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: А. М. Титеев
Преподаватель: Н. А. Зацепин
Группа: М8О-408Б
Дата:
Оценка:
Подпись:

Москва, 2020

Условие

Необходимо реализовать два известных метода сжатия данных для сжатия одного файла.

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

Как и требуется в условии запуск программы аналогичен запуску утилиты gzip: ./main <ключи> <файлы> <ключи> <файлы> ...

Препроцессинг

На первом этапе работы программа определяет наличие в поступившей строке ключей, директорий и файлов. При обработке ключей учитывается их взаимоперекрываение, как в утилите gzip: l и r имеют наибольший приоритет, далее идёт ключ t, после чего остальные. В случае, если новый ключ перекрывает по логике утилиты некоторые из уже имеющихся, то эти ключи деактивируются.

Если полученное слово из стандартного ввода не является ключом, то программа проверяет наличие директории с таким именем. Если такой директории нет, то считается, что это имя файла, и оно заносится в список файлов. Если директория с таким именем существует и подключён ключ r, то все файлы внутри этой директории добавляются в список.

После с файлами ведётся работа согласно введённым ключам.

Арифметический алгоритм

В этой программе реализованная целочисленная арифметика. Причина проста - машинам лучше и проще работать с целыми числами, чем с числами с плавающими точками. Например, наличие машинного эпсилон.

Перед обяъснением алгоритма стоит упомянуть ряд моментов. Во-первых, так как мы имеем дело с целыми числами, это значит, что у нас нет возможности бесконечно переходить к меньшим отрезкам. Поэтому в данной реализации работа с отрезками сводится к их масштабированию, параллельно записывая нужные биты. Во-вторых, для работы внутри рассматриваемого отрезка используется таблица частот, которая содержит в себе информацию о том, насколько часто встречается тот или иной символ. Это необходимо для нужного разбиения на ещё более мелкие отрезки.

Данная реализация делится на несколько фаз:

- Подготовительная фаза - производится инициализация буфера, куда будут записываться биты, счётчиков, начального отрезка и таблицы частот. Отдельно стоит упомянуть, что начальные значения частот у каждого символа изначально стоят

как единицы. Это необходимо корректной работы с отрезком. Помимо этого также в файл заносится служебная информация, где указано, каким алгоритмом сжали и какой был размер исходного файла.

- Основная фаза - производится кодировка символов. Каждому символу в начале сопоставляется соответствующий маленький отрезок, с которым в дальнейшем происходит масштабирование и смещение, дабы избежать переполнения. В процессе заносятся соответствующие биты. После того, как дальнейшее масштабирование становится невозможным, кодировка символа завершается. Обновляется таблица частот в соответствии с символом.

- Завершающая фаза - когда символы закодированы, программа кодирует EOF и дозаписывает недостающие биты. На этом работа с файлом завершается.

В процессе декомпрессии алгоритм будет часто обращаться к таблице частот, поэтому надо максимально сильно сократь работу с ней, чтобы алгоритм не просел по времени. Для этого есть довольно простое решение - в процессе декомпресси во время обновления таблицы будет производиться её небольшая сортировка. Сортировка сводится к тому, чтобы часто попадающиеся символы были ближе к началу таблицы. Это и ускоряет декомпрессию.

Аналогично компрессии, декомпрессию можно разделить на несколько фаз:

- Подготовительная фаза - почти аналогично компрессии, но с некоторыми отличиями. Инициализируется специальная переменная, которая будет содержать в себе биты, которые были в сжатом файле. Также в файл ничего не записывается, а производится лишь чтение.

- Основная фаза - при помощи специальной переменной определяем символ. Проводим такие же манипуляции с отрезком, как и в сжатии, параллельно загружая новые биты из файла и отбрасывая уже лишние. После этого аналогично компрессии обновляется таблица частот.

- Завершающая фаза - после получения кода об EOF декомпрессия завершается. Весь результат декомпрессии по ходу заносится в новый файл.

