УДК 550.34.03

Сравнение кодов переменной длины: коды Элиаса и Фибоначчи применительно к вопросам сжатия данных

© 2008 г. Д.Б.Калошин, Е.В.Башкирев, В.Ю.Бурмин

Институт физики Земли им. О.Ю.Шмидта РАН, г. Москва, Россия

Анализ регистрируемой геофизической информации показывает, что в момент сейсмической активности присутствует высокий разброс вероятности показаний приборов. С целью уменьшения требований к информационным ресурсам ЭВМ и уменьшения уровня потребляемой энергии, при применении автономных систем на базе ЭВМ, используются алгоритмы кодирования с минимальными требованиями к загрузке центрального процессора ЭВМ, и как следствию его энергопотреблению: статический алгоритм Хаффмана, кода Элиаса, Райса, Голомба, Фибоначчи. Применение кодов Фибоначчи для кодирования данной информации позволяет получить выигрыш в размере сжатой информации от 10% до 30% относительно других методов кодирования.

Ключевые слова: сейсмические наблюдение, хранение данных, коды, сжатие данных.

Введение

Наблюдаемое в течение последних 20 лет стремительное увеличение количества используемых вычислительных машин (сопровождаемое неуклонным ростом их возможностей) привело к значительному расширению, как круга пользователей ЭВМ, так и сферы применения компьютеров. Одновременное развитие телекоммуникационных систем, рост пропускной способности и общего количества линий связи, появление и развитие глобальных компьютерных сетей, в настоящее время доступных сотням миллионов пользователей, вызвали быстрое увеличение объемов хранимой и передаваемой в информации.

Получение и хранение геофизической информации в цифровом виде требует все больше и больше ресурсов. Особенно это касается высокочастотных процессов, для которых степень дискретизации данных оказывается высокой. В частности, это относится к сейсмическим данным, особенно к данным высокочастотной сейсмики. Одним из видов сейсмических наблюдений является регистрация землетрясений с помощью автономных сейсмических станций. При таких наблюдениях автономность станций зависит от двух факторов. Первый – это энергопотребление аппаратуры и второй - емкость накопителей информации. Очевидно, что количество накапливаемой информации при автономном режиме можно существенно увеличить за счет сжатия данных. Кроме автономных наблюдений, сжатие данных становится насущным и при передаче данных по телекоммуникационным каналам, где стоимость канала пропорциональна количеству передаваемой информации.

Таким образом, задача хранения и передачи текстовой, графической, звуковой и другой информации в наиболее компактном виде достаточно актуальна.

Преобразование информации в более компактную форму называется сжатием данных. Сжатие данных состоит из двух стадий: моделирование и кодирование. От того насколько компактно будут представлены смоделированные данные, в основном и определяется эффективность алгоритма сжатия.

В настоящей статье представлены результаты сравнения кодов переменной длины: кодов Элиаса и Фибоначчи применительно к вопросам сжатия геофизических данных.

Классификация существующих методов кодирования

В настоящее время, как правило, большинство методы сжатия в сочетании с каким-либо методом энтропийного кодирования, заменяющего символы их кодовыми словами – строками нулей и единиц – так, что более часто встречающимся символам соответствуют более короткие кодовые слова.

Такие методы кодирования известны с конца 1940-х гг. и хорошо изучены. Их можно разбить на два больших класса: префиксные (Хаффмана, Шеннона, Шеннона–Фано) и арифметические.

Префиксные коды называются так потому, что ни одно кодовое слово не является полным началом (т.е. префиксом) никакого другого слова, что гарантирует однозначность декодирования.

Известно много способов построения префиксных кодов: коды Шеннона и Шеннона–Фано почти оптимальны, а код Хаффмана – оптимален среди префиксных кодов.

Так как длина каждого кодового слова выражается целым числом битов, то префиксные коды неэффективны на алфавитах малой мощности (2–8 символов) или при наличии символов с очень большой (более 30–50%) вероятностью появления и по качеству сжатия могут уступать арифметическим.

