УДК 004.63 550.3

АЛЬТЕРНАТИВНЫЙ ПОДХОД К СЖАТИЮ И ХРАНЕНИЮ ИСХОДНЫХ ВРЕМЕННЫХ РЯДОВ

Махмудов Е.Р. , Яблочкина К.А.

Камчатский филиал Геофизической службы РАН, mer@emsd.ru

Введение

Основные форматы данных, которые используются геофизиками (SEED/miniSEED, CSS, SEG-Y,WIN) были разработаны достаточно давно и используют идентичный подход к сжатию данных(SEED, CSS, WIN). Этот метод основывается на двух этапах: представление фрагмента временного ряда как списка разницей амплитуд его значений и представление их меньшим количеством битов.

В данной статье предлагается рассмотреть альтернативный подход. Основная идея заключается в представлении массива данных строчными выражениями, подвергающиеся упаковке современными алгоритмами сжатия.

Сравнение алгоритмов сжатия

Среди наиболее известных алгоритмов, свободно применяющихся без патентных ограничений, выделяются следующие:

- deflate - использует комбинацию алгоритма LZ77 и алгоритма Хаффмана, применяется в программах WinZip, gzip
- Bzip2 - реализация алгоритма Барроуза — Уилера блочно-сортировочного сжатия, применяется в программе bzip2 и тд
- LZMA - улучшенный алгоритм сжатия LZ77, дополненный алгоритмом интервального кодирования, применяется в программе 7-Zip
- PPMd – алгоритм основан на контекстном моделировании и предсказании, используется в программах 7-Zip, WinZip и т.д.

Основными характеристиками алгоритмов служат время упаковки и распаковки, степень сжатия фрагмента данных, в качестве основного критерия эффективности их соотношения.

Для оценки характеристик алгоритмов сжатия были взяты несколько суточных файлов.

1) Файл инфразвуковой станции PRT, размер 50.2 Мб, частота дискретизации 100 Гц. Результаты приведены в таблице 1.

Таблица 1. Сравнительные оценки основных параметров характеризующих возможности алгоритмов по данным инфразвуковой станции.

	Deflate	Bzip2	LZMA	PPMd
Коэффициент сжатия	3,19	4,44	3,89	4,78
Коэффициент сжатия, %	100	139	121	150
Время сжатия, с	103	133,1	107,6	29,7
Время сжатия, %	100	77,3	95,7	346
Время распаковки, с	0,76	4,4	1,6	30,5
Время распаковки, %	100	17,2	47,5	2,5
Оценка распространенности для разных языков программирования	5	4	3	1

На основании полученных результатов можно сделать следующие предварительные выводы:

- алгоритм PPMd не подходит для подобных задач ввиду большого времени распаковки

- алгоритм Bzip2 имеет отличный показатель по степени сжатия, но ему требуется больше времени как на упаковку, так и на распаковку.
- алгоритмы Deflate и LZMA имеют хорошее соотношение скорости упаковки/распаковки.

2) Файл сейсмической станции в «тихом» (фоновая активность) режиме работы, 42.2 Мб, частота дискретизации 128

Таблица 2. Сравнительные оценки основных параметров характеризующих возможности алгоритмов по данным «тихой» сейсмической станции.

	Deflate	Bzip2	LZMA
Коэффициент сжатия	7,87	10,09	9,27
Коэффициент сжатия, %	100	78	85
Время сжатия, с	141,2	108	95,2
Время сжатия, %	100	130	148
Время распаковки, с	0,3	3,7	0,6
Время распаковки, %	100	8	50

2) Файл сейсмической станции в «шумном» (перекрытие всего входного диапазона) режиме работы, 56.7 Мб, частота дискретизации 128

Таблица 3. Сравнительные оценки основных параметров характеризующих возможности алгоритмов по данным «шумной» сейсмической станции.

	Deflate	Bzip2	LZMA
Коэффициент сжатия	3,24	4,05	3,63
Коэффициент сжатия, %	100	125	112
Время сжатия, с	118,5	156,6	141,5
Время сжатия, %	100	75	83
Время распаковки, с	1	5,2	1,9
Время распаковки, %	100	19	52

На основании полученных результатов можно сделать следующие выводы:

- Для всех алгоритмов параметры сильно зависят от входящих данных и не могут быть оценены заранее
- алгоритм deflate однозначно лучше по скорости распаковки
- алгоритм Bzip2 имеет лучший коэффициент сжатия
- алгоритм LZMA занимает промежуточное положение, имея хорошие показатели.
- Несмотря на то, что алгоритм LZMA по характеристикам выглядит оптимальным, на сегодняшний день предпочтительнее использование алгоритма bzip2, однако общая тенденция к расширению применения алгоритма LZMA позволит в будущем говорить об его рекомендации к использованию.

