### Московский авиационный институт

**Факультет прикладной математики и физики**

**Лабораторная работа №4**

**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:

Размер координатного индекса до сжатия:260 957 Кбайт

Размер координатного индекса после сжатия: 223 445 Кбайт

Размер координатного блока до сжатия: 117 056 Кбайт

Размер координатного блока после сжатия: 113 559 Кбайт

Процент сжатия координатного индекса: 14.374900311523891%

Процент сжатия координатного блока: 2.987577424394823%

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 существенного увеличения нет, так как всегда можно найти непустое пересечение, а затем просто раскодировать это. Проблемы со скоростью возникают при работе с цитатным поиском и там время проседает из-за необходимости распаковывать поток данных. Как было сказано ранее, на операцию пересечения сжатие не оказывает влияния(алгоритм не блочный) так как сами значения остаются неизменны, всегда есть обратное однозначное преобразования сжатого представления в естественное, что является достаточным условием рабоыт самого алгоритма, а также обоснованием работы поиска