Арифметические коды не стоят явного соответствия между символами и кодовыми словами; они основаны на других принципах, и обычно применяются в сочетании с методами статистического моделирования для кодирования символов в соответствии с предсказанными вероятностями. Арифметическому кодированию посвящено много публикаций, так как его качество лучше, чем у посимвольного префиксного кодирования, и близко к теоретическому минимуму и при малой мощности алфавита, и при очень неравномерном распределении вероятностей появления символов.

С другой стороны, кодирование и декодирование арифметических кодов при достаточно большой мощности кодируемого алфавита ($|\Sigma| \geq 100$) заметно медленнее кодирования и декодирования префиксных кодов, а разница в качестве сжатия обычно незначительна и не превышает 1%; по этим и ряду других причин считается, что префиксное кодирование более предпочтительно для практического использования.

Классический алгоритм Хаффмана

Алгоритм Хаффмана является одним из классических алгоритмов, известных с 1960-х годов. Использует только частоту появления одинаковых байтов в изображении. Сопоставляет символам входного потока, которые встречаются большее число раз, цепочку бит меньшей длины. И, напротив, символам, встречающимся редко – цепочку большей длины. Для сбора статистики требуется два прохода, что приводит к буферизации входного потока. Классический алгоритм Хаффмана требует записи в файл таблицы соответствия кодируемых символов и кодирующих цепочек, что совершенно неприемлемо при кодировании небольших файлов, что ни только не приводит к отсутствию выигрыша в сжатии, но и порой и увеличивает размер сжатого файла.

Коды переменной длины

Для того чтобы избежать использования таблицы декодирования успешно применяются коды переменной длины: Элиаса, Райса, Голомба, Фибоначчи [*Ватолин и др.,* 2002]. Данные схемы кодирования, в отличии от схемы Хаффмана, изначально не ставят задачу "как минимум не увеличить файл" (данная посылка у Хаффмана подразумевает только сам кодированный блок, без таблицы декодирования), а отталкиваются из посылки резкого разброса вероятности появления символов алфавита.

Авторами было произведено исследование с применением модели Шеннона, на основании которой было выявлено интересная особенность: поведение потока данных с резким разбросом вероятности появления символов алфавита характерно для выполняемых файлов, которые составляют до 50% от среднестатистического набора сжимаемых файлов.

Используя схемы кодирования кодами переменной длины можно эффективно сжимать выполняемые файлы, причем простота реализации декодировщика, в отличие от алгоритмов Хаффмана и тем более, арифметического кодирования, позволяет распаковывать такие файлы при выполнении файла. При использовании файлов, самораспаковывающихся в момент выполнения, можно решить следующие задачи:

– Значительное уменьшение размера выполняемого файла.

– Степень сжатия выполняемого файла с последующей распаковкой файла в момент выполнения будет превышать степень сжатия классических универсальных архиваторов, так как будут учитываться особенности формата выполняемых файлов.

– Увеличит скорость выполнения файла, за счет уменьшения времени считывания самого файла с носителя.

– Позволит обезопасить от неавторизованного изменения и изучения алгоритма работы упакованного файла.

Коды Элиаса γ и Фибоначчи

В основополагающей работе в области кодирования кодами переменной длины П.М.Фенвика "Кодирование целых чисел "разорванными" кодами Элиаса" [*Fenwick*, 1996] предлагается использовать для кодирования коды Элиаса γ. Данные коды позволяют кодировать число K однозначно декодируемой последовательностью размером $2*K+1$ бит.

Данный метод кодирования используется в алгоритме сжатия LZB, а также реализован в архиваторе выполняемых файлов aPack Ю.Ибсена. В настоящий момент, упомянутый архиватор является наилучшим в своем классе программ.

Авторами данной статьи произведен анализ алгоритма, произведено исследование и реализован алгоритм, позволяющий произвести более качественное кодирование с использованием меньших ресурсов компьютера (текст программы дан в приложении, а саму программу можно будет скачать на сайте dsrl.ifz.ru).

