Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.627

Шумилов
Евгений Анатольевич

Алгоритм сжатия данных на основе сортировки

АВТОРЕФЕРАТ
на соискание степени магистра технических наук
по специальности 1-40 80 03 Вычислительные машины и системы

_____

Научный руководитель
Одинец Дмитрий Николаевич
доцент, кандидат технических наук

_____

Минск 2017

ВВЕДЕНИЕ

В современном обществе сжатие данных – одна из самых актуальных задач информационного мира.

Может показаться что уменьшение размера информации на один процент – не значительно. Но на самом деле это не так, в больших масштабах даже один процент может вытекать в довольно большие суммы для хранения и передачи. То же самое применимо и к передаче данных, особенно если учитывать широкое распространение потоковых и аналогичных сервисов по передаче изображений и видео.

Если для сжатия медиа данных может быть применимо сжатие с потерей информации, то для остальных видов данных потеря даже части может быть равнозначна потери всей информации.

Так как рост глобальных сетей и спектр предоставляемых ими услуг постоянно увеличивается, то вслед за ними и увеличиваются объёмы передаваемой и хранимой информации. Так же успешно выполняются проекты по оцифровке содержимого больших хранилищ информации, например, библиотек, выставочных залов и художественных галерей. Вполне разумно ожидать что текущий рост будет увеличиваться с такой же или большей скоростью.

Таким образом задача хранения и передачи всех типов информации, будь то графическая, текстовая или звуковая в наиболее компактном виде является актуальной.

Существует несколько методов для сжатия без потерь – их можно разделить на категории: статистические и трансформирующие, поточные и блочные методы.

Статистические работают по принципу присваивания более коротких кодов для часто встречающихся последовательностей, в результате этого средняя длина последовательности становится короче. В трансформирующих методах статистические свойства данных используются косвенно. Так же существуют смешанные методы, но их количество не велико.

Все поточные методы могут быть применены и к блокам, но обратное будет неверно. Блочные методы неприменимы к потокам, так как для начала их выполнения необходимо задать длину блока, заполненного данными, которые требуется сжать.

При сжатии данных можно применять разные стратегии, ниже перечислены основные из них.

Первая стратегия – трансформация потока, в её случае описание поступающих данных происходит через уже обработанные («скользящее

окно-словарь»). Сюда будут входить методы для потоков «слов» (LZ-методы), когда комбинации поступающих элементов предсказуемы по уже обработанным комбинациям, и преобразование по таблице (RLE, LPC, DC, MTF) для потоков «элементов», когда нет смысла рассматривать комбинации длиной два и более элемента или запоминать их.

В этом случае не вычисляются никакие вероятности и в результате трансформации может быть сформировано несколько потоков. В случае увеличения суммарного объёма потоков их структура будет улучшена и последующее сжатие может быть осуществлено проще, быстрее и лучше.

Вторая стратегия – статистическая стратегия, используя её два типа методов могут быть применены – адаптивные (поточные) и блочные.

В случае адаптивных методов вычисление вероятностей для поступающих данных будет происходить на основании статистики по уже обработанным данным. Кодирование происходит с использованием этих вычисленных вероятностей. Для потоков «слов» используется семейство PPM-методов, а для потоков «элементов» – адаптивные варианты методов Хаффмана и Шеннона-Фано, арифметического кодирования.

При использовании блочных методов статистика сжатого блока будет отдельно закодирована и добавлена к нему. Для потоков «элементов» могут быть использованы статические варианты методов Хаффмана, Шеннона-Фано и арифметического кодирования, для «слов» – статическое CM.

И третья стратегия – трансформация блока, при её использовании входящие данные будут разбиты на блоки, которые будут трансформированы целиком, а в случае блока однородных данных лучше взять блок целиком. Используемыми методами будут методы сортировки блоками (ST, BWT, PBS).

Как и в случае трансформации потоков несколько блоков может быть сформировано и их структура будет существенно улучшена, даже если их суммарная длина не уменьшится.

Следует заметить, что в случае большой длине однородных данных эффективнее будут блочные методы, в противном случае эффективнее ведут поточные методы. Чем источник сложнее, тем сильнее улучшит сжатие оптимальная трансформация, чем источник проще, тем эффективней прямолинейное статистическое решение.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель данной работы заключается в разработке высокоэффективных методов и алгоритмов неискажающего сжатия данных, которые одновременно с высокой скоростью обладают и высоким качеством сжатия.

Особое значение придавалось достижению сравнительно высокой средней скорости работы алгоритма сжатия на типичном тексте, которая имеет наибольшее практическое значение, при приемлемой скорости сжатия в наихудшем случае.

Качеству сжатия также уделялось пристальное внимание. В некоторых случаях допускалось ухудшение качества на 0.5-1.5%, если это ускоряло алгоритм в 1.5-2 раза.

В данной работе рассматривались только методы сжатия, допускающие эффективную программную реализацию на последовательных ЭВМ, поскольку область применения аппаратных решений очень узкая.

