УДК 004.627

© 2021 г. **Н.В. Безденежных,**
В.Э. Иванов, канд. техн. наук,
А.В. Левенец, д-р техн. наук,
Чье Ен Ун, д-р техн. наук

(Тихоокеанский государственный университет, Хабаровск)

ОЦЕНКА ВЛИЯНИЯ ПРОЦЕДУРЫ ИЗМЕНЕНИЯ РАЗРЯДНОСТИ ИЗМЕРИТЕЛЬНЫХ ДАННЫХ НА ЭФФЕКТИВНОСТЬ ИХ СЖАТИЯ

В статье рассматривается процедура предварительной обработки данных перед проведением сжатия, состоящая в принудительном изменении исходной разрядности данных. Оценивается влияние такой процедуры на эффективность сжатия, приведены результаты исследования, показывающие перспективность предложенного подхода.

Ключевые слова: измерительные данные, предварительная обработка, сжатие данных, разрядность, информационно-измерительная система.

DOI: 10.22250/isu.2021.68.109-115

Введение

Применение алгоритмов сжатия данных в настоящее время является традиционной практикой в случае необходимости экономии пропускной способности канала связи, при ограниченных объемах запоминающих устройств и т.д. Необходимо отметить, что современные методы сжатия цифровых данных практически вплотную приблизились к предельной границе теоретически возможного сжатия в условиях существующего общего подхода к задаче сжатия [1].

Эффективность сжатия обеспечивается, кроме особенностей собственно алгоритмов сжатия, процедурами предварительной обработки, приводящими свойства исходных данных к наиболее оптимальному варианту. Следует отметить, что процедуры предварительной обработки часто ориентированы на конкретный тип данных, обладающих некоторыми специфическими особенностями, в силу чего не применяются к данным других типов, так как

в лучшем случае не повышают эффективность процедур сжатия, либо их применение имеет ограниченный характер в силу недостаточно высокой эффективности.

Наиболее распространенной предварительной обработкой можно считать преобразование Барроуза – Уиллера [2], которое применяется в различных архиваторах и предназначено в основном для обработки текстовых данных и в меньшей степени – данных мультимедиа. При этом преобразование неэффективно с практической точки зрения при работе с измерительными данными, для которых повторение одного и того же значения возможно только в том случае, когда динамический диапазон данных существенно меньше объема анализируемой выборки.

Фактически в настоящее время в качестве относительно универсальной предварительной обработки измерительных данных можно предложить только методы выделения постоянной составляющей и линейного или периодического тренда. Наиболее простой обработкой такого вида является построение разностного ряда, что позволяет избавиться от постоянной составляющей и, как правило, существенно снизить динамический диапазон, обеспечивая тем самым более высокую эффективность сжатия. К сожалению, такая обработка практически неприменима к нестационарным измерительным данным, обладающим наиболее высокой информационной ценностью – по крайней мере, для задач исследовательского типа.

Основные положения

В основу широко распространенных современных алгоритмов предварительной обработки и сжатия данных положен принцип нерушимости границ слов исходных (сжимаемых) данных. Другими словами, если исходные данные представляют собой набор n -разрядных слов, то дальнейшая работа алгоритмов предварительной обработки и сжатия будет связана с обработкой именно таких слов, однако при этом выходные данные могут быть и другой разрядности. С одной стороны, подобный подход вполне оправдан, поскольку предполагается, что источник информации выдает последовательность отсчетов, коррелирующих друг с другом, и задачей соответствующих алгоритмов является определение степени взаимосвязи отсчетов, а также удаление избыточной информации в каждом отсчете.

Однако можно предположить, что корреляционные связи могут возникать не только между самими отсчетами, но и между частями отсчетов. Наиболее наглядным примером этого может служить тот факт, что если поток данных от источника будет представлен, например, восьмиразрядными

числами, причем изменение значений соседних отсчетов незначительно (например, в пределах ± 2), то старшие биты отсчетов в значительном числе случаев не будут изменяться, т.е. будут сильно коррелированы. Можно предположить, что на такую коррелированность битовых ансамблей будут влиять как параметры измерительного сигнала, так и способ кодирования отсчетов этого сигнала, который может обеспечить требуемое в рамках предлагаемого подхода поведение соответствующих битов каждого отсчета.

Таким образом, в качестве некоторой обобщенной предварительной обработки данных перед сжатием можно предложить разбиение исходного их набора, подвергающегося сжатию, на слова, разрядность которых позволит алгоритму сжатия работать более эффективно. Соответствующая процедура состоит из следующих основных шагов:

- исходный набор данных представляются в виде битового массива, без явного выделения границ слов;

- оценивается корреляционная взаимосвязь внутри полученного битового массива, например, с помощью битовой корреляционной функции [3];

- по полученной оценке определяется оптимальная (субоптимальная) разрядность слов;

- битовый массив, полученный на первом шаге, разбивается на слова соответствующей разрядности;

- производится операция упорядочивания, – например, с помощью преобразования Барроуза – Уиллера;

- полученный массив передается соответствующему алгоритму сжатия.

