ОЦЕНКА ЭФФЕКТИВНОСТИ АЛГОРИТМОВ СЖАТИЯ ТЕКСТОВОЙ ИНФОРМАЦИИ

К. В. Лупан, к. т. н. П. С. Сафронов

Одесский национальный политехнический университет
Украина, г. Одесса
sps@opu.ua

Провден анализ алгоритмов сжатия текстовой информации, содержащей неслучайные данные и обладающей избыточностью. Такие алгоритмы позволяют в несколько раз повысить скорость обмена по каналам связи и во столько же раз сэкономить объем памяти запоминающего устройства, поскольку данные, хранимые на различных носителях и передаваемые по каналам связи, обычно обладают значительной избыточностью, при этом эффективность декодирования достигается за счет словарей фрагментов переменной длины.

Ключевые слова: сжатие данных, избыточность, декодирование, словарь.

Существует проблема эффективного сжатия текстовой информации, обладающей каким-либо лингвистическим признаком: одинаковым алфавитом, одинаковыми популярными комбинациями символов. Критериями качества алгоритмов сжатия, являются степень сжатия данных и скорость их распаковки. Принципиальная возможность сжатия текстовой информации связана с тем, что составляющие текста, буквы и словоформы сильно отличаются по частоте встречаемости в тексте, в то время как их длины слабо связаны между смежными элементами текста. На языке теории информации Шеннона это означает, что энтропия текста (информационная емкость кодирующих текстовых символов) отлична от максимально возможной. Согласно эмпирическому закону Ципфа наблюдается примерно обратная зависимость между частотой встречаемости слов в тексте и их рангами, то есть номерами в порядке убывания частоты. Неравномерность появления элементов текста имеет место не только на уровне букв и слов, но и на уровне частей слов (суффиксов, префиксов) или групп слов.

Целью данной работы является анализ эффективности алгоритмов сжатия текстовой информации, содержащей неслучайные данные и обладающей избыточностью.

Основной принцип алгоритмов сжатия базируется на том, что в любом файле, содержащем неслучайные данные, информация частично повторяется. Используя статистические математические модели, можно определить вероятность повторения определенной комбинации символов. После этого можно создать коды, обозначающие выбранные фразы, и назначить самым часто повторяющимся фразам самые короткие коды. Для этого используются разные техники, например энтропийное кодирование, кодирование повторов и сжатие при помощи словаря [1, 2].

Для выбора того или иного алгоритма сжатия данных необходимо провести анализ типа данных, которые подлежат сжатию. Характерной особенностью большинства «классических» типов данных, с которыми традиционно работают люди, является определенная избыточность. Степень избыточности зависит от типа данных. Степень избыточности видеоданных, как правило, в несколько раз больше, чем графических, а избыточность графических данных, в свою очередь, в несколько раз больше, чем текстовых. Кроме того, степень избыточности данных зависит от принятой системы кодирования. Так, например, можно сказать, что кодирование текстовой информации средствами украинского языка дает в среднем избыточность на 20—30% больше, чем кодирование идентичной информации средствами английского языка. Определение избыточности информации необходимо проводить программным способом, для чего нужно разработать алгоритм, по которому будет разрабатываться программное средство. При обработке информации избыточность также играет важную роль, например, при преобразовании или селекции информации избыточность используют для повышения ее качества. Однако, когда речь идет не об обработке, а о хранении готовых документов или их передаче, то избыточность можно уменьшить, что дает эффект сжатия данных [3, 4].

Адаптивные алгоритмы сжимают текст в процессе однократного его сканирования. Кодирование заключается в нахождении повторяющихся участков текста и замене каждого участка указателем, адресованным той части текста, где такой участок уже встречался. Характерной чертой адаптивных алгоритмов является достаточная их универсальность, то есть возможность работать с любыми, не только текстовыми, данными, ненужность начальной информации о характере данных и их статистике. Это их свойство снижает эффективность сжатия, и достигаемое сжатие, как правило, меньше полученного другими методами. Поскольку адаптивные алгоритмы не используют априорных сведений о статистике, они сравнительно медленны, а произведение времени кодирования на достигаемый коэффициент сжатия на текстовых данных у них в целом ниже, чем у статических алгоритмов. Одним из характерных параметров является так называемая длина адаптивности, которая характеризует «память» алгоритма к ранее введенной части текста. В итоге статистически неоднородные тексты сжимаются неудовлетворительно [2].

Под обслуживающим систему сжатия словарем понимается таблица, используемая алгоритмом сжатия для кодирования или декодирования текста. Словарь может быть локальным и глобальным. Локальный словарь строится для каждого отдельного текста и хранится вместе со сжатым его вариантом. Глобальный составляется на основе нескольких эталонных текстов и служит для сжатия любого текста, по предположению обладающего статистическими свойствами эталонной группы. Простейшей формой словаря является кодовая таблица символов алфавита, ставящая в соответствие каждому символу свой код. Коды выбираются с таким расчетом, чтобы общая длина закодированного ими текста была минимальной. Другой формой словаря может являться словарь фрагментов переменной длины. Такой словарь может быть составлен на основе морфологического анализа текста или с помощью статистического алгоритма, выделяющего в соответствии с некоторым критерием группы символов с наибольшей информационной избыточностью и помещающего их в словарь. Выбранным фрагментам присваиваются коды, которые затем занимают их место в тексте. Коды выбираются для достижения максимального сжатия [4].

Анализ эффективности сжатия текстовой информации показал, что если оценивать алгоритмы сжатия по критерию быстродействия, то необходимо отдельно учитывать скорости кодирования и декодирования. Эффективность алгоритма во многом определяется машинной реализацией и настроечными параметрами. Методы с глобальными словарями дают значительно большую скорость кодирования, чем методы с локальными словарями или адаптивные. С другой стороны, методы, использующие коды фиксированной длины, дают большую скорость декодирования. Если скорости кодирования и декодирования одинаково важны, как это имеет место в системах связи, то наиболее целесообразно использование словарей небольших размеров и кодов фиксированной длины со встроенными указателями, используемыми в адаптивных алгоритмах и алгоритмах, основанных на словарях произвольных фрагментов.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. Шустер Г. Детерминированный хаос: Введение. — Москва: Мир, 1988.

2. Сэломон Д. Сжатие данных, изображения и звука.— Москва: Техносфера, 2004.

3. Цымбал В. П. Теория информации и кодирование.— Киев: Вища школа, 1992.

4. Поддубный А. П., Холуев М. А., Галактионов Н. С. Использование файла в качестве избыточного словаря для препроцессинга данных на основе словарных методов сжатия // Известия высших учебных заведений. Поволжский регион. Технические науки.— 2010 — № 4 (16).— С. 47–54.

———————

K. V. Lupan, P. S. Safronov

Efficiency estimation of text compression algorithms

The authors analyze compression algorithms of the text information containing non-random data and has redundancy. Such algorithms allow increasing the data rate via communication channels and save the memory capacity of the storage device by the same amount, since the data stored on different media and transmitted through communication channels usually have significant redundancy, wherein the decoding efficiency is achieved due to the code books with variable length fragments.

Keywords: data compression, redundancy, decoding, code book.