Основные усилия были направлены на изучение, анализ и дальнейшее развитие основных методов, используемых для сжатия изображений: методов первичной обработки, контекстной обработки, методов обхода плоскости, методов сжатия одномерных данных без контекстной корреляции.

Особое внимание уделено методам сортировки:

- сортировка параллельных блоков;
- фрагментирование.

Сортировка производится для уменьшения контекстной корреляции и в общем случае зависит только от размера данных, но не от их содержания.

В процессе исследований были использованы основные положения теории информации, теории кодирования дискретных источников сообщений, комбинаторики.

Как правило, при разработке алгоритмов использовался тот факт, что длина сжимаемых данных естественным образом ограничена сверху: например, ввиду ограниченных объемов носителей информации длина кодируемой последовательности заведомо меньше 2^{64} (около $1.6*10^{19}$).

Предложенные алгоритмы не обязательно являются асимптотически оптимальными, однако близки к таковым с практической точки зрения, обладая при этом рядом преимуществ.

Все разработанные новые методы и новые варианты существующих методов позволяют существенно улучшить как степень сжатия данных, на

которые они ориентированы, так и общую эффективность сжатия таких данных современными вычислительными машинами, в среднем на 5-10 процентов по сравнению с существующими аналогами.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В общей характеристике работы отображены цели и задачи исследования, актуальность темы, описаны методы исследования.

Введение дает увидеть освещение степени разработанности темы и оценку современного состояния решаемой задачи, основание и исходные данные для разработки темы.

В следующей главе рассматривается кодирование источников данных типа «аналоговый сигнал»: линейно-предсказывающее кодирование и субполосное кодирование. Описаны в основном те варианты методов, которые показали высокую эффективность на практике.

Метод LPC применяется обычно для сжатия аналоговых сигналов. И как правило, каждый элемент сигнала отклоняется от своего предсказываемого значения не только из-за «сильных» обусловленных изменений – эволюции, но и из-за «слабых» фоновых колебаний, то есть шума. Поэтому возможно два противоположных типа моделей:

- вклад шума невелик по сравнению с вкладом эволюции;
- вклад эволюции невелик по сравнению с вкладом шума.

В первом случае мы будем предсказывать значение S[i] на основании сложившейся линейной тенденции, во втором – как равное среднему арифметическому h предыдущих элементов.

В главе «Обобщенные методы сортирующих преобразований» описываются методы сортирующих преобразований: сортировка параллельных блоков и фрагментирование. Рассматриваются как простейшие «демонстрационные» варианты, так и достаточно сложные, более эффективные в некоторых важных частных случаях.

Два блока A и B называются параллельными, если каждому элементу A[i] первого блока поставлен в соответствие один элемент B[i] второго блока, и наоборот. Длины блоков L_A и L_B равны: $L_A = L_B = L$. Размеры элементов блоков R_A и R_B могут быть разными.

Основная идея метода PBS состоит в сортировке элементов In[i] входного блока In и их раскладывании в несколько выходных блоков Out_j на основании атрибутов A[i] этих элементов. Атрибут A[i] есть значение функции A, определяемой значениями предшествующих элементов In[j] и/или элементов P[k] из параллельного блока P.

При декодировании осуществляется обратное преобразование: элементы из нескольких блоков Out_j собираются в один результирующий, соответствующий несжатому блоку In.

Чем лучше значения In[i] предсказуемы по значениям A[i], тем эффективнее последующее сжатие блоков Out_j с помощью простых универсальных методов.

В главе «Кодирование источников данных без памяти» рассмотрены некоторые методы сжатия источников данных без памяти: SEM, VQ, ENUC. Показано, что SEM (разделение мантисс и экспонент) является большим семейством методов сжатия, содержащим методы универсального кодирования как частный случай. Обсуждаются четыре основных варианта SEM:

- фиксированная длина экспоненты – фиксированная длина мантиссы;
- фиксированная длина экспоненты – переменная длина мантиссы;
- переменная длина экспоненты – переменная длина мантиссы;
- переменная длина экспоненты – фиксированная длина мантиссы.

ВЫВОДЫ

В работе показано, что:

• разработанные автором способы оптимизации существующих алгоритмов сжатия и соответствующих им алгоритмов разжатия данных на основе сортировки могут их улучшить как по скорости выполнения, так и по количеству потребляемой памяти;

• разработанные автором вариации существующих алгоритмов сжатия данных на основе сортировки в большинстве случаев позволяют улучшить как степень сжатия данных, на которые они ориентированы, так и общую эффективность сжатия таких данных современными вычислительными машинами, в среднем на 5-10 процентов по сравнению с существующими аналогами;

• описаны обобщенные методы сортировки, оптимизированные варианты методов PBS и фрагментирования существенно лучше «демонстрационных» по скорости и требованиям к объему рабочей памяти;

• использование нумерующего кодирования является очень перспективным направлением исследований, в частности в задаче сжатия данных без потерь.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

[1-А.] Шумилов, Е. А. Алгоритм сжатия данных на основе сортировки / Е. А. Шумилов // Алгоритм сжатия данных на основе сортировки : Тезисы докл. к конф. – Гомель , 2016 – С. 20-21.