Y_j = y_1 y_2 \ldots y_i \ldots y_n$ на основе двоичных (n,k)-биномиальных чисел X_j, используя теорему 1 и функцию (3), состоит из следующих этапов.

Этап 1. Определяется в n-разрядной равновесной комбинации $Y_j = y_1 y_2 \ldots y_i \ldots y_n$, имеющей число k единиц, значение последнего разряда y_n.

Этап 2. Если $y_n = 0$, то:

$$X_j = Y_j / 00 \ldots 0 = y_1 y_2 \ldots y_i \ldots y_{n-1} 0 / 00 \ldots 0 = x_1 x_2 \ldots x_i \ldots x_{r-1} 1,$$

т. е. от комбинации $Y_j = y_1 y_2 \ldots y_i \ldots y_{n-1} 0$ отбрасываются все нулевые разряды, начиная с $y_n = 0$, до появления первой двоичной единицы $y_r = 1$, которая будет представлять значение последнего разряда $x_r = y_r = 1$ искомого (n,k)-биномиального числа $X_j = x_1 x_2 \ldots x_i \ldots x_{r-1} 1$. В противном случае:

$$X_j = Y_j / 11 \ldots 1 = y_1 y_2 \ldots y_i \ldots y_{n-1} 1 / 11 \ldots 1 = x_1 x_2 \ldots x_i \ldots x_{r-1} 0,$$

т. е. от комбинации $Y_j = y_1 y_2 \ldots y_i \ldots y_{n-1} 1$ отбрасываются все единичные разряды, начиная с $y_n = 1$, до появления первого двоичного нуля $y_r = 0$, который будет представлять значение последнего разряда $x_r = y_r = 0$ искомого (n,k)-биномиального числа $X_j = x_1 x_2 \ldots x_i \ldots x_{r-1} 0$. При этом в обоих случаях значения остальных разрядов остаются без изменений: $x_1 = y_1$, $x_2 = y_2, \ldots, x_{r-1} = y_{r-1}$.

Теорема 2, которую приведем без доказательства, приводит свойства отображения f_b^{-1} и способ его практической реализации. Введем теперь в рассмотрение операцию конкатенации, которую обозначим как «$++$». Данное действие является обратным по отношению к операции декатенации.

Теорема 2 Всякому двоичному (n,k)-биномиальному числу $X_j = x_1 x_2 \ldots x_i \ldots x_r$, $X_j \in X[n,k]$, $r < n$, можно поставить в соответствие единственную двоичную равновесную комбинацию $Y_j = y_1 y_2 \ldots y_i \ldots y_n$, $Y_j \in Y[n,k]$, $j = \overline{1, C_n^k}$, составленную из n разрядов y_i, сумма значений которых равна k, с помощью функции $Y_j = f_b^{-1}(X_j)$ вида:

$$Y_j = y_1 y_2 \ldots y_i \ldots y_n = \begin{cases} x_1 x_2 \ldots x_i \ldots x_{r-1} 0 + +11 \ldots 1, \\ x_1 x_2 \ldots x_i \ldots x_{r-1} 1 + +00 \ldots 0. \end{cases} \qquad (4)$$

Моделирование процесса восстановления f_b^{-1} двоичных равновесных комбинаций $Y_j = y_1 y_2 ... y_i ... y_n$ на основе двоичных (n,k)-биномиальных чисел X_j, используя теорему 2 и функцию (4), состоит из следующих этапов.

Этап 1. Определяется в двоичном (n,k)-биномиальном r-разрядном числе $X_j = x_1 x_2 ... x_i ... x_r$, $X_j \in X[n,k]$, $r < n$, значение последнего разряда x_r.

