Московский авиационный институт
(национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Курсовая работа по курсу «Дискретный анализ»: Методы сжатия данных

Студент: В. В. Гринин
Преподаватель: Н. А. Зацепин
Группа: М8О-408Б
Дата:
Оценка:
Подпись:

Москва, 2020

Условие

Необходимо реализовать два известных метода сжатия данных: LZW и Арифметическое кодирование.

Формат запуска должен быть аналогичен формату запуска программы gzip, должны быть поддержаны следующие ключи: -c, -d, -k, -l, -r, -t, -1, -9. Должно поддерживаться указание символа дефиса в качестве стандартного ввода.

Метод решения

Обработка входных данных

Для начала программа обрабатывает аргументы командной строки. Таковыми могут быть ключи или наименования файлов и директорий. Чтобы отделять ключи от обычных наименований, в качестве первого символа для ключей используется «-». В ходе изучения работы ключей программы gzip, были выявлены закономерности в их поведении. А именно:

1. Если был введён ключ -t, то он делает ключи -c, -k, -d, -1, -9 недействительными как сейчас, так и при их будущих вводах.

2. Если был введён ключ -l, то он делает ключи -c, -k, -d, -t, -1, -9 недействительными как сейчас, так и при их будущих вводах.

3. В то же время сочетания -cd, -c1, -c9, -kd, -k1, -k9, а так же сочетания ключа -r со всеми остальными ключами не являются взаимоисключающими.

4. Если же во время обработки ключей встречается неизвестный ключ, то программа прекращает свою работу с соответствующей ошибкой, как и утилита gzip.

В случае если начальным символом не является «-», то оно заносится в красно-чёрное дерево имён файлов/директорий.

Работа с файлами

Проверяется пустота дерева имён файлов/папок – если дерево пустое то программа завершается.

Если дерево не пустое, каждый его элемент сначала рассматривается как директория, затем как файл. Чтобы отделить директорию от файла, используется функция `opendir` из библиотеки `dirent.h`.

Если элемент является директорией, то проверяется активность ключа -r: ключ активирован – идёт работа с директорией, нет – пропуск файла. Если элемент является файлом, то никаких проверок не проводится и начинается непосредственная работа с ним.

Подготовка к сжатию

Если у имени файла есть суффикс «.gz», то файл не обрабатывается. Программа по итогу выведет соответствующее сообщение. Если этого суффикса нет, то, если отсутствует ключ -c, проверяется наличие файла с тем же именем, но при этом ещё и с суффиксом. Если такой файл существует – пользователю предлагают перезаписать этот файл. Если пользователь откажется, то работа с данным файлом прекращается.

Если отсутствуют ключи -1 и -9, то файл сжимается сразу двумя методами. По итогу сжатия получаем два временных файла. Далее сравниваются размеры файлов, где выбирается наименьший, который впоследствии и становится результатом работы программы. Тот файл, который больше, просто удаляется. Если применяется один из этих ключей, то программа использует один из двух алгоритмов. Для ключа -1 это LZW, а для ключа -9 это арифметическое кодирование.

В случае если какой-то из алгоритмов дал сбой, то прекращается работа с файлом и выводится соответствующая ошибка.

Подготовка к распаковке

Для распаковки используется ключ -d. Проверяется наличие у файла суффикса «.gz». Если он не имеется, то работа с файлом прекращается и выводится соответствующее сообщение. Если имеется, то при отсутствии -t и -c проверяется наличие файла с тем же именем, но без суффикса «.gz». При наличии такого файла, программа запрашивает разрешение на перезапись. В случае отказа завершается работа с файлом. Когда такого файла нет или пользователь дал согласие, то работа продолжается. Каждый архив, сжатый этой программой имеется первые несколько служебных байт, которые содержат в себе информацию о алгоритме сжатия и размере исходного файла. При считывании первого байта определяется алгоритм сжатия. Для LZW это символ «L», для арифметического кодирования это «A». Если это другой символ, то работа с файлом прекращается и выводится уведомление об ошибке. В случае неудачного завершения алгоритма, выводится соответствующее сообщение. При отсутствии ключей -t и -c удаляется файл, в который записывались данные после распаковки, и работа с файлом прекращается. Далее при отсутствии ключей -t, -c и -k, удаляется изначальный архив, а при отсутствии ключей -t и -c временный файл для распаковки переименовывается и получает имя изначального архива без расширения «.gz».

