

УДК 621.327

В.В. Баранник, А.В. Коломийцев, О.Н. Карпенко

Харьковский университет Воздушных Сил им. И. Кожедуба

## ОБОСНОВАНИЕ НЕОБХОДИМОСТИ СОВЕРШЕНСТВОВАНИЯ ТЕХНОЛОГИИ СЖАТИЯ ДАННЫХ

*Рассматриваются основные особенности функционирования существующих технологий обработки данных от различных источников информации. Обосновывается направление дополнительного повышения степени сжатия без внесения погрешности.*

*архивация данных, графические форматы, структурные закономерности*

### Введение

В настоящее время разработано несколько десятков методов компактного представления, часть из которых составляет основу для построения стандартизированных технологий сжатия аудиоданных, видеоданных, текстовых документов и двоичных последовательностей [1]. Это позволяет снизить объемы обрабатываемых и передаваемых цифровых

данных в информационных системах. В тоже время даже на компактное представление данных затрачивается большое количество двоичных разрядов. Усредненные значения объемов цифровых данных, представленных различными форматами, приведены в табл. 1. Поэтому задача организации дополнительного снижения объемов данных без внесения погрешности является актуальной.

Таблица 1

Усредненные значения объемов цифровых данных, представленных различными форматами, в Мбит

Тип данных	Объем данных
Книги в формате djvu (400 – 900стр.)	56 – 400
Книга в формате pdf, 500 стр.	384 – 560
Архив (rar) книги в формате doc с графическими документами (150 – 300 стр.)	16 – 56
Реалистическое изображение с борта КА в формате jpeg 1048 × 915 × 24 бита	0,3
Реалистическое изображение с борта КА в формате bmp 2295 × 1536 × 24 бита	84
Реалистическое изображение в формате tiff 1118 × 1105 × 16 бит	3,5

### 1. Анализ известных технологий компактного представления данных

Существующие технологии компактного представления данных можно классифицировать в зависимости от видов обрабатываемой информации [2 – 4]:

1) *технология сжатия статических изображений*; основными представителями данной технологии являются методы формата JPEG, алгоритм RLE, алгоритм LZW;

2) *технология сжатия динамических изображений*, базируется на методах форматов MPEG 2 и MPEG 4;

3) *технология сжатия аудиоданных и речи*, включает в себя методы формата MPEG 3 вокодеры;

4) *технология архивации данных*, включает такие архиваторы, как rar, zip и arj, данные архиваторы строятся с использованием алгоритмов арифметического кодирования, кодирования Хаффмена, методов LZW и RLE.

Значения коэффициента сжатия для различных технологий компактного представления приведены в табл. 2, табл. 3. Анализ данных, представленных в табл. 2 и 3, показывает следующее.

1. Файлы, содержащие документы, представленные в графическом виде имеют, большие объемы, достигающие десятков Мбит. Это обусловлено недостаточными значениями степени сжатия, которые обеспечиваются технологиями компактного представления изображений (коэффициент сжатия реалистических изображений изменяется в среднем от 2 до 5 раз).

2. Существующие технологии архивирования данных обеспечивают незначительное сжатие до 2 раз для файлов формата bmp. Для файлов форматов jpeg в результате архивации происходит увеличение первоначальных объемов в среднем на 2%.

3. Компактное представление архивов на основе существующих архиваторов не дает увеличения степени сжатия.

4. Существующие технологии компактного представления данных для существующих скоростей выполнения машинных операций и скоростей передачи данных по каналу связи [5] **не обеспечивают** своевременную передачу реалистических изображений и файлов, содержащих документы представленные в графических форматах, исходный объем которых превышает 10 Мбит. Суммарное

время на обработку и передачу данных изменяется от нескольких десятков секунд до нескольких часов.

Таким образом, существующие технологии ком-

пактного представления не обеспечивают требуемых значений степени сжатия данных, полученных от различных источников информации.