Этап 2. Если $x_r = 0$, то:

$$Y_j = X_j + +11...1 = x_1 x_2 ... x_i ... x_{r-1} 0 + +11...1 = y_1 y_2 ... y_i ... y_{n-1} 1,$$

т. е. к двоичному биномиальному числу $X_j = x_1 x_2 ... x_i ... x_{r-1} 0$ присоединяются единичные разряды $11...1$: $y_{r+1} = y_{r+2} = ... = y_n = 1$ так, чтобы общее количество разрядов искомой двоичной равновесной комбинации Y_j составило n, $Y_j \in Y[n,k]$. В противном случае:

$$Y_j = X_j + +00...0 = x_1 x_2 ... x_i ... x_{r-1} 1 + +00...0 = y_1 y_2 ... y_i ... y_{n-1} 0,$$

т. е. к двоичному биномиальному числу $X_j = x_1 x_2 ... x_i ... x_{r-1} 1$ присоединяются нулевые разряды $00...0$: $y_{r+1} = y_{r+2} = ... = y_n = 0$ так, чтобы общее количество разрядов искомой двоичной равновесной комбинации Y_j составило n, $Y_j \in Y[n,k]$. При этом в обоих случаях значения остальных разрядов остаются без изменений: $x_1 = y_1$, $x_2 = y_2$, $\ldots$, $y_r = x_r = 1$.

Отображение $f_b : Y[n,k] \to X[n,k]$ является биективным, поскольку соответствия $X_j = f_b(Y_j)$ и $Y_j = f_b^{-1}(X_j)$ есть функциональные (теоремы 1 и 2), т. е. каждый элемент $Y_j \in Y[n,k]$ имеет единственный образ $X_j \in X[n,k]$, а каждый элемент $X_j \in X[n,k]$ – единственный прообраз $Y_j \in Y[n,k]$.

Отображения вида f_b и f_b^{-1} оперируют с двоичными n-разрядными равновесными комбинациями $Y_j \in Y[n,k]$, т. е. число k единиц есть величина постоянная. При этом следует учесть, что, исходя из свойств двоичных (n,k)-биномиальных чисел [2], $k_{\min} = 1$ и $k_{\max} = n-1$. Более общим является случай, когда k может принимать любые значения из заданного диапазона $0 \le k \le n$, а сжимаемый массив A представляет собой множество:

$$A = \bigcup_{k=0}^{n} Y[n,k] \text{ и } A_j \in A = \{0,1\}^n, \quad j = \overline{1,2^n},$$

двоичных n-разрядных последовательностей A_j, для которых отсутствует ограничение вида по числу k единиц.

С учетом того, что двоичные (n,k)-биномиальные числа X_j являются префиксными только для постоянного значения k [6], то для однозначного восстановления $A_j \in A = \{0,1\}^n$ из чисел X_j следует дополнительно использовать значение k единиц, выраженного в двоичном виде $\mathrm{Bin}k$.

При сжатии A_j необходимо использовать функцию f_w, ставящую в соответствие исходной последовательности A_j выборку (k, Y_j), где $Y_j = A_j$. Далее, если полученное значение k удовлетворяет неравенству $0 < k < n$, то для сжатия равновесной комбинации Y_j, соответствующей A_j, используется кодирование f_b на основе двоичных биномиальных чисел. При этом к сжатым комбинациям для однозначного восстановления добавляется $\mathrm{Bin}k$, т. е. выполняется дополнительно кодирование вида f_k. Если же значение k удовлетворяет системе равенств $(k=0) \vee (k=n)$, то кодируемая результирующая комбинация будет состоять только из $\mathrm{Bin}k$, т. е. используется единственно метод кодирования f_k.

Таким образом, рассмотрим отображение вида:

$$f_{bg} : A \to Z,$$

которое задается соответствующей функцией:

$$Z_j = f_{bg}(A_j),$$

где $A_j = a_1 a_2 \ldots a_i \ldots a_n$, $A_j \in A = \{0,1\}^n$, $Z_j = (\mathrm{Bin}\,k, X_j)$ или $Z_j = \mathrm{Bin}\,k$, $Z_j \in Z$, $j = \overline{1, 2^n}$. Нижеследующая теорема 3, которую приведем без доказательства, приводит свойства отображения f_{bg} и способ его реализации.

Теорема 3. Всякой двоичной последовательности $A_j = a_1 a_2 \ldots a_i \ldots a_n$, $A_j \in A = \{0,1\}^n$, $j = \overline{1, 2^n}$, можно поставить во взаимно однозначное соответствие двоичную комбинацию $Z_j \in Z$ следующего вида:

1) если $0 < k < n$, то:

$$Z_j = \mathrm{Bin}\,k {+}{+} X_j, \tag{5}$$

где $k = \sum_{i=1}^{n} a_i$ и $X_j = f_b(Y_j)$, $X_j \in X[n,k]$, $Y_j \in Y[n,k]$;

2) в противном случае, если $(k=0) \vee (k=n)$, то:

$$Z_j = \mathrm{Bin}\,k. \tag{6}$$

Отображение $f_{bg} : A \to Z$ также является биективным, поскольку каждый элемент A_j имеет единственный образ, а каждый элемент Z_j − единственный прообраз для всех $A_j \in A$ и $Z_j \in Z$.

