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

Отчет по лабораторной работе №5
«Поиск цитат, координатный индекс»
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.

Выполнение лабораторной работы

В этой лабораторной работе за основу была взята предыдущая лабораторная работа.

Для булева поиска нужна была структура

{

hash(word): [doc_id1, doc_id2, …]

...

}

Теперь же нужно хранить еще и позиции в документе. «Сырой» обратный индекс будет таким:

{

hash(word):

{

doc_id: [pos1, pos2, …]

...

},

...

}

Но в итоге, наш обратный индекс должен иметь структуру, такую же, как и в предыдущей ЛР:

{

hash(word): (offset, size_of_block)

}

Достигается это следующим образом — создается дополнительный файл, в котором будут храниться позиции в документах. Документ, который хранит последовательность doc_id, изменит свою структуру: если раньше он хранил

последовательность doc_id, то теперь один элемент — это (doc_id, смещение_в_файле_с_позициями, сколько_нужно_считать).

Процесс построения поисковой выдачи: так как в цитате все объекты соединяются операцией &, то здесь не будет дерева выражений. Алгоритм будет следующим:

1. Получить слова из цитаты и шаг

2. Получить словарь для слова, в котором ключи — doc_id, значения — позиции

3. Взять слово i и результат, полученный до шага i (cur_res).

3.1 Идти по ключам cur_res и искать их в словаре для i(поиск за O(1), так как используется хеш-таблица). Если нашли пересечение, то идем по позициям и ищем пересечение позиций в диапазоне допустимого шага.

4. Возвращаем результат для цитаты.

Проверка

Проверку осуществлять уже проще, так как поиск того-же PyCharm ищет цитаты, что эквивалентно созданному поиску с шагом 1.

Запросы для проверки цитатного поиска:

«мастер спорта»/1 — 262 статьи было получены за 0.016 секунды, с выводом 0.26 секунды.

```
(base) ivan@ivan-G5:~/study/search/search_MAI/l7_coordinate$ python l7.py
Поиск выполнен за 0:00:00.016003
Id: 1050.Заголовок: 1981_год_в_спорте. Url: https://ru.wikipedia.org/wiki/1981_год_в_спорте
Id: 27.Заголовок: 2012_год_в_спорте. Url: https://ru.wikipedia.org/wiki/2012_год_в_спорте
Id: 2076.Заголовок: Лапта. Url: https://ru.wikipedia.org/wiki/Лапта
Id: 29.Заголовок: Белоглазов,_Сергей_Алексеевич. Url: https://ru.wikipedia.org/wiki/Белоглазов,_Сергей_Алексеевич
Id: 30.Заголовок: Ильенко,_Наталья_Никитична. Url: https://ru.wikipedia.org/wiki/Ильенко,_Наталья_Никитична
Id: 1062.Заголовок: 1975_год_в_спорте. Url: https://ru.wikipedia.org/wiki/1975_год_в_спорте
Id: 39.Заголовок: Челышев,_Анатолий_Васильевич. Url: https://ru.wikipedia.org/wiki/Челышев,_Анатолий_Васильевич
Id: 2088.Заголовок: Фадеев,_Александр_Николаевич. Url: https://ru.wikipedia.org/wiki/Фадеев,_Александр_Николаевич
Id: 2092.Заголовок: Хоменков,_Леонид_Сергеевич. Url: https://ru.wikipedia.org/wiki/Хоменков,_Леонид_Сергеевич
Id: 1070.Заголовок: Кузьмин,_Кирилл_Константинович. Url: https://ru.wikipedia.org/wiki/Кузьмин,_Кирилл_Константинович
Id: 1073.Заголовок: Билялетдинов,_Динияр_Ринатович. Url: https://ru.wikipedia.org/wiki/Билялетдинов,_Динияр_Ринатович
Id: 2098.Заголовок: Вараев,_Башир_Магомедович. Url: https://ru.wikipedia.org/wiki/Вараев,_Башир_Магомедович
Id: 1077.Заголовок: Аухадов,_Хамзат_Ахмаевич. Url: https://ru.wikipedia.org/wiki/Аухадов,_Хамзат_Ахмаевич
Id: 2103.Заголовок: Михайлин,_Вячеслав_Вячеславович. Url: https://ru.wikipedia.org/wiki/Михайлин,_Вячеслав_Вячеславович
Id: 2108.Заголовок: Соколова,_Октябрина_Александровна. Url: https://ru.wikipedia.org/wiki/Соколова,_Октябрина_Александровна
```

«мастер по самбо»/2 — 3 статьи за 0.12 секунды, с выводом также за 0.12

```
(base) ivan@ivan-G5:~/study/search/search_MAI/l7_coordinate$ python l7.py
Поиск выполнен за 0:00:00.116252
Id: 621.Заголовок: Емельяненко,_Фёдор_Владимирович. Url: https://ru.wikipedia.org/wiki/Емельяненко,_Фёдор_Владимирович
Id: 917.Заголовок: Козлов,_Геннадий_Андреевич. Url: https://ru.wikipedia.org/wiki/Козлов,_Геннадий_Андреевич
Id: 1022.Заголовок: Калмыцкая_борьба. Url: https://ru.wikipedia.org/wiki/Калмыцкая_борьба
Articles count: 3
get_search_res_for_quotes works 0:00:00.118048
```

«мастер спорта федерации»/3 — немного искусственный запрос, что бы просто продемонстрировать работу поиска через шаг.

```
(base) ivan@ivan-G5:~/study/search/search_MAI/l7_coordinate$ python l7.py
Поиск выполнен за 0:00:00.116252
Id: 621.Заголовок: Емельяненко,_Фёдор_Владимирович. Url: https://ru.wikipedia.org/wiki/Емельяненко,_Фёдор_Владимирович
Id: 917.Заголовок: Козлов,_Геннадий_Андреевич. Url: https://ru.wikipedia.org/wiki/Козлов,_Геннадий_Андреевич
Id: 1022.Заголовок: Калмыцкая_борьба. Url: https://ru.wikipedia.org/wiki/Калмыцкая_борьба
Articles count: 3
get_search_res_for_quotes works 0:00:00.118048
```

Оценка точности

Точность считается как

$$P = \frac{a}{a + b}.$$

a — количество релевантных документов

b — число нерелевантных документов.

Возьмем 5 верхних статей из выдачи для 5ти запросов из ЛР2. Все запросы ищем цитатным поиском с шагом 1.

"спорт экспресс" — 0.6

"виды спорта" — 1

"активный отдых" — 0.8

"профессиональный бокс" — 1

"боевые искусства" — 1

Средняя точность — 0.88

Как видно, точно выросла значительна из-за того, что все запросы являются более естественными для цитатного поиска, так как в этих запросах слова по отдельности искать смысла почти нет.

Суммарный объем файлов, полученный для тестового множества из 3 тысяч статей, составляет около 110МБ.

Добавление булево поиска

Так как мы изменили структуру бинарных файлов, то получение данных для булева поиска изменилось. Из файла, который теперь содержит вместо

последовательности doc_id элементы (doc_id, offset, k) считываем все элементы и берем первое значение — получили список doc_id. Для того, что бы разделять запросы, строим регулярное выражение, ищем отдельно цитаты и слова для булева поиска. Процесс получения итого пула doc_id не изменился.

Вывод

В ходе лабораторной работы был реализован смешанный поиск, содержащий в себе как цитатный поиск, так и булев. Как видно из оценки точности, то качество поиска выросло.