LZ77

Описание файлов программы

Код программы разбит на 9 файлов:

1. Arithmetic.h - Описывает класс Arithmetic, в котором заключён соответствующий алгоритм кодирования.

2. Arithmetic.cpp - Реализует класс Arithmetic.

3. Globals.h - Файл с общими библиотеками.
4. LZ77.h - Описывает класс LZ77, в котором заключён соответствующий алгоритм кодирования.
5. LZ77.cpp - Реализует класс LZ77.
6. main.cpp - Основной файл, отвечающий за чтение входных данных и принятие действий относительно каждого файла.
7. Makefile - Файл для сборки программы.
8. preprocessing.h - Содержит прототипы функций, необходимых для обработки данных перед компрессией/декомпрессией.
9. preprocessing.cpp - Реализует все функции из соответствующей библиотеки.

Основные типы данных

1. Arithmetic - Реализует соответствующий алгоритм.
2. LZ77 - Реализует соответствующий алгоритм.

Описание методов и функций программы

Основные свойства и методы класса Arithmetic

public:

1. bool Compress (const char*, const char*) - сжатие файла;
2. bool Decompress (const char*, const char*) - распаковка файла;
3. Arithmetic() - 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 буфер.
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 буфер.
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 TLZ77

public:

1. LZ77() - стандартный конструктор
2. LZ77(IStruct s) - конструктор через вспомогательную структуру IStruct
3. InitEncode() - инициализирует данные необходимые для сжатия
4. Compress(std::string in_str, std::string out_str) - сжатие файла
5. Decompress(std::string in_str, std::string out_str) - распаковка файла
6. ~LZ77() - деструктор

private:

1. LoadDict(unsigned int dictpos) - загрузка словаря из файла в циклический буфер на позиции dictpos
2. DeleteData(unsigned int dictpos) - удаления всех ссылок на удаляемый сектор с началом в dictpos
3. HashData(unsigned int dictpos, unsigned int bytestodo) - хэширование и запись ссылок на возможное преведущее совпадение в словаре
4. FindMatch(unsigned int dictpos, unsigned int startlen) - поиск максимального совпадения в словаре с позицией dictpos, не меньше чем startlen
5. DictSearch(unsigned int dictpos, unsigned int bytestodo) - кодирование считанного сектора с началом в dictpos и длинной bytestodo
6. SendChar(unsigned int character) - кодирование символа character
7. SendMatch(unsigned int matchlen, unsigned int matchdistance) - кодирование пары <matchlen, matchdistance>
8. ReadBits(unsigned int numbits) - считывание numbits битов из файла
9. SendBits(unsigned int bits, unsigned int numbits) - отправка numbits битов записанных в bits в файл
10. const int compressFloor - минимальное совпадение, для записи в виде<length,offset>
11. const int comparesCeil - максимальное число раз которое ищется совпадение в FindMatch
12. const int CHARBITS - сколькими битами кодируется символ

13. const int MATCHBITS - сколькими битами кодируется длина совпадения
14. const int DICTBITS - сколькими битами кодируется длина словаря(offset)
15. const int HASHBITS - сколько бит в хэше
16. const int SECTORBITS - сколько бит в секторе
17. const unsigned int MAXMATCH - максимальная кодируемая длина совпадения
18. const unsigned int DICTSIZE - размер словаря
19. const unsigned int HASHSIZE - размер хэша
20. const unsigned int SHIFTBITS - на сколько происходит сдвиг при хэшировании
21. const unsigned int SECTORLEN - размер сектора
22. const unsigned int SECTORAND - нужен для определения к какому сектору относится то или иное место в словаре
23. unsigned char* dict - ссылка на словарь размером DICTSIZE
24. unsigned int *hash - ссылка на хэш размером HASHSIZE
25. unsigned int *nextlink - ссылка на массив, на каждой позиции которого хранится позиция предыдущего вхождения подстроки с совпадающим хэшем
26. unsigned int matchlength - длина совпадения, применяется в FindMatch, DictSearch
27. unsigned int matchpos - позиция совпадения, применяется там же
28. unsigned int bitbuf - буфер, который используется для записи и чтения бит из файла
29. unsigned int bitsin - сколько битов находится в буфере в данный момент
30. unsigned int masks[17] - маски для побитового чтения/записи
31. FILE *infile, *outfile; - файлы из которых идёт считывание/запись

Прочие функции

1. void Compress(std::string) - Создает необходимые для компрессии данные (временные файлы).
2. void Decompress(std::string) - Создает необходимые для декомпрессии данные (классы и временные файлы).