Таблица 2

Значение  $k_{сж}$  для разных технологий сжатия при обработке реалистических изображений

Метод	RLE	LZW	Формат PCX	JPEG, q=0	JPEG, h ≥ 30дБ	Архиватор rar	Архиватор zip
$k_{сж}$	0,7	1,1	0,9	2	5,3	1,5	1,35

Таблица 3

Значение  $k_{сж}$  для разных технологий сжатия и типов данных

Метод сжатия	Виды исходных данных						
	Файл rar	Файл bmp	Файл jpeg	Файл djvu	Файл видео	Файл pdf	Файл mp3
Архиватор rar	1	1,9	0,98	1	1,02	1	1,047
Архиватор zip	1	1,22	0,99	0,99	1,01	1,0001	1,045

## 2. Обоснование выбора направления разработки новых технологий компактного представления данных

Для того, чтобы сформулировать направление создание новых технологий компактного представления необходимо выявить недостатки существующих подходов к сжатию данных.

Технологии сжатия статических и динамических изображений, обеспечивают наибольшие степени сжатия за счет сокращения психовизуальной избыточности и последующего статистического кодирования компонент трансформант ортогональных преобразований. Психовизуальная избыточность сокращается в результате обнуления высокочастотных составляющих компонент трансформант. Основными недостатками данных технологий являются:

- возможные потери информации, которые возникают на этапе самого преобразования и на этапе квантизации их компонент;

- зависимость эффективности сжатия от характеристик источника информации.

По этим причинам методы данной технологии нельзя использовать для сжатия данных, полученных от различных источников информации и требующих различной степени достоверности (для некоторых приложений требуется проводить обработку без внесения погрешностей).

Технологии, обеспечивающие обработку различных типов данных без потери информации, реализуются на основе различных типов архиваторов. В этом случае обработка данных состоит из двух основных этапов:

- на первом этапе с использованием алгоритмов RLE или LZW осуществляется выявление цепочек одинаковых элементов (сокращается простейшая структурная избыточность);

- на втором этапе проводится арифметическое кодирование или кодирование Хаффмана длин цепочек одинаковых элементов. На данном этапе сжатие достигается за счет исключения статистической избыточности.

Основными недостатками существующих архиваторов являются:

- сильная зависимость степени сжатия данных от количества и размеров областей, содержащих одинаковые элементы, от степени неравномерности распределения вероятности появления символов алфавита. В тоже время большая часть данных от различных источников информации характеризуются нестационарностью свойств. Это приводит к низким степеням сжатия на основе существующих архиваторов;

- то, что отдельные этапы обработки данных в архиваторах совпадают с этапами обработки данных на основе технологий сжатия изображений. В этом случае эффективность архиваторов зависит от тех же факторов, что и эффективность технологий сжатия видеоданных. Следовательно, архивация файлов, полученных с выхода кодеров статических, динамических изображений, будет сопровождаться увеличением первоначального объема;

- то, что выходные кодовые комбинации архиваторов характеризуются отсутствием областей, содержащих одинаковые элементы, а законы распределения вероятности появления символов стремятся к равномерному закону. Это делает невозможность дополнительного повышения степени сжатия архивных файлов;

- низкая помехоустойчивость кодовых комбинаций с ошибкам в канале связи. Это обусловлено сильной степенью размножения ошибок в канале связи на значительную часть восстанавливаемых данных. Данная особенность объясняется тем, что:

- ошибка в разделителе кодовых слов приводит к восприятию его как кодового слова. Отсюда все последующие кодовые слова будут декодированы неправильно;

- ошибка в кодовой комбинации арифметического кода приведет к неправильному восстановлению всех последующих элементов;

- арифметические коды несут информацию о длине цепочек одинаковых элементов. Поэтому ошибка в одном элементе приводит к сдвигу всех последующих восстанавливаемых элементов исходных данных.

