УДК 004:681.3

Новые методы сжатия временных рядов экологических показателей

Чуприн В.И., Родригес Залепинос Р.А.
Донецкий национальный технический университет
chuprin.vladislav@gmail.com, rodriges@csm.donntu.edu.ua

Abstract

Chuprin V.I., Rodriges Zalipynis R.A. "New time series compression methods of ecologic parameters". *The paper analyzes storage peculiarities of satellite Earth remote sensing data time series. We propose methods for their compression based on the discovered peculiarities exploiting different schemes of Huffman coding. One of the proposed methods reaches 6% increase in the compression ratio (93%) in contrast to the deflate method used in Java SE6 (87%), for a time series of aerosol optical thickness derived from MODIS radiometer of TERRA satellite. Further improvement can be achieved by using the entropy coding of floating point numbers.*

Keywords: time series, lossless compression, Huffman coding, arithmetic compression, floating-point data, Aura, TERRA, MODIS, OMI.

Введение

Анализ временных рядов является одним из важнейших инструментов оценки и прогнозирования состояния окружающей среды [1].

Новые средства мониторинга, генерируют большие объемы данных [2]. Например, космический исследовательский аппарат Aura предоставляет 26 гигабайт информации ежедневно [3].

Большие объемы данных являются причиной увеличения времени считывания и передачи информации по сети. Возрастают требования к размеру дисковых накопителей. Появляется необходимость использования полос высоких пропускных способностей, что часто приводит к повышению финансовых расходов. Одним из решений приведенных выше проблем является использование методов сжатия данных.

В статье проанализированы особенности временных рядов оптической толщины аэрозоля, полученной с радиометра MODIS спутника TERRA [4], а также концентрации озона, определенной радиометром OMI спутника Aura [5]. Предложены модификации существующих методов с использованием приведенных особенностей, за счет чего достигнуто увеличение степени сжатия.

Постановка задачи

Для решения поставленных проблем с использованием сжатия предлагаются методы, удовлетворяющие следующим требованиям:

– обеспечение сжатия данных без потерь точности;

– принятие во внимание особенностей рассмотренных экологических показателей, включая представление значений числами с плавающей точкой;

– наличие возможности использования блочных методов;

– допустимо использование методов с ресурсоемкой декомпрессией, так как нагрузка переносится на клиентские машины, которые обладают достаточными вычислительными мощностями.

Следует определить основные источники, порождающие избыточность, и подобрать нужный комплекс методов, удовлетворяющий поставленным требованиям.

Обзор существующих методов

Литература по вопросу сжатия временных рядов с плавающей точкой довольно разрежена. Сообщество разработало ряд схем сжатия неструктурированных данных [6], которые являются эффективными для 3D моделей основанных на точечном представлении. Приведенные алгоритмы не оправдывают себя при сжатии временных рядов экологических показателей, так как не учитывают существующие в них закономерности.

Одним из требований к сжатию временных рядов, является отсутствие потерь, что обусловлено накоплением большой погрешности при анализе. В большинстве случаев гораздо важнее получить точные показатели, чем сэкономить время при вычислениях. В статье не рассмотрены методы, которые хорошо зарекомендовали себя при сжатии многомерных данных с потерями, такие

как метод выделения структур, фрактальные методы, дискретное преобразование Фурье, дискретное косинусное преобразование.

Структура и особенности временных рядов

Временные ряды доступны для каждой ячейки регулярной широтно-долготной решетки земной поверхности за определенный период с заранее заданным шагом по временной шкале. Обладают низкой энтропией. Наблюдается дублирование данных как для близких по времени значений, так и при изменении положения внутри плоскости выбранного участка. Большая часть данных является неопределенной, по причине наличия облачности во время сканирования территории радиометром. Для хранения показателей обычно используются числа с плавающей точкой одинарной точности.

