Московский авиационный институт
Факультет прикладной математики и физики

Лабораторная работа №5

по курсу:
«Информационный поиск»
по теме:
«Цитатный поиск»
2 семестр

Студент: Ахмед С. Х.
Преподаватель: Калинин А. Л.
Группа: 8О-106М

Москва, 2019 г

Постановка задачи

В этом задании необходимо расширить язык запросов булева поиска новым элементом – поиском цитат. Синтаксис этого элемента следующий:

• [«что где когда»] – кавычки, включают режим цитатного поиска для терминов внутри кавычек. Этому запросу удовлетворяют документы, содержащие в себе все термины что, где и когда, причём они должны встретиться внутри документа ровно в этой последовательности, без каких либо вкраплений других терминов.

• [«что где когда» / 5] – аналогично предыдущему пункту, но допускаются вкрапления других терминов так, чтобы расстояние от первого термина цитаты до последнего не превышало бы 5.

Новый элемент может комбинироваться с другими стандартными средствами булева поиска, например:

• [«что где когда» && друзь]

• [«что где когда» || квн]

• [«что где когда» && !«хрустальная сова»]

Для реализации цитатного поиска нужно использовать координатный индекс, т.е. для каждого вхождения термина в документ построить и сохранить список позиций внутри документа, где этот термин встречался.

В отчёте нужно описать формат координатного индекса.

Привести статистические данные:

• Размер получившегося индекса.

• Время построения индекса.

• Общее количество позиций. Среднее количество позиций на термин и на пару термин документ.

• Скорость индексации (кб входных данных в секунду)

• Время выполнения поисковых запросов.

• Примеры долго выполняющихся запросов. Кроме того, нужно привести примеры запросов и результаты их выполнения. В выводах должны быть указаны недостатки работы, приведены примеры их решения. Что можно сделать, чтобы ускорить «долгие» запросы?

Ход решения

Полученный размер индекса:

1) представление деревом: ключ - хэш строки, значение - координатный блок – 335 МБ
2) Словарь и координатный индекс в разных файлах: Размер координатного индекса 254 МБ

Битовое представление – изменилось только представление координатного индекса

Время построения индекса: 3 минуты.

На килобайт данных: `0.015621185302734375 секунд`

`Общее количество позиций: 37007340`

`Среднее кол-во позиций на термин: 29.559097683825115 (30 позиций)`

`Проверка корректности поиска заключалась в проверке вхождения слов и цитат запроса в документ.`

`К реализованному в 4 лабораторной работе пришлось навешать доп обработку для цитат, а именно предварительное формирование словаря значений для цитат – то есть, координатные блоки для цитаты. Примеры запросов:`

- `<<из за>>/5 && !<<на карте>> латвия- выполнялся 3.54 s`
- `<<обзор пляжа>> <<юрмала март>>/6 (литва !<<эстонии названия>>) 33.9 ms`
- `<<обзор пляжа>> <<юрмала март>>/6 11.9 ms`
- `<<петропа вловская кре пость>>/3 <<санкт петербурге на заячьем острове>>/7 12.4 ms`

`Сложно найти очень долго выполняющиеся запросы для этой реализации, к таким запросам можно отнести первый запрос.`

`Одной из проблем моей реализации является наивное сопоставление цитат, избыточная предобработка входного потока и избыточный размер словаря. Первая проблема решается хранением n-грамм, она также решает проблему скорости операции пересечения и размера операции объединения. Вторая проблема решается модернизацией дерева выражений, а третья сжатием словаря.`

`Оценка качества поиска: не изменилась, так как не менялся сам алгоритм булева поиска, а лишь появился алгоритм поиска цитат с помощью координатного блока.`