Получение информации об архиве

В случае указания ключа -l производятся следующие действия. Для начала программа считывает первый байт. В случае, если он не совпадает с символами, указывающими на метод сжатия, то выводится сообщение об ошибке и завершается работа с файлом.

Далее читается 8 байт, в которые помещается размер файла до сжатия. После этого программа считывает размер архива, и вычисляется процент сжатия. Далее выводятся размер сжатого файла, размер до компрессии, процент сжатия в полуинтервале

$[-100\%; 100\%)$ и имя файла до архивации (если файл имеет расширение «.gz», то имя выводится без этого расширения, в противном случае выводится имя архива).

Далее незакодированный файл будет упоминаться как файл, а закодированный файл как архив.

LZW компрессия

Компрессия методом LZW происходит по следующему принципу: из размера файла определяется верхняя граница буфера, в котором будут храниться слова, и каким количеством байт будут кодироваться слова, после чего строится префиксное дерево из всех односимвольных слов-символов ASCII. В архив записывается 9 байт информации, первый из которых – это указание метода архивации, а остальные 8 – размер изначального файла. Далее читается первый символ файла, и в архив записывается код соответствующего слова. Полученный символ указывает на узел, в который будет добавлен следующий символ. Затем символы считываются до создания новой вершины в префиксном дереве, а в архив записывается код вершины, предшествующей новой. Последняя буква, полученная до добавления вершины заносится в буфер. После чего из корня ищется вершина, к которой ведёт эта буква, и процесс повторяется вплоть до окончания символов в файле или создания максимального количества вершин, которое сможет прочитать декомпрессор. Если произошло второе, то в архив записывается 0, (никак иначе на этом этапе он быть записан не может), ранее установленным количеством байт, из префиксного дерева удаляются все вершины кроме корневой и потомков первого рода, после чего процесс компрессии начинается заново, но позиции в исходном файле и архиве не получают откат. Если на каком-либо этапе компрессии возникает ошибка, его работа прекращается, и выводится соответствующая ошибка.

LZW декомпрессия

Декомпрессия начинается с прочтения размера изначального файла из архива, который необходим для определения нужно количества байт для прочтения слова и проверки на безошибочность декомпрессии. Далее, из архива считываются только коды слов определённого ранее размера. После считывания первого слова создаётся красно-чёрное дерево, и в него записываются все односимвольные слова из ASCII символов. Далее, по полученному коду в дереве находится необходимая строка, и она записывается в файл. После чего этот символ записывается во временное слово. Далее алгоритм считывает коды из архива. При обработке кодов возможны 4 ситуации:

1. Код входит в список полученных слов. В этом случае в красно-чёрном дереве ищется слово с необходимым кодом, и это слово записывается в файл, после чего в красно-чёрное дерево записывается новое слово, которое является предыдущим словом, к которому добавили первую букву только что полученного. Декомпрессия продолжается.

2. Код не входит в красно-чёрное дерево, но он является следующим по счёту, следовательно это слово можно интерпретировать как предыдущее, к которому дописали в конце букву, с которой оно начинается. Полученное слово записывается в файл и в красно-чёрное дерево, после чего декомпрессия продолжается.

3. Код не входит в красно-чёрное дерево и не является следующим на подходе, следовательно архив повреждён. Декомпрессия прекращается, и выводится соответствующее сообщение.

4. Код равен 0. Красно-чёрное дерево очищается, и процесс декомпрессии начинается сначала, но позиции в файле и архиве не получают откат.

По окончании чтения архива, количество байт, которое было в изначальном файле, сверяется с тем, сколько было записано в его новую версию. При несовпадении выводится соответствующее сообщение, и декомпрессия завершается неудачно.

Арифметическая компрессия

В теории арифметическое сжатие описывается достаточно просто. У нас имеется промежуток от 0 до 1. Имеется таблица частот, в которой содержится информация о том, как часто встречается тот или иной символ. Промежуток разделяется на множество отрезков, каждый из которых представляет собой какой-либо символ. При считывании символа, мы переходим к его отрезку. Далее цикл повторяется, но уже с новыми границами, заданными этим символом. На практике же мы неизбежно сталкиваемся с машинным эпсилон, поэтому следует попробовать написать всё в целых числах.