Способы практической реализации отображений f_{bg} и f_{bg}^{-1}, указанных в теореме 3 в (5) и (6), могут быть различными. Выбранные подходы к построению f_{bg} и f_{bg}^{-1}, методы кодирования и формируемые для них модели процессов в конечном итоге влияют на быстродействие и объем аппаратно-программных затрат при их практической реализации.

Отображение $f_{bg}: A \to Z$ называется обобщенным методом сжатия на основе двоичных (n,k)-биномиальных чисел (или обобщенным биномиальным сжатием), которое задается следующей сложной функцией вида:

$$f_{bg} = \begin{cases} f_k \circ f_w,\ (k=0) \vee (k=n), \\ f_k \circ f_b \circ f_w,\ 0 < k < n, \end{cases} \tag{7}$$

где Z − множество результирующих последовательностей Z_j:

$$\begin{aligned} &Z = Z_o \cup Z_b, \\ &Z_o = Q \times \varnothing = \{Z_j / Z_j = \mathrm{Bin}\,k, (k=0) \vee (k=n)\}, \\ &Z_b = Q \times X[n,k] = \{Z_j / Z_j = (\mathrm{Bin}\,k, X_j), 0 < k < n-1\}, \\ &Q = \{\mathrm{Bin}\,k / 0 \le k \le n\},\ Y_i \in Y[n,k],\ Y_j = f_w(A_j); \end{aligned}$$

f_k − функция $Z_j = f_k(A_j)$, ставящая в соответствие исходной последовательности $Y_j = A_j$ двоичную запись $\mathrm{Bin}\,k$ числа k единиц, где $(k=0) \vee (k=n)$, и которая определяет отображение вида:

$$f_k : Y[n,k] \to Z_o,$$

или функция $Z_j = f_k(X_j)$, ставящая в соответствие двоичному (n,k)-биномиальному числу результирующую последовательность $Z_j = \mathrm{Bin}\,k + +X_j$, если $0 < k < n$, и которая определяет отображение вида:

$$f_k : X[n,k] \to Z_b;$$

f_w − функция $Y_j = f_w(A_j)$, ставящая в соответствие исходной последовательности A_j упорядоченную выборку вида (k, Y_i), где $Y_j = A_j$, и определяющая отображение вида:

$$\begin{aligned} &f_w : A \to M, \\ &M = \{(k, Y_i) / 0 < k < n, Y_j \in Y[n,k]\}. \end{aligned}$$

В свою очередь, обратное отображение $f_{bg}^{-1} : Z \to A$, которое задается обратной сложной функцией:

$$f_{bg}^{-1} = \begin{cases} f_w^{-1} \circ f_k^{-1}, \ (k=0) \vee (k=n), \\ f_w^{-1} \circ f_b^{-1} \circ f_k^{-1}, \ 0 < k < n, \end{cases} \tag{8}$$

представляет собой восстановление с учетом имеющегося значения k исходных двоичных последовательностей A_j. В случае $0 < k < n$ восстановление A_j производится на основе $\mathrm{Bin}\,k$ и двоичных (n,k)-биномиальных чисел $X_j \in X[n,k]$, а в случае $k = 0$ или $k = n$ – на основе $\mathrm{Bin}\,k$ путем генерирования n нулей или единиц, соответственно. Такой вид восстановления будем называть обобщенным биномиальным.

Способы реализации сложных функций (7), (8) на подобласти определения $k \in \{1,2,...,n-1\}$ для сжатия f_{bg} являются аналогичными как для функций (3), (4) сжатия f_b на всей области значений k, т. е. на всем диапазоне $0 < k < n$. Способы реализации сложных функций (7), (8) на подобласти определения $k \in \{0,n\}$ определяются простой операцией вычисления k единиц при сжатии f_{bg} и формированием исходной нулевой или единичной последовательности A_j при восстановлении f_{bg}^{-1}.