Отсюда следует, что для повышения степени сжатия и сохранения заданной степени достоверно-

сти создаваемая технология компактного представления данных от различных источников информации, должна удовлетворять следующим требованиям:

1) организовывать процесс сжатия данных на основе устранения структурной избыточности;

2) учитывать в процессе сжатия структурные закономерности одновременно по двум признакам. При этом для обеспечения дополнительного сжатия признаки должны быть взаимозависимыми;

3) обрабатывать двоичные данные. Двоичное представление является универсальным для представления информации от различных источников.

Выбор признаков структурной природы обусловлен тем, что [6, 7] они имеют количественную меру; учитывают закономерности, позволяющие исключить избыточность без потери качества; используются только целочисленные операции, что сокращает время обработки; не требуется знание априорной информации для кодирования и является более устойчивым к нестационарности изображений.

Отсюда следует наличие противоречия на уровне разработки технологии компактного представления, которое состоит в том, что:

– с одной стороны устранение избыточности одновременно по нескольким признакам позволяет достичь больших степеней сжатия;

– с другой стороны отсутствуют теоретические основы и методы сокращения структурной двухпризнаковой избыточности.

Значит, требуется выбрать конкретные взаимозависимые информативные признаки структурной природы, удовлетворяющие требованиям, предъявляемые процессом компактного представления данных от различных источников информации.

### 3. Структурное представление данных

Структурная избыточность исключается в результате выявления структурных закономерностей в двоичных последовательностях (одномерная последовательность, двумерный массив или трехмерная структура) по некоторому признаку [6, 7]. Суть такого кодирования заключается в формировании кода-номера всей двоичной последовательности с заданным значением структурного признака. Наиболее информативными структурными признаками являются вектор  $S$  запретов появления на определенной позиции единичного элемента ( $S = \{s_i\}$ ,  $i = \overline{1, m}$ ;  $s_i$  – признак запрета появления на  $i$ -й позиции единичного элемента; если  $s_i = 0$ , то на  $i$ -й позиции запрещено появление единицы и наоборот) и количество серий единиц  $\theta$  в двоичной последовательности [7].

Кодирование с учетом запретов на позиции единиц соответствует полиадическому представлению двоичных данных. В этом случае двоичное число рассматривается как полиадическое число, элементы которого принимают значения  $\{0; 1\}$ . Тогда основания двоичного полиадического числа на-

ходятся по формуле  $\lambda_i = s_i + 1$ , где  $i = \overline{1, m}$ . В этом случае позиции  $i$ , для которой запрещено появление 1, соответствует  $\lambda_i = 1$  и  $s_i = 0$ . Достоинствами методов кодирования с учетом вектора ограничений  $S$  являются: относительно небольшое количество операций на обработку; возможность распараллеливания обработки; повышение степени сжатия  $k_{сж}$  с ростом количества позиций с запретом появления единичных элементов. Основным недостатком такого кодирования является резкое снижение степени сжатия при уменьшении количества запретов на появление единичных элементов и при увеличении длины двоичной последовательности. Степень сжатия в зависимости от степени насыщенности единичными элементами изменяется от 1,2 до 2,5 раз.

Достоинствами кодирования данных с учетом ограниченного количества серий единиц  $\theta$  являются: уменьшение зависимости степени сжатия от закона распределения и от длины серии одинаковых двоичных элементов; возможность повышения степени сжатия с увеличением единичных элементов и с ростом длины двоичной последовательности; относительно небольшим количеством операций на нахождение количества серий единиц. В тоже время основными недостатками являются: отсутствие параллельной схемы кодирования; возможны варианты уменьшения степени сжатия даже при уменьшении единичных элементов. Степень сжатия в зависимости от количества серий единиц изменяется от 1,3 до 3 раз.