На вход мы получаем файл. Создаём свой файл, в который мы заносим первые 9 байт. 1 байт обозначит тип сжатия, остальные 8 байт - размер исходного файла. По умолчанию у каждого символа частота выставлена на единицу.

Каждый символ кодируется по следующей схеме:

1. Рассчёт границ символа по частоте его появления;

2. Кодировка символа посредством цепочки манипуляций над границами. Если отрезок лежит в верхней половине допустимых значений - пишем бит равный единице. Если лежит в нижней половине - пишем бит равный нулю. Если лежит где-то по центру - увеличиваем счётчик битов, которые будут выставлены вслед за следующим битом с отличным от него значением. Если не выполняется ни одно из этих условий, т.е. получившийся отрезок достаточно большой, то кодировка завершается. Иначе - увеличиваем границы в 2 раза. По сути это аналогично побитовому сдвигу влево.

3. Обновление таблицы частот. Если случилось переполнение, то масштабируем частоты, деля их на два и пересчитывая накопленные частоты. После этого производится сортировка таблицы, чтобы ускорить работу с ней.

Арифметическая декомпрессия

В теории мы получаем число, в котором закодированы символы. Далее мы определяем в каком отрезке лежит это число, благодаря чему узнаём о закодированном символе. Далее мы выбираем новые границы, а именно границы того отрезка. Разбиваем этот отрезок также на несколько частей, следуя таблице частот и аналогично узнаём следующий символ. Однако в текущей реализации используются целые числа, поэтому и декодирование немного отличается.

В первую очередь мы получаем такое же число. По нему мы также определяем границы, однако способ их нахождения несколько отличается - вместо того, чтобы сразу их узнать, мы сначала находим накопленную частоту и уже по ней определяем границы и закодированный символ. После этого мы по аналогичной схеме, как в кодировании символа, проводим манипуляции над границами и таким образом убираем ненужные биты. Далее мы начинаем декодировать следующий символ.

После декодирования символа мы также обновляем таблицу частот.

Описание файлов программы

Код программы разбит на 13 файлов:

1. ACC.h - Содержит перечисление методов и описание класса TACC, необходимого для работы арифметической компрессии и декомпрессии.
2. ACC.cpp - Содержит реализацию всех методов класса TACC.
3. BFile.h - Содержит перечисление методов и описание класса TOutBinary и класса TInBinary, необходимых для записи в файл и чтения из файла соответственно.
4. BFile.cpp - Содержит реализацию всех методов классов TOutBinary и TInBinary.
5. Globals.h - Содержит в себе все необходимые глобальные переменные и библиотеки используемые несколькими файлами.
6. LZW.h - Содержит перечисление методов и описание класса TLZW, необходимого для работы алгоритма LZW.
7. LZW.cpp - Содержит реализацию всех методов класса TLZW.
8. main_help.h - Содержит в себе перечисление и описание всех функций необходимых для препроцессинга перед началом работы алгоритмов компрессии и декомпрессии.
9. main_help.cpp - Содержит реализацию всех функций, необходимых для препроцессинга, описанных в файле main_help.h.
10. Prefix.h - Содержит перечисление методов и описание класса TPrefix, необходимого для работы LZW компрессии.

11. Prefix.cpp - Содержит реализацию всех методов класса TPrefix.
12. main.cpp - Содержит в себе алгоритм чтения файлов и ключей.
13. Makefile - Файл для сборки программы.

Основные типы данных

1. TArithmetic - класс, описывающий работу арифметического алгоритма компрессии и декомпрессии.
2. TOutBinary - класс обеспечивающий запись необходимого количества байт в файл.
3. TInBinary - класс обеспечивающий считывание необходимого количества байт из файла.
4. TLZW - класс, описывающий работу алгоритма LZW.
5. TPrefix - класс, обеспечивающий построение префиксного дерева для LZW сжатия.