В большинстве случаев обрабатываются заранее подготовленные данные, при этом не требуется передача показателей в реальном времени. Потребность в накоплении данных делает более приоритетным использование блочных алгоритмов сжатия. Однако, дополнительно некоторый прирост коэффициента сжатия, можно получить используя поточные алгоритмы сжатия чисел с плавающей точкой [7]. Степень прироста напрямую зависит от использованного блочного алгоритма.

Благодаря низкой энтропии данных эффективными будут статистические методы. Прирост может дать предварительная модификация структуры при помощи преобразующих методов. Чем сложнее структура временного ряда, тем сильнее оптимальное преобразование данных улучшит сжатие. Следует рассмотреть метод разделения мантисс и экспонент для увеличения однородности показателей. Увеличение длины одинаковых фраз можно достигнуть благодаря использованию различных способов обхода координатной сетки. При этом, следует выделить обход методом зигзаг-сканирования.

Метод Хаффмана

Классический блочный алгоритм сжатия на основе статистических показателей, разработан Девидом Хаффманом в 1952 году. Идея алгоритма состоит в том, что символы с наибольшей частотой получают короткие коды, а с наименьшей частотой более длинные. Этот метод, требует дополнительного прохода по входному блоку для сбора статистики. Существуют адаптивные варианты, не требующие дополнительного прохода [8]. Алгоритм построен так, что присваивает каждому символу код с целым количеством бит. Это порождает наилучшие коды переменной длины, когда вероятности символов алфавита являются степенями числа 2.

Благодаря однозначному отображению исходных элементов на биты сжатой последовательности, метод позволяет использовать оптимизацию, основанную на преобразовании цепочек. Цепочка кодируется идентификатором начала, длиной и символом, из которого эта цепочка состоит. Маркер начала не должен встречается в исходном алфавите. Когда сжатая последовательность занимает больше места, чем исходная цепочка, в выходной поток записываются исходные символы.

Арифметический метод

Арифметический метод присваивает код не отдельным символам, а всей входной последовательности. Это позволяет решить проблему эффективности, когда относительные частоты не являются степенями числа 2 [9].

При данном подходе длинные цепочки одинаковых элементов, лишь косвенно влияют на коэффициент сжатия, путем увеличения вероятности встречаемого символа. Метод может быть обобщён для алфавита с произвольной длинной символа.

Преобразующие оптимизации

Оптимальное преобразование данных дает возможность увеличить эффективность сжатия, за счет увеличения однородности. Приведенные в данной статье оптимизации позволяют достичь прирост коэффициента сжатия при использовании метода Хаффмана с кодированием цепочек. Для арифметического метода результаты останутся неизменными в силу описанных выше особенностей.

В связи с малой энтропией данных для большинства показателей наблюдается одинаковое значение экспонент. Благодаря хранению экспонент отдельно от мантисс удалось получить длинные цепочки одинаковых значений.

Методы обхода координатной сетки

Данные временного ряда экологических показателей можно представить в виде трехмерной сетки значений для пространственно-временных координат. Указанная сетка ограничена прямоугольным параллелепипедом. Стороны соответствуют долготе, широте и значению времени для показателя.

В исходном виде, выполняется обход по временной шкале для определенного значения

координат. Использование иных обходов позволяет достигнуть большей однородности обрабатываемых данных.

Обход строками по координатной плоскости

В сравнении с временной шкалой, данные показателей, с большей вероятностью, имеют одинаковые значения для ближних точек внутри координатной плоскости широты-долготы. Указанную закономерность подтверждает рисунок 1.

На рисунке 1b приведена координатная сетка для значения 0,0782 временного ряда концентрации озона радиометра OMI спутника Aura. Наблюдаемая закономерность выражается в виде вытянутых вдоль плоскости цилиндрических скоплений (рис. 1a). С учетом этого, целесообразно применить обход строками по координатной плоскости.