Из теорем 1 и 2, которые обосновывают методы реализации соответствия, формулируемого теоремой 3, а также из самой теоремы 3 следуют модели процессов обобщенного биномиального сжатия f_{bg} и восстановления f_{bg}^{-1} двоичных последовательностей.

Моделирование процесса сжатия f_{bg} двоичных последовательностей $A_j = a_1 a_2 ... a_i ... a_n$, $A_j \in A = \{0,1\}^n$, $j = \overline{1,2^n}$ осуществляется на основе теорем 1 и 3, функции (3) и состоит из следующих этапов.

Этап 1. Определяется количество s разрядов для двоичного представления $\mathrm{Bin}\,k$ числа k единиц, $0 \le k \le n$, исходной n-разрядной последовательности $A_j = a_1 a_2 ... a_i ... a_n$:

$$s = \lceil \log_2 (n+1) \rceil.$$

Этап 2. Производится вычисление числа k двоичных единиц в исходной n-разрядной последовательности $A_j = a_1 a_2 ... a_i ... a_n$:

$$k = \sum_{i=1}^{n} a_i,$$

тем самым реализуя функцию $f_w(A_j) = (k, Y_j)$ и определяя класс равновесных комбинаций $Y[n,k]$, к которому относится A_j, $A_j = Y_j \in Y[n,k]$.

Этап 3. Выполняется преобразование числа k единиц к его двоичному виду $\mathrm{Bin}\,k$, состоящему из s разрядов.

Этап 4. Если число k удовлетворяет системе равенств $(k=0) \vee (k=n)$, т. е. $k \in \{0,n\}$, то результирующей будет являться комбинация вида $Z_j = \mathrm{Bin}\,k$, $Z_j \in Z_o$.

В противном случае имеющееся значение n и вычисленное значение k представляют собой параметры двоичной (n,k)-биномиальной системы счисления и осуществляется переход к последующим этапам для реализации кодирования $f_k\left(f_b\left(Y_j\right)\right)=Z_j$.

Этап 5. Определяется в n-разрядной равновесной комбинации $Y_j=y_1y_2...y_i...y_n$, имеющей число k единиц, значение последнего разряда y_n.

Этап 6. Если $y_n=0$, то:

$$X_j=Y_j/00...0=y_1y_2...y_i...y_{n-1}0/00...0=x_1x_2...x_i...x_{r-1}1,$$

т. е. от комбинации $Y_j=y_1y_2...y_i...y_{n-1}0$ отбрасываются все нулевые разряды, начиная с $y_n=0$, до появления первой двоичной единицы $y_r=1$, которая будет представлять значение последнего разряда $x_r=y_r=1$ (n,k)-биномиального числа $X_j=x_1x_2...x_i...x_{r-1}1$. В противном случае:

$$X_j=Y_j/11...1=y_1y_2...y_i...y_{n-1}1/11...1=x_1x_2...x_i...x_{r-1}0,$$

т. е. от комбинации $Y_j=y_1y_2...y_i...y_{n-1}1$ отбрасываются все единичные разряды, начиная с $y_n=1$, до появления первого двоичного нуля $y_r=0$, который будет представлять значение последнего разряда $x_r=y_r=0$ (n,k)-биномиального числа $X_j=x_1x_2...x_i...x_{r-1}0$. При этом в обоих случаях значения остальных разрядов остаются без изменений: $x_1=y_1$, $x_2=y_2$, …, $x_{r-1}=y_{r-1}$.

Этап 7. Выполняется конкатенация двоичных значений $\mathrm{Bin}\,k$ и (n,k)-биномиального числа X_j, т. е. кодирование вида $f_k\left(X_j\right)=Z_j$ для случая $0<k<n$ или $k\in\{1,2,...,n-1\}$:

$$Z_j=\mathrm{Bin}\,k++X_j,$$

тем самым получая результирующую комбинацию $Z_j\in Z_b$.

