### Московский авиационный институт

**Факультет прикладной математики и физики**

**Лабораторная работа №6**

**по курсу:**

**«Информационный поиск»**

**по теме:**

**«Сжатие координатного индекса»**

**2 семестр**

Студент: Ахмед С. Х.

Преподаватель: Калинин А. Л.

Группа: 8О-106М

**Москва, 2019 г**

Постановка задачи

В этом задании необходимо применить алгоритмы сжатия к координатным блокам. Исследовать изменения в размерах частей индекса, влияние на скорость индексации и поиска.

В отчёте нужно указать:

• Выбранный метод сжатия. Привести побитовую схему хранения данных в индексе. Описать причины, по которым был выбран именно этот метод сжатия.

• Влияние сжатия на размер и скорость прохождения по координатным блокам всех терминов, редких терминов, терминов средней частотности и высокочастотных терминов.

• Обосновать, почему поиск после внедрения сжатия работает корректно. Как производилось тестирование?

Ход решения

В качестве исходного алгоритма был выбран Variable Bytes из-за простоты реализации его на питон, а также минимального числа изменений при работе с сжатым координатным индексом:

Битовое описание:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  | … |  |
| M байтов – (Сжатое) количество документов (Len(doc\_ids)) | | | | K байт – doc\_id(сжатое) | | | | P байтов (сжатое) – количество координат (N) | | | | N байт – координаты (сжаты) | | |
|  | | | | Повторение блока Len(doc\_ids) раз | | | | | | | | | | |

Размер координатного индекса до сжатия: 343 957 Кбайт

Размер координатного индекса после сжатия: 223 445 Кбайт

Размер координатного блока до сжатия: 197 056 Кбайт

Размер координатного блока после сжатия: 113 559 Кбайт

Процент сжатия координатного индекса: 35.04%

Процент сжатия координатного блока: 42.40%

В первоначальных замерах я сравнивал свое сжатие с сжатием библиотекой pickle, от того, величина сжатия была очень мала. После обнаружения данного казуса я пересчитал с учетом корректной размерности несжатого индекса. Следует отметить, данный алгоритм сжатия имеет недостаток, который отражается в числах: на малочастотных словах он способен увеличить размер индекса, в то самое время он может срезать размер частотных терминов. Так же к минусам алгоритма можно отнести его работу с крупными значения. По отношению к работе координатных блоков без индексов, то тут можно заметить существенное ускорение алгоритмов, так как теперь индекс (его часть) умещается в память, а расстояние между объектами в пространстве памяти уменьшилось. Корректность работы проверялась сравнением результата поиска по сжатому индексу с несжатым индексом, в результате чего и убедился в корректности работы алгоритма(результаты совпали)