

МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)
Факультет информационных технологий и прикладной математики
Кафедра вычислительной математики и программирования

Отчет по лабораторной работе №6

«Сжатие»

по курсу

«Информационный поиск»

Группа: 80-106М

Выполнил: Демин И.А.

Преподаватель: Калинин А.Л.

Москва, 2019

Задание

В этом задании необходимо расширить язык запросов булева поиска новым элементом – поиском цитат.

VariableByte

Байтовое кодирование переменной длины (variable byte encoding VB, или variable byte coding — VBC) использует для кодирования интервалов целое количество байтов. Последние 7 бит в каждом байте являются "полезной нагрузкой" и кодируют часть интервала. Первый бит байта является битом продолжения (continuation bit). Он равен единице у последнего байта закодированного интервала и нулю в остальных случаях.

Иными словами, сжатие будет получены на кодирование «близких» больших чисел, когда их разница меньше самих чисел. В других случаях эффективности сжатия не будет.

Применение

Прежде чем приступить к написанию кода нужно понять, где можно получить сжатие. После анализа я нашел 3 места для оптимизации по памяти:

1. Использовать для записи чисел вместо Long Long(8 байт) unsigned int(4 байта)
2. Применить VB к последовательности doc_id для каждого слова
3. Применить VB к последовательности pos_in_files для каждого документа

Но если первый пункт прост в реализации, то второй и третий требуют перестроения в бинарных файлах.

Сейчас в файле bin_file записаны элементы (doc_id, offset, k), что не будет эффективно сжато при помощи VB. Файл offset_blocks подходит для эффективного сжатия, для корректной работы от него придется отказаться и перестроить файл bin_file.

Таким образом bin_file будет содержать следующие элементы для каждого ключа в обратном индексе:

vbcode_doc_ids, write_freq, for_write_pos_in_file

vbcode_doc_ids — doc_id сжатые VB

write_freq — частоты встречи слова в каждом файле

for_write_pos_in_file — позиции в файлах сжатые VB

Работа поиска не изменилась. Вот пример поискового запроса «мастер спорта» из предыдущей ЛР:

```
(base) ivan@ivan-G5:~/study/search/search_MAI/l8_compression$ python l8.py
Id: 0.Заголовок: Каратаев,_Николай_Дмитриевич. Url: https://ru.wikipedia.org/wiki/Каратаев,\_Николай\_Дмитриевич
Id: 1027.Заголовок: Сидоренко,_Александр_Игнатьевич. Url: https://ru.wikipedia.org/wiki/Сидоренко,\_Александр\_Игнатьевич
Id: 2051.Заголовок: Бобаренко,_Николай_Семёнович. Url: https://ru.wikipedia.org/wiki/Бобаренко,\_Николай\_Семёнович
Id: 2055.Заголовок: Рожков,_Геннадий_Федосеевич. Url: https://ru.wikipedia.org/wiki/Рожков,\_Геннадий\_Федосеевич
Id: 10.Заголовок: Норманов,_Азамат_Турдиевич. Url: https://ru.wikipedia.org/wiki/Норманов,\_Азамат\_Турдиевич
Id: 19.Заголовок: 1981_год_в_спорте. Url: https://ru.wikipedia.org/wiki/1981\_год\_в\_спорте
Id: 20.Заголовок: Цыплаков,_Александр_Викторович. Url: https://ru.wikipedia.org/wiki/Цыплаков,\_Александр\_Викторович
Id: 1043.Заголовок: Аухадов,_Апти_Хамзатович. Url: https://ru.wikipedia.org/wiki/Аухадов,\_Апти\_Хамзатович
Id: 1051.Заголовок: Комплексное_единоборство. Url: https://ru.wikipedia.org/wiki/Комплексное\_единоборство
Id: 2078.Заголовок: Армейский_рукопашный_бой. Url: https://ru.wikipedia.org/wiki/Армейский\_рукопашный\_бой
Id: 34.Заголовок: Романов,_Михаил_Иванович. Url: https://ru.wikipedia.org/wiki/Романов,\_Михаил\_Иванович
Id: 2083.Заголовок: Каспаров,_Гарри_Кимович. Url: https://ru.wikipedia.org/wiki/Каспаров,\_Гарри\_Кимович
Id: 1066.Заголовок: Соколова,_Октябрина_Александровна. Url: https://ru.wikipedia.org/wiki/Соколова,\_Октябрина\_Александровна
Id: 2093.Заголовок: Вараев,_Шарип_Магомедович. Url: https://ru.wikipedia.org/wiki/Вараев,\_Шарип\_Магомедович
```

Оценка результатов

Сжатие было получено и весьма значительное. Сжатию подверглись файлы bin_file и offset_files, которые теперь содержатся в файле bin_file.

Размер bin_file и offset_files из предыдущих ЛР — 110мб.

Размер bin_file, полученного в ходе выполнения ЛР — 10мб.

Вывод

В ходе лабораторной работы я изучил алгоритм VB, изменена структура бинарных файлов на более оптимальную и выполнено сжатие. По результатам видно, что сжатие весьма значительно, но, я думаю, здесь так же сыграло роль то, что до этого для каждого числа отводилось 8 байт, которые были выделены неоправдано. В итоге значительное пространство отводилось под информацию, которая не несла в себе полезной информации.