Описание методов и функций программы

Основные свойства и методы класса TACC

public:

1. bool Compress (const char*, const char*) - сжатие файла;
2. bool Decompress (const char*, const char*) - распаковка файла;
3. TACC() - конструктор, в котором задаются начальные значения для последующей работы со сжатием/распаковкой файла;

private:

1. bool chError - флаг ошибки при распаковке файла;
2. unsigned char indexToChar [NO_OF_SYMBOLS] - таблица перевода из индексов к символам;
3. int charToIndex [NO_OF_CHARS] - таблица перевода из символов в индексы;
4. int cumFreq [NO_OF_SYMBOLS + 1] - массив накопленных частот. Нужен для определения границ;
5. int freq [NO_OF_SYMBOLS + 1] - массив частот. В нём хранится число появлений тех или иных символов;

6. long low - нижняя граница отрезка;
7. long high - верхняя граница отрезка;
8. long value - число, которое лежит в отрезке;
9. long bitsToFollow - количество бит, которые надо пустить в след за следующим выставляемым битом;
10. int buffer - буффер для работы с файлом;
11. int bitsToGo - число битов, которые ещё можно загрузить в буффер;
12. int garbageBits - счётчик плохих битов при распаковке файла. Как только их становится слишком много - распаковка отменяется и выводится сообщение об этом;
13. FILE *out - файл, в который мы записываем;
14. FILE *in - файл, из которого мы считываем;
15. void UpdateModel (int) - обновление модели под новый символ;
16. void StartInputingBits() - подготовка к побитовому вводу;
17. void StartOutputingBits() - подготовка к побитовому выводу;
18. void EncodeSymbol (int) - кодировка символа;
19. void StartEncoding() - подготовка к сжатию;
20. void DoneEncoding() - завершение кодирования. Загрузка последних битов в буффер;
21. void StartDecoding() - подготовка к распаковке;
22. int DecodeSymbol() - распаковка символа;
23. int InputBit() - получение одного бита из файла;
24. void OutputBit(int) - отправление одного бита в файл;
25. void DoneOutputingBits() - отправление последних битов в файл;
26. void OutputBitPlusFollow(int) - вывод указанного бита и отложенных ранее;

Основные свойства и методы класса TOutBinary

public:

1. TOutBinary() - задаёт начальные значения. Файл не будет открыт.
2. bool Open(std::string*) - открывает файл;
3. bool Close() - закрывает файл;
4. bool Write(const char*, size_t) - запись в файл;
5. bool WriteBin(size_t bit) - запись бита в файл;
6. unsigned long long SizeFile() - подсчёт размера файла;
7. friend bool operator « (TOutBinary& file, size_t const &bit) - запись бита в файл;

private:

1. std::ofstream out - файл вывода;
2. std::string name - имя файла;
3. unsigned char head - маска для заноса бита в block;
4. unsigned char block - временный буффер для хранения и записи битов в файл;

Основные свойства и методы класса TInBinary

public:

1. TInBinary() - задаёт начальные значения. Файл не будет открыт.
2. bool Open(std::string*) - открывает файл;
3. bool Close() - закрывает файл;
4. bool Read(char*, size_t) - считывает из файла некоторое количество байт;
5. bool ReadBin(char* bit) - считывает из файла один бит;
6. unsigned long long SizeFile() - подсчёт размера файла;
7. friend bool operator » (TInBinary& iFile, char &bit) - получение бита из файла;

private:

1. std::ifstream in - файл вывода;

2. std::string name - имя файла;
3. unsigned char head - маска для заноса бита в block;
4. unsigned char block - временный буффер для хранения битов из файла, через него получают биты;

Основные свойства и методы класса TLZW

public:

1. TLZW(TInBinary*, TOutBinary*) - Конструктор класса. Передаются файл для чтения и файл для записи.
2. bool Compress(std::string) - Производит компрессию данных. На вход получает имя файла для компрессии. В случае успешного выполнения возвращает true, иначе false.
3. bool Decompress(std::string) - Производит декомпрессию данных. На вход получает имя файла для декомпрессии. В случае успешного выполнения возвращает true, иначе false.
4. ∼TLZW() - Стандартный деконструктор.

private:

1. TInBinary* ForRead - Файл для чтения.
2. TOutBinary* ForWrite - Файл для записи.
3. TPrefix* CompressionTree - Префиксное дерево для хранения слов при компрессии.
4. std::map<unsigned long long int, std::string> DecompressionTree - Красно-чёрное дерево для хранения слов при декомпрессии.

Основные свойства и методы класса TPrefix

public:

1. TPrefix(TInBinary*, TOutBinary*) - Конструктор для корневой вершины. Передаются файл для чтения и файл для записи.
2. TPrefix() - Конструктор для всех прочих вершин.
3. int Update(char) - Добавление вершины из других вершин. Возвращает коды ошибок или успехов.