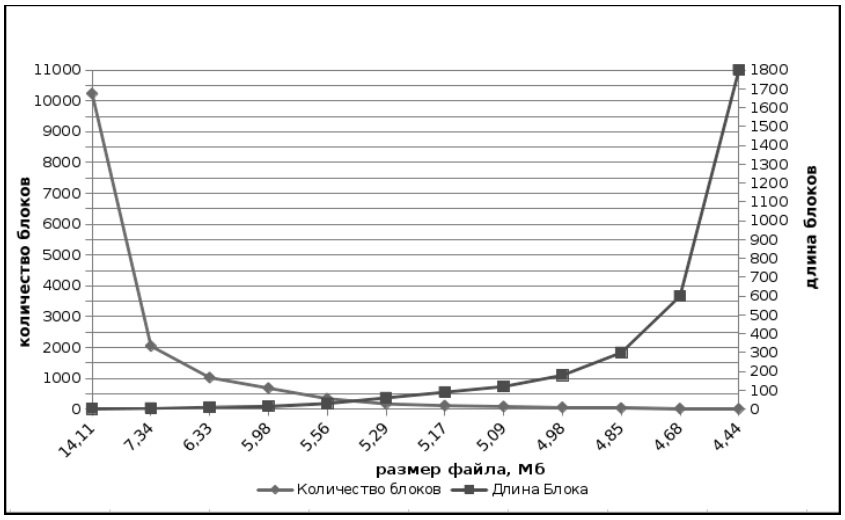
Формат TCTiSe

Результатом реализации идеи стал формат TCTiSe (от Text Compressed Time Series в переводе "сжатые текстовые временные серии", произносится как ТиСиТайз).

Формат имеет очень простую и прозрачную блочную структуру. Каждый блок начинается со строчного идентификатора. Информационные блоки MAIN и META имеют фиксированный размер,

блоки DATA и CUST переменный. Подробная техническая документация представлена на странице в интернете по адресу https://bitbucket.org/john_16/tctise.

Последовательно первым должен следовать основной блок MAIN, который содержит метку однозначно идентифицирующий тип файла, технические сведения и вспомогательную информацию о назначении записи. Затем должны располагаться информационные блоки META, содержащие в себе информацию о канале регистрирующей станции, характеристиках измеряемого параметра. Следующие блоки с данными DATA и блоки расширения CUST могут чередоваться. Блок DATA состоит из контрольной метки и упакованных значений временного ряда. Блок CUST это блок-расширение, структура определяется свободно, при соблюдении базовых полей, может быть использован для хранения второстепенных регистрируемых данных, являться уведомлением о происходящих сбоях и т.п.



Продолжительность блока DATA задается исходя из критерия времени. На рисунке 1 представлен типовой график зависимости количества и длины блоков от размера файла, пересечение линий которых, дает представление об оптимальной длине блока в 45 секунд, однако для удобства рекомендовано использование 30 или 60 секундного блока.

Рис.1 Типовой график зависимости количества и длины блоков от размера файла.

Сравнительное тестирование и выводы

В качестве сравнительного тестирования использованы популярный формат SEED в его вариации miniSEED и формат WIN применяемый в регистраторах Datamark. В таблице 4 указаны значения размеров получаемых файлов для каждого из типа исходных сигналов и форматов.

Таблица 4. Сравнение значений размеров получаемых файлов для каждого из типа исходных сигналов и форматов.

Формат / Сигнал	SEED	WIN	TCTiSe_Gzip	TCTiSe_Bzip2	TCTiSe_LZMA
Сейсмика тихая	6,56	-	6,3	4,5	5,73
Сейсмика шумная	15,0	-	19,4	15,4	17,6
Инфразвук	13	-	17,9	14,5	15,0
Наклономер	-	0,815	0,629	0,499	0,565

Полученные результаты свидетельствует об ощутимом преимуществе формата TCTiSE в двух случаях (40 и 60 процентов) и незначительном проигрыше в двух других (3 и 11 процентов).

Заключение

Формат TCTiSe осуществляет эффективное сжатие временных рядов (в ряде случаев высокую эффективность) за счет современных высокопроизводительных алгоритмов.

Формат TCTiSe имеет крайне простую и ясную внутреннюю структуру, легкую техническую реализацию за счет общедоступных библиотек алгоритмов сжатия, что значительно упрощает процесс интеграции в действующие программные проекты.