Очевидно, что на первом этапе следует оценить принципиальную возможность применения предложенной процедуры к реальным измерительным данным и оценить ее эффективность. Следует учесть, что любое преобразование обеспечивает максимальную эффективность сжатия при использовании специализированных алгоритмов сжатия, наиболее полно реализующих предоставляемые таким преобразования возможности. В данном случае подобная оценка может быть произведена только относительно некоторой идеальной процедуры сжатия, в качестве которой далее используется алгоритм RLE и предполагается, что данные на его входе упорядочены оптимальным способом.

В качестве оценки эффективности применялся коэффициент сжатия $k_{сж}$, равный отношению числа битов на выходе процедуры сжатия $n_{сж}$ к числу битов исходных данных $n_{исх}$. Алгоритм RLE предполагает, что сжатая последовательность, состоящая из m одинаковых символов (чисел), представляется двумя битовыми полями, первое из которых описывает длину последова-

тельности, а второе – сам символ. Таким образом, если сжимаемая последовательность состоит из N слов, а словарь слов – из K символов, причем некоторый k -й символ встречается в сжимаемом наборе m_k раз, то общую длину закодированной последовательности можно рассчитать следующим образом:

$$n_{\text{сж}} = N(\log_2 K + \log_2 [\max_k m_k]).$$

Число битов исходных данных определяется как произведение объема сжимаемого набора данных N и исходной разрядности слов n_N . Здесь не учитываются накладные расходы на упорядочивание данных перед сжатием и передачу необходимой информации приемной стороне, так как производится оценка общего подхода, а не конкретного алгоритма.

Исходные данные

Исследование было проведено на телеметрических данных, полученных от пяти различных энергетических объектов Дальневосточного региона (электростанции, подстанции) и представляющих собой наборы показаний датчиков (кадры), объединенные единым временем получения. Такие наборы данных далее будут обозначаться символом S с соответствующим номером. Объемы наборов данных различны и находятся в диапазоне от 550 килобайт до 1,1 мегабайта, что связано с различающейся структурой кадров, которые для разных объектов состоят из разного числа датчиков. Разрядность слов для всех наборов равна восьми.

Еще одной группой наборов, по которым проводилось исследование, были последовательности отсчетов, получаемых от одного датчика. В этом случае тестовые последовательности обозначались символом D . Число таких наборов данных было равно семи. Датчики отличались размером словаря символов (динамическим диапазоном). Так, для набора $D1$ размер словаря был равен шести, а для остальных – 10, 16, 21, 23, 48 и 57 соответственно.

Таким образом, полученные наборы данных позволяют оценить эффективность предложенной процедуры как при работе с кадрами данных, когда в битовую последовательность с последующей разбивкой на выбранную разрядность преобразуются наборы отсчетов от разных источников (датчиков), так и при работе с последовательностью отсчетов от одного датчика.

Важно подчеркнуть, что в первом случае фактически оценивалась возможность выявления битовых корреляционных связей между различными источниками информации.

Результаты исследования

В процессе исследования оценивался коэффициент сжатия при разбиении исходных данных на слова следующей разрядности: 2, 4, 8, 16, 24, 32. Выбор набора разрядностей был произвольным. Полученные в ходе исследования результаты усреднялись по 20 реализациям. Следует подчеркнуть, что здесь меньшие значения означали более высокую эффективность. Объем анализируемых данных составил 768 отсчетов соответствующей разрядности.

Результаты проведенного исследования, полученные для наборов данных типа S, представляющих собой наборы кадров данных, представлены в табл. 1.

Таблица 1

Набор данных	Разрядность					
	2	4	8	16	24	32
S1	0,0274 (2,44)	0,06 (1,11)	0,0663 (1)	0,0692 (0,96)	0,1392 (0,48)	0,0754 (0,88)
S2	0,0266 (12,5)	0,0593 (5,26)	0,3143 (1)	0,4016 (0,79)	0,2946 (1,07)	0,2448 (1,28)
S3	0,028 (3,7)	0,0617 (1,63)	0,1016 (1)	0,0929 (1,09)	0,1822 (0,56)	0,0908 (11,01)
S4	0,0257 (5,26)	0,0584 (2,32)	0,135 (1)	0,1372 (0,99)	0,2168 (0,62)	0,1374 (0,98)
S5	0,0267 (4,54)	0,0594 (2,08)	0,1223 (1)	0,103 (1,19)	0,0964 (1,26)	0,0819 (1,49)

Помимо абсолютных значений, в табл. 1 в скобках приведены относительные величины, причем в качестве базовых использованы значения для исходной разрядности восемь. Анализ полученных результатов показывает, что исходная разрядность данных (восемь) никогда не становится оптимальной с точки зрения сжатия. При этом уменьшение разрядности всегда приводит к уменьшению коэффициента сжатия и, следовательно, к увеличению эффективности процедуры сжатия. Увеличение разрядности до 16 и 24, т.е. объединение стоящих рядом двух и трех байтов соответственно, не приводит к снижению коэффициента сжатия, а разрядность 32 в ряде случаев дает хорошие результаты. Эти факты можно объяснить тем, что объединение байтов, принадлежащих разным источникам информации, в общем случае не приводит к сокращению размера словаря символов, а аномальные значения для разрядности 32 объясняются особенностями группирования датчиков в кадрах рассматриваемых наборов.