4. int UpdateForRoot() - Добавление вершины из корня. Возвращает коды ошибок или успехов.
5. void Clear(bool) - Очистка дерева после переполнения.
6. ~TPrefix() - Стандартный деконструктор.

private:

1. std::vector<std::pair<char, TPrefix*>> Next - Вектор потомков вершины и путей в них.
2. unsigned long long int NumberOfWord - Номер слова в данном узле.
3. static char LastLetter - Последняя прочитанная буква. Необходима для построения нового слова.
4. static unsigned long long int NeedToRead - Вспомогательная переменная для чтения нужного кол-ва символов.
5. static unsigned long long int LastNumber - Номер следующего добавленного слова.
6. static unsigned long long int Border - Максимальная граница количества слов перед очисткой дерева.
7. static TInBinary* ForRead - Файл для чтения.
8. static TOutBinary* ForWrite - Файл для записи
9. static unsigned short int Bites - Количество байт, необходимое для кодирования слова.

Прочие функции

1. bool KeyManager(std::string) - Обрабатывает полученные ключи. В случае получения неизвестного ключа возвращает false, иначе true.
2. bool DifferensOfSizes(TInBinary*, std::string) - вывод для каждого файла размера сжатого, оригинального, коэффициента сжатия(%) и имя оригинального файла(ключ l). В случае повреждения. архива возвращает false, иначе true.
3. void WorkWithDirectory(std::string) - работает с директорией (ключ r).
4. void WorkWithFile(std::string) - работает с файлом (определяет наличие файла, принимает решение о компрессии или декомпрессии, выполняет прочие ключи).
5. bool IsDirectory(std::string, bool) - Проверяет, является ли файл директорией. Если файл является директорией, возвращает true, иначе false.

6. void PrintDirectoryErrors(std::string) - Уведомляет об ошибках.
7. bool IsArchive(std::string) - Проверяет, является ли файл архивом. Если файл является архивом, возвращает true, иначе false.
8. void Rename(std::string, std::string) - Изменяет название файла после успешной компрессии или декомпрессии.
9. void Delete(std::string) - Удаляет временный файл.
10. void MainDecompress(TInBinary*, std::string) - Отвечает за подготовку декомпрессинга.
11. void MainCompress(TInBinary*, std::string) - Отвечает за подготовку компрессинга.
12. unsigned long long int LZWCompress(TInBinary*, std::string, TOutBinary*) - Подготавливает LZW компрессию. Возвращает размер нового файла.
13. unsigned long long int ArithmeticCompress(TInBinary*, std::string) - Подготавливает арифметический компрессию. Возвращает размер нового файла.
14. void KeepSmall(unsigned long long int, unsigned long long int, std::string) - Сохраняет архив самого малого размера.
15. int main(int, char*) - Осуществляет чтение входных данных.

Исходный код

Тест производительности

В качестве теста производительности использовался файл размером, наполненный случайными символами английского алфавита.

	Размер сжатого файла (б)	Эффективность сжатия	Максимальное потребление памяти (Мб)	Время компрессии (с)	Время декомпрессии (с)
gzip					
Оба алгоритма					
LZW					
Арифм. кодирование					

Выводы

В процессе выполнения данной работы я получил некоторые знания и навыки связанные с компрессией и декомпрессией файлов. Так же я закрепил полученные ранее знания о работе с префиксными деревьями. Были освоены новые приёмы работы с файлами и получены базовые навыки для работ с директориями.

LZW кодирование получилось понять и реализовать за гораздо меньшее время и написав меньше строчек кода чем с арифметическим кодированием, но при этом арифметическое кодирование сжимает данные гораздо лучше.

Список литературы

1. Арифметическое кодирование [Электронный ресурс]: mf.grsu.by URL: http://mf.grsu.by/UchProc/livak/po/comprsite/theory_arithmetic.html (дата обращения 14.07.2020)

2. Арифметическое кодирование [Электронный ресурс]: habr.com URL: https://habr.com/ru/post/130531/ (дата обращения 26.08.2020)

3. Алгоритм LZW [Электронный ресурс]: mf.grsu.by URL: http://mf.grsu.by/UchProc/livak/po/comprsite/theory_lzw.html (дата обращения 26.08.2020)