РАЗРАБОТКА ПРОГРАММЫ-АРХИВАТОРА НА ОСНОВЕ АЛГОРИТМА СЖАТИЯ ДАННЫХ БЕЗ ПОТЕРЬ

Э.Р. Муртазин
Научный руководитель: А.О. Савельев
Томский политехнический университет
erm2@tpu.ru

Введение

Последние 10 лет явно прослеживается тенденция роста объема информации. В отчете исследовательской компании IDC (International Data Corporation), занимающейся изучением мирового рынка информационных технологий и телекоммуникаций, за 2017 год говорится, что объем сгенерированных данных в 2012 году составлял 2,9 зеттабайт, в 2016 году эта цифра возросла до 16 зеттабайт, и к 2025 году предполагается, что объём данных составит 163 зеттабайт [1]. Безусловно, не весь объем информации относится к полезной и требует обработки и анализа, однако объем полезной информации также растет. Именно поэтому возрастает потребность в более рациональном способе хранения и передачи данных. Алгоритмы сжатия данных, а также программы-архиваторы, способные производить сжатие, могут поспособствовать решению этой проблемы.

Сжатие данных

Сжатие данных – это алгоритмическое преобразование данных, производимое с целью уменьшения занимаемого ими объема. Основано на устранении избыточности, содержащейся в исходных данных [2]. Простейшим примером избыточности является повторение в тексте фрагментов (например, слов естественного или машинного языка). Подобная избыточность обычно устраняется заменой повторяющейся последовательности ссылкой на уже закодированный фрагмент с указанием его длины.

Другой вид избыточности связан с тем, что некоторые значения в сжимаемых данных встречаются чаще других. Сокращение объёма данных достигается за счёт замены часто встречающихся данных короткими кодовыми словами, а редких — длинными (энтропийное кодирование). Сжатие данных, не обладающих свойством избыточности принципиально невозможно без потерь.

Метод сжатия без потерь

Метод сжатия без потерь реализуется следующим образом: исходных данных находят какую-либо закономерность и с учётом этой закономерности генерируют вторую последовательность, которая полностью описывает исходную. Например, для кодирования двоичных последовательностей, в которых много нулей и мало единиц, мы можем использовать следующую замену:

$00 \rightarrow 0$
$01 \rightarrow 10$
$10 \rightarrow 110$
$11 \rightarrow 111$

В данном случае исходная последовательность из шестнадцати бит:
00 01 00 00 11 10 00 00
преобразуется в последовательность из тринадцати бит:
0 10 0 0 111 110 0 0

Большинство алгоритмов сжатия без потерь работают в две стадии: на первой генерируется статистическая модель для входящих данных, вторая стадия отображает входящие данные в битовом представлении, используя модель для получения «вероятностных» (то есть часто встречаемых) данных, которые используются чаще, чем «невероятностные».

Коэффициент сжатия

Коэффициент сжатия — основная характеристика алгоритма сжатия. Она определяется как отношение объёма исходных несжатых данных к объёму сжатых данных: $k = S_O / S_C$, где k – коэффициент сжатия, S_O – объем исходных данных, S_C – объем сжатых данных.

Разработка приложения

Итоговая программа разработана на основе кодирования длин серий, написана на языке C# и имеет следующую реализацию:

1) Исходный файл преобразуется в массив байтов.
2) Программа проходит по этому массиву, ищет повторяющиеся байты, а также считает количество их повторений.
3) Программа генерирует новый массив байт, в котором содержится количество повторений и сами байт.

В результате чего исходный массив, состоящий из девяти байт, например,
253 253 253 253 5 5 149 149 149
будет преобразован в массив состоящий из шести байт: 4 253 2 5 3 149

Разархивирование файла происходит путем восстановления исходной цепочки байтов.

Алгоритм кодирования длин серий эффективен для данных с большой избыточностью, однако почти бесполезен для малоизбыточных массивов байт.

Пример работы приложения

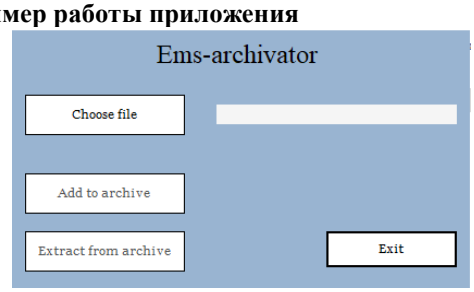

Рис. 1. Пользовательский интерфейс приложения

При попытке заархивировать файл с определенной долей избыточности данная программа архивирует файл с коэффициентом архивации k=2, при этом архиватор WinRar заархивировал с k=3,6.

Рис. 2. Результаты сжатия избыточного .doc файла

Также, были проведены опыты с другими форматами файлов:

Рис. 3. Результаты сжатия избыточного .xls файла

В данном случае k=2.7.

Однако, при попытке заархивировать малоизбыточный файл мы получаем коэффициент сжатия k=1.01, следовательно, в данном случае архивирование бесполезно. К слову, WinRAR хоть и показал больший коэффициент сжатия, однако мы можем наблюдать, что он также плохо справляется с малоизбыточными файлами и сжимает с k=1.11.

Рис. 4. Результаты сжатия малоизбыточного .doc файла

Рис. 5. Сравнение коэффициентов сжатия созданной программы и архиватора WinRAR

При разархивировании файла программа создает его копию с изначальным расширением и пометкой (new) в названии. Разархивированный файл идентичен исходному.

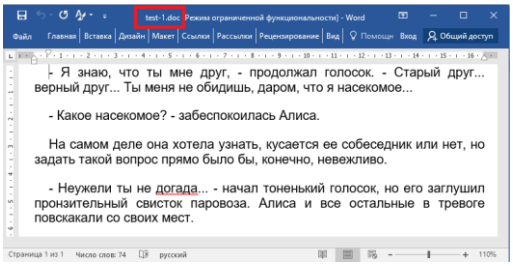

Рис. 6. Содержимое исходного файла

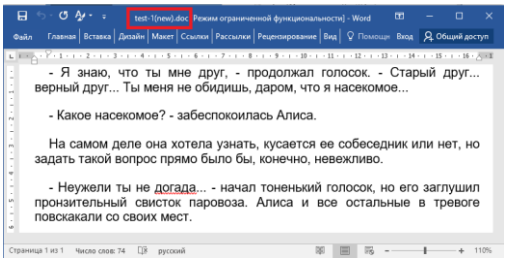

Рис. 7. Содержимое разархивированного файла

Заключение

В результате проведенной работы, был изучен материал по сжатию данных, а также создана программа-архиватор, реализованная на основе алгоритма кодирования длин серий. Следующий этап разработки – усовершенствование программы, путем внедрения различных более совершенных методов сжатия данных, которые будут наиболее эффективны при работе с определенными типами данных. В конечном итоге планируется создать встраиваемую библиотеку для архивации, которая позволит разработчикам эффективнее использовать ресурсы памяти и минимизировать затраты на хранение редко используемых приложением данных - метаданных, файлов конфигураций и т.д.

Список использованных источников

1. THE EVOLUTION OF DATA THROUGH 2025 [Электронный ресурс] / Сайт компании Seagate – URL: https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf (дата обращения 12.10.2018).
2. Д. Сэломон. Сжатие данных, изображений и звука. / Дэвид Сэломон. – М .: Техносфера, 2004. – 368 с.
3. Ватолин Д., Ратушняк А., Смирнов М., Юкин В. Методы сжания данных. Устройство архиваторов, сжатие изобажений и видео. / под общ. ред. О. А. Голубев – М .: ДИАЛОГ-МИФИ, 2003. – 384 с.