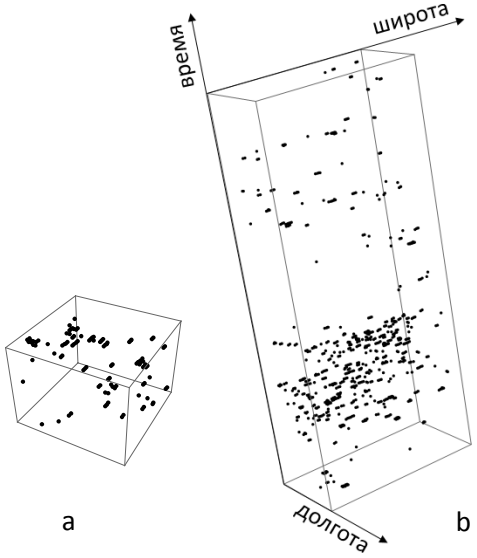

Рисунок 1. – Скопления одинаковых показателей

Метод зигзаг-сканирования

Метод эффективен в случаях, когда значения изменяются не вдоль строк, а по диагонали плоскости. Этот метод, используется в алгоритме сжатия изображений JPEG.

Обход массива показателей начинается с одного угла плоскости и заканчивается в противоположном (рис. 2).

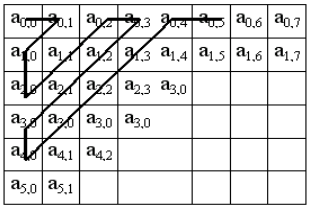

Рисунок 2. – Зигзаг-сканирование

Сравнение эффективности предложенных модификаций

Получены коэффициенты сжатия показателей для следующих временных рядов.

a) Оптическая толщина аэрозоля радиометра MODIS спутника TERRA для области: южная широта 30,0°, северная широта 60,0°, западная долгота 0,0°, восточная долгота 50,0° для уровня широтно-долготной решетки 1,0° x 1,0°, временным интервалом с 01.03.2000 по 01.09.2000 с шагом в 1 день.

b) Концентрация озона радиометра OMI спутника Aura для области: южная широта 43,0°, северная широта 54,0° западная долгота 20,0°, восточная долгота 42,0° для уровня широтно-долготной решетки 0,25° x 0,25° и временным интервалом с 01.01.2005 по 31.07.2005 с шагом в 1 день.

Ниже приведены использованные для сравнения методы.

1. Встроенный в Java SDK алгоритм сжатия без потерь, который основан на методе deflate [10].
2. Арифметический метод для сжатия однобайтовых элементов.
3. Обобщенный арифметический метод для чисел с плавающей точкой одинарной точности.
4. Метод Хаффмана для однобайтовых элементов.
5. Обобщенный метод Хаффмана для чисел с плавающей точкой одинарной точности.
6. Обобщенный метод Хаффмана с кодированием одинаковых последовательностей элементов (обход вдоль временной шкалы).
7. Оптимизированный метод Хаффмана с обходом по координатной плоскости.
8. Метод Хаффмана с обходом по плоскостям с использованием зигзаг-сканирования.

Таблица 1. – Полученные коэффициенты сжатия временных рядов

	Характеристики данных				Результаты работы алгоритмов							
	уникальность (%)	min	max	NA* (%)	1	2	3	4	5	6	7	8
a	0,2	238,2	495,0	46,5	87,91	70,93	93,5	61,74	92,26	92,86	**94,47**	94,44
b	0,6	-0,05	5,0	84,98	74,28	49,66	79,45	45,66	76,39	79,56	**80,79**	80,04

* - количество неопределенных значений

Чуприн В.І., Родрігес Заліпиніс Р.А.
Донецький національний технічний університет

Ниже приведены закономерности, выведенные на основе полученных результатов.

1. В общем случае, арифметический метод оказывается более эффективным чем метод Хаффмана. Это обуславливает большую его популярность на текущий момент.

2. Методы адаптированные для чисел с плавающей точкой работают значительно эффективнее чем те же методы для байтовых элементов. Во-первых, благодаря уменьшению количества бит, требуемых для кодирования каждого отдельного символа. При этом, прирост коэффициента сжатия достигается за счет уменьшения количества символов и увеличения вероятности их появления. Во-вторых, наблюдается уменьшение таблицы перекодировки, которая используется в обоих приведенных методах.