Моделирование процесса восстановления f_{bg}^{-1} последовательностей $A_j=a_1a_2...a_i...a_n$, $A_j\in A=\{0,1\}^n$, $j=\overline{1,2^n}$ из комбинаций-образов Z_j, $Z_j\in Z_o\cup Z_b$ осуществляется на основе теорем 2, 3 и функции (4). В модели, реализующей f_{bg}^{-1}, в качестве основных шагов используются этапы из модели восстановления f_b^{-1} при известном $\mathrm{Bin}\,k$ и вводятся шаги, формирующие последовательности только из нулей или только из единиц, когда $k=0$ или $k=n$ соответственно.

6. Результаты исследования

Результаты отображения $f_b:Y[n,k]\rightarrow X[n,k]$ и, следовательно, $f_b^{-1}:X[n,k]\rightarrow Y[n,k]$ при $n=8$ и $k=2$ для некоторых комбинаций приведены в табл. 1. Затененные ячейки в табл. 1 означают отбрасываемые разряды равновесных комбинаций Y_j в соответствии с (3) или прибавляемые разряды к двоичным биномиальным числам X_j в соответствии с (4), $1 \le j \le C_8^2$. Средний коэффициент сжатия равновесных комбинаций, составляющих массив данных в табл. 1, имеет значение 1,94. Среднее время сжатия f_b определяется как среднее время числа машинных тактов, требуемых для поиска и отбрасывания двоичных разрядов, и составляет 3,2 машинных такта.

Таблица 1

Соответствие между $Y[8,2]$ и $X[8,2]$ для некоторых равновесных комбинаций

j	Равновесный код $Y[8,2]$								Множество $X[8,2]$						
0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	
9	0	0	0	1	1	0	0	0	0	0	0	1	1		
14	0	0	1	1	0	0	0	0	0	0	1	1			
15	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0
18	0	1	0	0	1	0	0	0	0	1	0	0	1		
19	0	1	0	1	0	0	0	0	0	1	0	1			
20	0	1	1	0	0	0	0	0	0	1	1				
22	1	0	0	0	0	0	1	0	1	0	0	0	0	0	1
27	1	1	0	0	0	0	0	0	1	1					

В качестве примера проведем сжатие f_b исходной равновесной комбинации $Y_i \in Y[16,4]$ – $Y_j = 1000100001100000$. В соответствии с (3) получаем сжатое изображение Y_j в виде $X_j \in X[16,4]$:

$$X_j = 1000100001100000/00000 = 10001000011.$$

При этом длина результирующей комбинации сокращается с 16 двоичных разрядов до 11, т. е. в 1,45 раза.

Восстановление f_b^{-1} равновесной комбинации $Y_i \in Y[20,16]$, имея двоичное биномиальное число $X_j \in X[20,16]$ – $X_j = 010010$, производится согласно (4) следующим образом:

$$Y_j = 010010 + +111111111111 = 010010111111111111.$$

Результаты отображений $f_{bg} : A \to Z$ и $f_{bg}^{-1} : Z \to A$ для $n = 24$, где $A_j \in A = \{0,1\}^{24}$, $1 \le j \le 2^{24}$, приведены в табл. 2. Если $0 < k < 23$, то применяется сложная функция кодирования $f_k \circ f_b \circ f_w$, в противном случае, т. е. при $(k = 0) \vee (k = 24)$ – сложная функция кодирования $f_k \circ f_w$. Отношение длин двоичного представления A_j и Z_j для табл. 2 изменяется в пределах от 0,86 до 4,8, а их среднее значение для всей рассматриваемой таблицы составляет приблизительно 2,15. Среднее время T обобщенного биномиального сжатия f_{bg} для информационного массива табл. 2 определяется теперь как сумма времени t_1, необходимого для подсчета числа k единиц, и среднего времени t_2 сжатия массива. Принимая для наиболее простого, но наиболее затратного способа подсчета единиц, $t_1 = n = 24$, получаем $T = t_1 + t_2 = 24 + 6{,}6 = 30{,}6$ машинных такта.