В тоже время существуют двоичные последовательности, для которых выполняется неравенство  $N(m, \Lambda) \leq N(m, \theta)$  и наоборот  $N(m, \Lambda) \geq N(m, \theta)$ . Отсюда следует, что для повышения степени сжатия двоичных данных без внесения погрешности и снижения времени обработки требуется сформировать двоичные массивы и разработать кодирование, учитывающее одновременно закономерности в направлении столбцов по количеству серий единиц, а в направлении строк – на позиции с возможным появлением единичного элемента

$$0 \leq a_{ij} \leq s_i; \quad s_i = \max_{1 \leq j \leq n} \{a_{ij}\}; \quad a_{ij}; \quad s_i \in \{0; 1\}; \quad (1)$$

$$v(a_{ij})_{i=1, m} = \theta_j, \quad (2)$$

где  $a_{ij}$  –  $ij$ -й элемент двоичного массива;  $\max_{1 \leq j \leq n} \{a_{ij}\}$  –

оператор вычисления ограничения на позиции с возможным появлением единиц;  $\theta_j$  – количество серий единиц для  $j$ -го столбца;  $v(a_{ij})_{i=1, m}$  – оператор вычисления количества серий единиц для столбцов двоичного массива.

Таким образом, построение технологий компактного представления двоичных данных, на которые одновременно наложены ограничений на количество единичных серий и на позиции с запретом появления единичных элементов является перспективным направлением.

## Выводы

1. Технологии компактного представления данных показали, что для существующих скоростей выполнения машинных операций и скоростей передачи данных по каналу связи **не обеспечивают** своевременную передачу реалистических изображений и файлов, содержащих документы представленные в графических форматах, исходный объем которых превышает 10 Мбит. Суммарное время на обработку и передачу данных изменяется от нескольких десятков секунд до нескольких часов, что не удовлетворяет требованиям относительно своевременности доведения информации.

2. Обосновано, что для повышения степени сжатия и сохранения заданной степени достоверности новая технология компактного представления данных должна удовлетворять следующим требованиям: организовывать процесс сжатия данных на основе устранения структурной избыточности; учитывать в процессе сжатия структурные закономерности одновременно по двум признакам. При этом для обеспечения дополнительного сжатия признаки должны быть взаимозависимыми; обрабатывать двоичные данные. Двоичное представление является универсальным для представления информации от различных источников.

3. Обосновано, что для повышения степени компактного представления двоичных данных с заданной степенью достоверности необходимо орга-

низовывать построение технологии сжатия данных на основе двухпризнакового кодирования в двоичном полиадическом пространстве.

## Список литературы

1. Ватолин В.И., Ратушняк А., Смирнов М., Юкин В. *Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео.* – М.: ДИАЛОГ – МИФИ, 2002. – 384 с.
2. Bell T.C. *Text compression.* Englewood Cliffs, N. J.: Prentice – Hall, 1990. – 210 p.
3. Rissanen J.J. *Universal coding, information, prediction, and estimation* // *IEEE Transactions on Information Theory.* – 1984. – Vol. 30, N 4. – P. 629-636.
4. Ziv J., Lempel A. *Compression of individual sequences via variable-rate coding* // *IEEE Transactions on Information Theory.* – 1978. – Vol. 24, N 5. – P. 530-536.
5. Стеклов В.К., Бирюков Н.Л. *Транспортные сети и системы электросвязи.* – К.: Зв'язок, 2003. – 352 с.
6. Баранник В.В., Юдин А.К. *Двоичное полиадическое кодирование по количеству серий единиц* // *Радиоэлектроника и информатика.* – 2005. – № 2. – С. 56-63.
7. Баранник В.В. *Метод сжатия изображений комбинированным полиадическим кодированием* // *Інформаційно-керуючі системи на залізничному транспорті.* – 2000. – № 2. – С. 66-69.

Поступила в редколлегию 06.04.2007

**Рецензент:** д-р техн. наук, проф. П.Ф. Поляков, Украинская государственная академия железнодорожного транспорта, Харьков.