3. Для большого количества цепочек метод Хаффмана оказывается более эффективным благодаря оптимизации, позволяющей кодировать эти цепочки в более короткие, исключая дублирование. Арифметический метод лишь косвенно учитывает длинные последовательности повторяющихся символов.

4. Прирост эффективности сжатия приносят методы обхода координатной плоскости на основе закономерностей представления данных, что достигается оптимальным хранением однородных цепочек.

Выводы

Благодаря практическим исследованиям на основе данных оптической толщины аэрозоля, полученной с радиометра MODIS спутника TERRA и концентрации озона определенной радиометром OMI спутника Aura, продемонстрирована эффективность подхода основанного на методе Хаффмана. Указанный метод был обобщен для элементов произвольной длины с целью применения к 4х байтным числам с плавающей точкой. Для минимизации объема, который занимают дублированные данные, была предложена схема кодирования повторяющихся элементов. Был также апробирован метод обхода данных, позволяющий увеличить длины цепочек одинаковых элементов и тем самым повысить однородность сжимаемых данных. Приведенная схема позволяет увеличить коэффициент сжатия без потери точности исходных данных.

В дальнейшем следует рассмотреть следующие оптимизации:

– различные методы обходов координатной сетки, которые используют статистические показатели, полученные на основе сжимаемых данных при помощи методов динамического программирования;

– кодирование энтропии чисел с плавающей точкой существующими поточными методами сжатия;

– применение подходов, которые бы позволяли предварительно оценивать эффективность использования приведенными методами. В случае высокой неоднородности данных, следует использовать алгоритмы более эффективные в общем случае.

Приведенные подходы могут быть также эффективны для данных обладающих малой энтропией и состоящих из элементов произвольного размера.

Литература

1. Rodriges Zalipynis, R.A. Representing Earth remote sensing data as time series / R.A. Rodriges Zalipynis // Donetsk National Technical University. Series: System analysis and information technology in environmental and social sciences, No.2, 2012. – 212 с. – С. 57 – 71.
2. Родригес Залепинос Р.А. Данные и методы интеллектуального анализа данных для исследования окружающей природной среды [Текст]: научн. журн. / Родригес Залепинос Рамон Антонио // Наук. праці Донецького національного технічного університету. – Сер.: Системный анализ и информационные технологии в науках о природе и обществе, No.1, 2011. – 214 с. – С. 94 – 107.
3. Up in the Air [Электронный ресурс] – Режим доступа: http://www.iwu.edu/magazine/2004/winter/aura.html (14.11.2012).
4. MODIS Website [Электронный ресурс] – Режим доступа: http://modis.gsfc.nasa.gov/ (10.10.2012).
5. OMI Instrument Science [Электронный ресурс] – Режим доступа: http://aura.gsfc.nasa.gov/instruments/omi.html (10.09.2012).
6. Devillers O., Gandoin P.-M. Geometric Compression For Interactive Transmission [Электронный ресурс] – Режим доступа: – http://hal.inria.fr/docs/00/07/27/43/ PDF/RR-3910.pdf (18.10.2012).
7. Lindstrom P., Isenburg M. Fast and Efficient Compression of Floating-Point Data [Электронный ресурс] – Режим доступа: http://dl.acm.org/citation.cfm?id=1187859 (14.10.2012).
8. Ватолин Д., Ратушняк А., Смирнов М. Методы сжатия данных. М.: Диалог-МИФИ, 2003. – 381 с. – С. 31.
9. Сэломон Д. Сжатие данных, изображений и звука, М.: Техносфера, 2004. – 367 с. – С. 62.
10. Class GZIPInputStream [Электронный ресурс] Режим доступа:http://docs.oracle.com/javase/1.5.0/docs/api/java/util/zip/GZIPInputStream.html (14.10.2012).