Исходное кодируемое число N раскладывается на сумму чисел Фибоначчи f_i (f_1=1, f_2=2, $f_i = f_{i-1} + f_{i-2}$). Известно, что любое число однозначно представимо в виде суммы чисел Фибоначчи. Поэтому можно построить код числа как последовательность битов, каждый из которых указывает на факт наличия в N определенного числа Фибоначчи. Заметим также, что если в N есть f_i, то в нем не может быть f_{i+1}. Поэтому если единичное значение бита указывает на использование какого-то числа, то мы можем обозначить конец записи текущего кода и начала следующего последовательностью из двух единиц. Сравнение двух кодов приведено в таблице.

Число	Кол-во бит	Коды Элиаса γ	Число	Кол-во бит	Коды Фибоначчи
0	1	1	0	2	11
1	3	010	1	3	011
2	3	011	2	4	0011
3	5	00100	3	4	1011
6	5	00111	4	5	00011
7	7	0001000	6	5	01011
14	7	0001111	7	6	000011
15	9	000010000	11	6	101011
30	9	000011111	12	7	0000011
31	11	00000100000	19	7	0101011
62	11	00000111111	20	8	00000011
63	13	0000001000000	32	8	10101011
126	13	0000001111111	33	9	000000011
127	15	000000010000000	53	9	010101011
254	15	000000011111111	54	10	0000000011
			87	10	1010101011
			88	11	00000000011
			142	11	01010101011
			143	12	000000000011
			231	12	101010101011
			232	13	0000000000011
			233	13	1000000000011
			254	13	1000001000011
			1595	16	1010101010101011

Как видно из приведенных таблиц кодирования 255 символ при кодировании кодами Элиаса γ требует 17 бит, что несколько затруднительно для реализации на 16-битных процессорах. Диапазон же кодирования на 16-битных процессорах для кодов Фибоначчи ограничен 1595, что в 6 раз больше, чем у кодов Элиаса γ. Такой широкий диапазон крайне удобен для кодирования алфавитов с нативным размером большим один байт.

Выводы

Данное исследование показало, что использование кодов Фибоначчи более эффективно в использовании, чем коды Элиаса γ.

В настоящее время авторами статьи проводятся исследования в области автоматизированной регистрации и оптимизации хранение регистрируемой геофизической информации. С целью уменьшения требований к информационным ресурсам ЭВМ и уменьшения уровня потребляемой энергии, при применении автономных систем на базе ЭВМ, авторами используются алгоритмы кодирования с минимальными требованиями к загрузке центрального процессора ЭВМ, и как следствию его энергопотреблению: статический алгоритм Хаффмана, кода Элиаса, Райса, Голомба, Фибоначчи. Анализ регистрируемой геофизической информации показывает, что в момент сейсмической активности присутствует высокий разброс вероятности показаний приборов. Применение кодов Фибоначчи для кодирования данной информации позволяет получить выигрыш в размере сжатой информации от 10% до 30% относительно других методов кодирования.

Литература

Ватолин Д., Ратушняк А., Смирнов М., Юкин В. Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео. М.: ДИАЛОГ-МИФИ, 2002.

Fenwick P.M. Punctured elias codes for variable-length coding of the integers. Auckland, Tech. Rep. / Dept. of Comp. Sci., Auckland Univ. 1996. N 137.