Таблица 2

Соответствие между некоторыми двоичными A_j и Z_j при n=24

Двоичная последовательность A_j	Двоичная комбинация Z_j		Вид f_{bg}
	Bin k	X_j или пустая строка	
000000000000000000000000	00000	–	$f_k \circ f_w$
000000000000000000001001	00010	0000000000000000000100	$f_k \circ f_b \circ f_w$
000000001000011000000000	00011	000000001000011	$f_k \circ f_b \circ f_w$
001100001110011100100000	01001	0011000011100111001	$f_k \circ f_b \circ f_w$
011100010000000000000000	00100	01110001	$f_k \circ f_b \circ f_w$
011111011001010000101100	01100	0111110110010100001011	$f_k \circ f_b \circ f_w$
100000000000000000000000	00001	1	$f_k \circ f_b \circ f_w$
111000010001110111111111	10000	111000010001110	$f_k \circ f_b \circ f_w$
111111111111111111111110	10111	1111111111111111111111	$f_k \circ f_b \circ f_w$
111111111111111111111111	11000	–	$f_k \circ f_w$

Анализ результатов биномиального сжатия равновесных комбинаций на основе двоичных биномиальных чисел демонстрирует высокую скорость кодирования и декодирования при достаточно хорошем коэффициенте сжатия. Время биномиального сжатия или восстановления будет ограничиваться величиной $t \le n - k$, если $k \le \lceil n/2 \rceil$, или $t \le k$, если $k > \lceil n/2 \rceil$.

Анализ результатов обобщенного биномиального сжатия двоичных последовательностей, которые могут иметь переменное значение единиц, также показывает относительно высокую скорость обработки информации при достаточно хорошем коэффициенте сжатия. Но ко времени t_2=t собственно сжатия здесь добавляется время t_1 вычисления числа k единиц, которое можно сделать существенно меньшим чем значение n [15].

Следует отметить, что при $n \to \infty$ коэффициент сжатия для методов, которые используют двоичные биномиальные числа, будет асимптотически стремиться к значению обратному величине энтропии исходного источника информации [9].

Для обоих методов сжатия требуемый объем аппаратно-программных затрат при практической реализации является низким, поскольку при сжатии используются простые операции просмотра разрядов сжимаемых комбинаций, проверки их значений и подсчета числа двоичных единиц.

7. SWOT-анализ результатов исследований

Strengths. Достоинствами методов сжатия на основе двоичных биномиальных чисел являются следующие.

1. Использование двоичных биномиальных чисел для сжатия позволяет увеличить производительность информационных систем. Это происходит за счет уменьшения времени передачи сжатой информации и уменьшения требуемого объема памяти для ее хранения. При этом коэффициенты сжатия данных в информационной системе являются достаточно высокими.

2. Методы сжатия на основе двоичных биномиальных чисел обладают высоким быстродействием при низком уровне аппаратно-программных затрат. Это позволяет работать рассматриваемым методам в реальном масштабе времени, а объем стоимостных затрат на их внедрение обеспечивается минимальным.

3. Исследованные методы сжатия имеют универсальный характер применения и направлены на обработку распространенных двоичных последовательностей, для которых необходимо вычислить только количество содержащих в них единиц.

4. Сжатые образы, получаемые при сжимающем кодировании, обладают свойствами чисел, что дает дополнительный положительный эффект при применении рассмотренных методов сжатия. Над ними можно проводить различные арифметические операции, операции сравнения и упорядочивания без восстановления исходных данных, что является очень полезным качеством, например, для распределенных баз данных.

Weaknesses. Недостатками методов сжатия на основе двоичных биномиальных чисел являются следующие.

1. В случае переменного значения количества единиц в двоичных последовательностях необходимо использование служебного слова для каждой сжатой комбинации с целью ее дальнейшего однозначного восстановления, что несколько снижает степень биномиального сжатия.

2. Степень сжатия также будет снижаться и может быть меньше единицы при приблизительно равном количестве двоичных нулей и единиц, а также при их равномерном расположении в разрядной сетке сжимаемых последовательностей.

Opportunities. Возможности и перспективы дальнейших исследований сжатия на основе двоичных биномиальных чисел являются следующие.

1. Обнаружение и использование граничных значений числа единиц, при которых сжатие на основе двоичных биномиальных чисел является целесообразным, позволит существенно уменьшить время кодирования и улучшить степень биномиального сжатия.

2. Системы кодообразующих ограничений для двоичных биномиальных чисел предоставляют возможность контролировать появление ошибок при сжатии двоичных последовательностей, а также при их передаче по каналу связи и хранении в памяти. Это позволяет повысить помехоустойчивость информационных систем.