3. bool ActivateKeys(std::string) - Активирует или деактивирует указанные в аргументах ключи. Из-за некорректного ключа функция возвращает false и завершает работу программы.

4. void ArchiveInfo(std::string) - Выводит информацию об архиве.

5. unsigned long long int CompressA(std::string) - Создаёт класс для арифметического кодирования и работает с ним. При некорректной работе алгоритма возвращает 0, иначе количество байт в получившемся файле.

6. unsigned long long int CompressL(std::string) - Создаёт класс для кодирования LZ77 и работает с ним. При некорректной работе алгоритма возвращает 0, иначе количество байт в получившемся файле.

7. bool ArchiveCheck(std::string) - Проверка на наличие у файла суффикса «.gz».

8. bool DirectoryCheck(std::string, bool) - Проверка на работоспособность директории.

9. void GetFiles(std::string, std::map<std::string, int>*) - Записывает все имена файлов директории в отдельное красно-чёрное дерево.

10. void SaveBest(std::string, unsigned long long int, unsigned long long int) - При отсутствии ключей 1 и 9 выбирает наименьший из полученных компрессией файлов и удаляет наибольший.

11. void ShowErrors(std::string) - Вывод сообщений о возникших ошибках во время подготовки данных для алгоритмов.

12. void Mv(std::string, std::string) - Выполняет команду mv для переименования временного файла.

13. void Rm(std::string) - Выполняет команду rm для удаления файла получившегося в процессе компрессии/декомпрессии и содержащего битые данные из-за ошибки алгоритма или удаляет поступивший алгоритму файл в случае отсутствия ключа k.

Исходный код

Тест производительности

Файл	Размер исходного файла	Алгоритм	Время сжатия (с)	Время декомпрессии (с)	Размер сжатого файла	Коэффициент сжатия
world95.txt		LZ77				
enwik8		LZ77				
enwik9		LZ77				
world95.txt		Арифметика				
enwik8		Арифметика				
enwik9		Арифметика				
world95.txt		gzip				
enwik8		gzip				
enwik9		gzip				

- Центральный процессор -
- Графический адаптер -
- Оперативная память -

Выводы

Благодаря выполнению данного проекта я научился основным принципам работы с файлами и директориями во время компрессии и декомпрессии. Так же я освоил основные правила и принципы работы компрессии и декомпрессии данных. Два построенных алгоритма дали мне необходимый базис навыков для работы с кастомными буферами. Так же я значительно улучшил свои навыки написания комплексных программ построения утилит, к примеру я научился реализовывать поддержку ключей и возможность работы нескольких алгоритмов.

Помимо этого стоит отметить, что статическое арифметическое кодирование будет проигрывать зачастую динамическому. Причина кроется в том, что в процессе чтения файла будут складываться ситуации, когда какой-то символ попадается достаточно часто, так что по сути его можно будет закодировать меньшим числом бит. Но если у нас статическое арифметическое кодирование, то возможна ситуация, когда под такой символ понадобится больше битов для кодирования, чем нужно. По сути динамический вариант может хорошо адаптироваться под какие-то временные особенности текста, что благоприятно влияет на качество, но взамен немного бьёт по времени работы.

Список литературы

1. Арифметическое кодирование - Arithmetic coding [Электронный ресурс]: ru.qwe.wiki URL: https://ru.qwe.wiki/wiki/Arithmetic_coding (дата обращения

28.08.2020)

2. Arithmetic Coding [Электронный ресурс]: users.cs.cf.ac.uk URL: https://users.cs.cf.ac.uk/Dave.Marshall/Multimedia/node213.html (дата обращения 16.09.2020)

3. LZ77 на C, реализация алгоритма LZ77 на C [Электронный ресурс]: algor.skyparadise.org URL: https://algor.skyparadise.org/read/14 (дата обращения 16.09.2020)