ПРИЛОЖЕНИЕ

Исходные тексты программ

1

```
#define MAX_EGAMMA    255 // max in 15 bits
struct
    {
    int encoded;
    int len;
    }egamma_encoded[MAX_EGAMMA];

void init_egamma(void)
    {
#define MAX_EGAMMA_NUMBER     8 // 15
    long int egamma_numbers[MAX_EGAMMA_NUMBER]; // В egamma_numbers будут базовые числа

    egamma_numbers[0]=1l;
    for(int i=1;i<MAX_EGAMMA_NUMBER;i++)
        {
        egamma_numbers[i]=1l<<i;
        egamma_numbers[i]+=egamma_numbers[i-1];
        }

// В egamma_encoded будут уже закодированные числа при помощи чисел Элиаса Гамма
    for(i=0;i<MAX_EGAMMA;i++)     // Цикл для всех egamma_encoded
        {
        int k=i;                  // Сделано специально: резерв 1 символ для 0
        for(int j=0;j<MAX_EGAMMA_NUMBER;j++)
            {
            if(k < egamma_numbers[j])
                { // Наш диапазон
                egamma_encoded[i].encoded = 1 << j;           //  экспонента;
                egamma_encoded[i].encoded|=k - (j ? egamma_numbers[j-1] : 0) << j+1 ;
                                                              // мантисса (собственное значение)
                egamma_encoded[i].len=2*j+1;
                break;
                }
            }
        }
    }

int egamma(int normal_number, int &egamma_number)
    {
    egamma_number=egamma_encoded[normal_number].encoded;
    return egamma_encoded[normal_number].len;
    }
```

2

```
#define MAX_ENCODED_FIBO    35245771    // max in 31+1 bits
struct
    {
    int encoded;
    int len;
    }fibo_encoded[MAX_ENCODED_FIBO];

void init_fibonacci(void)
    {
#define MAX_FIBO_NUMBER      32
    int fibo_numbers[MAX_FIBO_NUMBER]; // В fibo_numbers будут числа Фибоначчи

    fibo_numbers[0]=1;
```

```
    fibo_numbers[1]=2;
    for(int i=2;i<MAX_FIBO_NUMBER;i++)
        {
        fibo_numbers[i]=fibo_numbers[i-1]+fibo_numbers[i-2];
        }

// В fibo_encoded будут уже закодированные числа при помощи чисел Фибоначчи
    for(i=0;i<MAX_ENCODED_FIBO;i++)     // Цикл для всех fibo_encoded
        {
        int k=i+1;                      // Сделано специально: резерв 1 символ для 0
        for(int j=MAX_FIBO_NUMBER-1;j>=0;j--)
            {
            if(k/fibo_numbers[j])
                {
                if(!fibo_encoded[i].encoded)
                    {
                    fibo_encoded[i].encoded|=1<<(j+1); // Флаг окончания числа
                    fibo_encoded[i].len=j+2;
                    }
                fibo_encoded[i].encoded|=1<<j;
                }
            if(!(k=k%fibo_numbers[j])) break;
            }
        }
    }

int fibonacci(int normal_number, int &fibo_number)
    {
    fibo_number=fibo_encoded[normal_number].encoded;
    return fibo_encoded[normal_number].len;
    }
```

Сведения об авторах

КАЛОШИН Денис Борисович – научный сотрудник ИФЗ РАН, 123995, ГСП-5, Москва, Д-242, ул. Большая Грузинская, д. 10. Тел.: (495)-254-68-95.

БУРМИН Валерий Юрьевич – доктор физико-математических наук, заведующий лабораторией ИФЗ РАН, 123995, ГСП-5, Москва, Д-242, ул. Большая Грузинская, д. 10. Тел.: (495)-254-68-95. E-mail: burmin@ifz.ru

БАШКИРЕВ Евгений Валерьевич – научный сотрудник ИФЗ РАН, 123995, ГСП-5, Москва, Д-242, ул. Большая Грузинская, д. 10. Тел.: (495)-254-68-95

Comparison of codes with different length: Elias and Fibonachi codes in application for data compression

D.B. Kaloshin, E.V. Bashkirev, V.Yu. Burmin

Schmidt Institute of Physics of the Earth, Russian Academy of Sciences, Moscow, Russia

Abstract. The analysis of recorded geophysical information shows that at the moment of seismic activity instrument readings are characterized by high data range. Coding algorithms with minimal requirements to the power of the central processor, and therefore its power consumption, are used to reduce the requirements to the informational resources of the computer and power consumption. These include Haffman static algorithm, Elias, Rice, Goloumb, and Fibonachi codes. Application of Fibonachi codes for this information encoding leads to a gain in compression ratio of 10% to 30% relative to other encodings methods.