Хорошую перспективу использования сжатия на основе двоичных биномиальных чисел в различных информационных системах с целью повышения их производительности, улучшения их технико-экономических характеристик и снижения стоимости обработки данных определяют такие характеристики разработанных моделей:

– быстродействие;

– универсальность по отношению к типам данных;

– низкий объем затрат при практической реализации;

– числовые характеристики сжатых образов.

Threats. Отрицательным фактором на внедрение методов сжатия на основе двоичных биномиальных чисел является преимущественное использование в системе двоичных данных, в которых вероятности появления нулей и единиц приблизительно равны. Это в существенной степени снижает эффективность применения рассматриваемых методов и может ограничить повышение производительности информационной системы.

8. Выводы

1. Построена математическая модель метода сжатия равновесных комбинаций на основе двоичных биномиальных чисел. Основу полученной математической модели составляют теоремы 1, 2 и разработанные модели процессов биномиального сжатия и восстановления для последовательностей с постоянным числом единиц.

2. Построена математическая модель обобщенного биномиального метода сжатия двоичных последовательностей с переменным числом единиц. Основу полученной математической модели составляют теорема 3 и разработанная модель процесса обобщенного биномиального сжатия (модель процесса обобщенного биномиального восстановления основывается на модели восстановления для равновесных комбинаций).

Разработанные модели процессов сжатия и восстановления на основе двоичных биномиальных чисел характеризуются небольшим числом простых операций, что обеспечивает высокое быстродействие кодирования и декодирования, а также низкий объем аппаратно-программных затрат.

Литература

1. Sayood Kh. Introduction to Data Compression. Morgan Kaufmann, 2017. 790 p.

2. Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео / Ватолин Д. и др. М.: ДИАЛОГ-МИФИ, 2003. 384 с.

3. Смирнов М. А. Обзор применения методов безущербного сжатия данных в СУБД. URL: http://compression.ru/download/articles/db/smirnov_2003_database_compression_review.pdf

4. Sayood K., Memon N. Lossless Compression Handbook. Academic Press, 2012. 488 p. doi: http://doi.org/10.1201/9781420041163-101

5. Борисенко О. А. Число і системи числення в електронних цифрових системах // Вісник СумДУ. 2007. № 4. С. 71–76.

6. Борисенко А. А., Кулик И. А. Биномиальное кодирование: монография / Сумы: СумГУ, 2010. 206 с.

7. Борисенко А. А. Биномиальный счет. Теория и практика: монография / Сумы: ИТД «Университетская книга», 2004. 170 с.

8. Кулик И. А., Чередниченко В. Б., Костель С. В. Алгоритм генерирования двоичных биномиальных чисел на основе минимальных систем кодообразующих ограничений // Вісник СумДУ. Серія «Технічні науки». 2008. № 2. С. 45–52.

9. Schalkwijk J. An algorithm for source coding // IEEE Transactions on Information Theory. 1972. Vol. 18, Issue 3. P. 395–399. doi: http://doi.org/10.1109/tit.1972.1054832

10. Cover T. Enumerative source encoding // IEEE Transactions on Information Theory. 1973. Vol. 19, Issue 1. P. 73–77. doi: http://doi.org/10.1109/tit.1973.1054929

11. Амелькин В. А. Методы нумерационного кодирования. Новосибирск: Наука, 1986. 155 с.

12. Амелькин В. А. Перечислительные задачи серийных последовательностей. Новосибирск: ИВМиМГ СО РАН, 2008. 317 с.

13. Кулик И. А., Борисенко А. А., Онориукпе Аджири. Модели сжатия и восстановления данных на основе двоичных биномиальных чисел // V Міжнародна науково-практична конференція «Методи та засоби кодування, захисту й ущільнення інформації», 19–21 квітня 2016 р.: тез. доп. Вінниця: Вінницький національний технічний університет, 2016. С. 101–105.

14. Кулик И. А., Скордина Е. М., Костель С. В. Генерирование кодов-сочетаний для решения информационных задач ИУС // АСУ и приборы автоматики. Всеукраинский межведомственный сборник. 2011. № 155. С. 15–23.

15. Cover T. Enumerative source encoding // IEEE Transactions on Information Theory. 1973. Vol. 19, Issue 1. P. 73–77. doi: http://doi.org/10.1109/tit.1973.1054929