Результаты исследования предложенной процедуры на наборах дан-

ных, представляющих собой временную последовательность показаний одного датчика, приведены в табл. 2. Здесь, как и в табл. 1, приведены не только значения коэффициента сжатия, но и относительные величины, позволяющие получить представление об эффективности проведенной предварительной обработки.

Следует отметить, что для такого типа наборов данных выбор разрядности большей, чем восемь не дает положительного эффекта, т.е. не уменьшает коэффициент сжатия. При этом переход на меньшую разрядность оказывается вполне оправданным, только в случае, когда размер словаря (динамический диапазон) исходных данных достаточно велик. Так, для разрядности два обработка становится эффективной при динамическом диапазоне исходных данных, равном 16, а для разрядности четыре эта величина должна быть не меньше 42.

Таблица 2

Набор данных	Разрядность					
	2	4	8	16	24	32
D1	0,0285 (0,27)	0,0198 (0,39)	0,0078 (1)	0,0149 (0,52)	0,0242 (0,32)	0,0344 (0,22)
D2	0,0276 (0,54)	0,0283 (0,53)	0,0149 (1)	0,0327 (0,45)	0,0426 (0,35)	0,0559 (0,27)
D3	0,0267 (0,98)	0,0508 (0,51)	0,0262 (1)	0,0793 (0,33)	0,1536 (0,17)	0,2195 (0,12)
D4	0,0267 (1,12)	0,0534 (0,56)	0,03 (1)	0,1015 (0,29)	0,1855 (0,16)	0,2161 (0,14)
D5	0,0279 (1,17)	0,0523 (0,63)	0,0329 (1)	0,1238 (0,26)	0,2076 (0,16)	0,2373 (0,14)
D6	0,0261 (3,88)	0,0621 (1,63)	0,1013 (1)	0,2801 (0,36)	0,3926 (0,26)	0,3309 (0,31)
D7	0,0262 (4,03)	0,0624 (1,69)	0,1058 (1)	0,3242 (0,32)	0,4004 (0,26)	0,3341 (0,32)

Такие результаты можно объяснить тем, что эффективность перехода на малые разрядности обусловлена уменьшением динамического диапазона, что в общем случае приводит к уменьшению величины $n_{сж}$ и, как следствие, к уменьшению коэффициента сжатия. Очевидно, что когда динамический диапазон при восьмиразрядном представлении несущественно превышает максимально возможный динамический диапазон слов меньшей разрядности, эффективность процедуры изменения разрядности ожидаемо становится низкой.

Приведенные здесь данные характерны для всех наборов аналогичного типа, проанализированных в рамках исследования, и подтверждают выдви-

нутое предположение о достаточно сильном влиянии процедуры изменения разрядности на эффективность сжатия измерительных данных.

Следует, однако, подчеркнуть, что полученный эффект повышения эффективности сжатия может быть заметно снижен при практической реализации предложенного подхода.

Заключение

Таким образом, результаты проведенного исследования позволяют утверждать, что предложенная процедура изменения разрядности измерительных данных может использоваться как предварительная обработка перед процедурой сжатия и обеспечивает снижение коэффициента сжатия, особенно в случае уменьшения исходной разрядности. Важно подчеркнуть, что можно добиться улучшения работы алгоритмов сжатия при работе не только с последовательностью данных от одного источника, но и с совокупностью отсчетов от разных источников информации, которые фактически рассматриваются как одна информационная структура, без фиксации границ отдельных отсчетов. Дальнейшее развитие предложенного подхода можно связать с проведением более глубоких исследований, в том числе на данных от промышленных и технических объектов разного типа. Кроме того, нужно разработать специализированный алгоритм сжатия, позволяющий снизить накладные расходы, связанные с необходимостью приведения обработанных данных к некоторому оптимальному виду, с точки зрения существующих алгоритмов сжатия.

ЛИТЕРАТУРА

1. *Сергеенко В.С., Баринов В.В.* Сжатие данных, речи, звука и изображений в телекоммуникационных системах. – М.: РадиоСофт, 2009.
2. *Ватолин Д., Ратушняк А., Смирнов М., Юкин В.* Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео. – М.: ДИАЛОГ-МИФИ, 2003.
3. *Чье Ен Ун, Левенец А.В., Нильга В.В.* Представление телемеханических данных однородными n-мерными структурами как предварительная обработка в задачах сжатия // Информационно-управляющие системы. – 2011. – №6. – С. 7-10.

E-mail:

Безденежных Никита Владимирович – 2016103161@pnu.edu.ru;

Иванов Виктор Эдуардович – ivanov72@list.ru;

Левенец Алексей Викторович. – 000621@pnu.edu.ru;

Чье Ен Ун – chye@ais.khstu.ru.