**Московский авиационный институт  
(национальный исследовательский университет)**

Факультет прикладной математики и физики  
Кафедра вычислительной математики и программирования

**Лабораторная работа №9  
по курсу «Информационный поиск»**

Зонный поиск

|  |  |  |
| --- | --- | --- |
|  | **Студент:** | Рожлейс И.А. |
| **Группа:** | М8О-108М-18 |
| **Преподаватель:** | Калинин А.Л. |
| **Дата:** | 11.09.2019 |

Москва, 2019

**Цель работы**

Необходимо добавить в поисковый индекс информацию о зонах, в которых встретились термины. Как минимум, нужно сделать отдельные зоны для заголовков документов. Так же, необходимо учесть эти зоны в ранжировании, причём таким образом, чтобы поиск стал искать лучше.

В отчёте нужно привести:

* побитовое описание индекса с зонами;
* формулу ранжирования, подобранные веса;
* оценку качества поиска после внедрения зон.

Есть ли запросы, по которым качество ухудшилось? Почему? Что можно сделать, чтобы качество поиска по ним улучшилось, а по остальным запросам – не ухудшилось бы?

**Ход работы**

Для выполнения данной работы была произведена модификация лабораторной работы №3 «Булев индекс». В программу добавлена опция построения зонного индекса. Формат файла зонного индекса заимствован от формата координатного индекса. На **рис. 1** представлена структурная схема формата файла зонного индекса.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Начало файла** | | | | | | | | | | | | | | |  | | | **Конец файла** | | | | | | | |
| **Заголовок** | | | | **Словарь** | | | | | **Данные зон** | | | | | | | | | | | | | | | | |
| 0xBCBCBCBC | Кол-во токенов | Размер словаря (байт) |  | ИД токена 1 | Смещение токена 1 | … | ИД токена k | Смещение токена k | **Документ 1** | | | | | | | | … | | **Документ n** | | | | | | |
| **Заголовок 1** | | | | Зона 1 | … | | Зона i | **Заголовок n** | | | | Зона 1 | … | Зона j |
| 0x9C9C9C9C | ИД токена 1 | ИД документа 1 | Кол-во зон | 0x9C9C9C9C | ИД токена k | ИД документа n | Кол-во зон |

**Рис. 1.** Структура файла

Координатный индекс может быть довольно легко преобразован в зонный индекс. Вместо словопозиций, зонный индекс хранит информацию о зонах в которых располагается термин. Каждый документ имеет три зоны: заголовок, первая половина документа и вторая половина документа. Алгоритм преобразования координатного в зонный индекс заключается в следующем: зная количество токенов документа и словопозицию токена программа определяет в какой половине документа находится токен. Тем самым словопозиции токенов вырождаются в зоны.

В текущем варианте программы были выбраны веса зон в следующем соотношении: заголовок документа 0.5, первая половина документа 0.4 и вторая половина документа 0.1.

В качестве итоговой формулы ранжирования был выбран вариант вычисления среднего арифметического значения по результатам двух ранжирующих функций: TF-IDF и зонного поиска. То есть сперва независимо вычисляется ранг каждого документа в соответствие с метрикой TF-IDF и зонным поиском, после чего вычисляется их среднее арифметическое:

Здесь представляет порядковый номер документа в неранжированном списке, а и значение ранга соответствующего документа по алгоритмам ранжирования TFIDF и зонного поиска соответственно.

В **табл. 1** представлен список из 30 поисковых запросов, и соответствующая оценка релевантности документов поисковой выдачи. Оценка релевантности находится в пределах 0…3, где ноль соответствует наименьшей релевантности документа.

|  |  |  |
| --- | --- | --- |
|  | **Запрос** | **Оценка** |
| **1** | физиология клетки | 1 2 1 1 2 |
| **2** | секвенирование генома | 3 3 3 3 3 |
| **3** | опасные заболевания | 3 1 2 2 0 |
| **4** | дивергентный характер эволюции | 3 3 3 2 1 |
| **5** | биологический взрыв | 2 0 1 2 2 |
| **6** | экологические проблемы | 2 2 2 0 0 |
| **7** | современные проблемы биологии | 1 0 0 2 2 |
| **8** | биологическое разнообразие | 2 2 0 3 0 |
| **9** | генетически предрасположенные заболевания | 2 3 3 3 0 |
| **10** | последствия бешенства | 2 2 3 0 3 |
| **11** | цитоплазматическая мембрана | 1 1 1 1 1 |
| **12** | огромные динозавры | 3 3 3 3 3 |
| **13** | звёздные путешествия | 2 0 3 1 0 |
| **14** | быстрорастущие деревья | 3 3 3 3 3 |
| **15** | генетические алгоритмы | 3 1 2 0 0 |
| **16** | бубонная чума | 2 3 2 2 0 |
| **17** | анаэробные бактерии | 3 0 1 0 1 |
| **18** | зарождение жизни | 0 3 3 3 2 |
| **19** | открытия в микробиологии | 1 0 2 2 1 |
| **20** | самые значимые открытия | 1 0 0 0 0 |
| **21** | возможности микроскопа | 3 2 3 1 3 |
| **22** | представители рода пасленовых | 0 0 0 0 0 |
| **23** | естественный отбор | 3 2 2 3 2 |
| **24** | орудие труда обезьян | 0 0 0 0 3 |
| **25** | приспосабливаемость приматов | 0 0 1 0 2 |
| **26** | генетическое разнообразие | 3 3 3 1 2 |
| **27** | вред курения | 1 0 0 0 0 |
| **28** | нейрофизиологические процессы | 1 0 0 1 2 |
| **29** | будущее биологии | 0 0 0 2 3 |
| **30** | устройство клетки | 0 0 0 0 3 |

**Табл. 1.** Оценка релевантности поисковой выдачи

Некоторые запросы были сформулированы таким образом, чтобы на них можно было дать лишь утвердительный ответ: «да» либо «нет». Пример одного из таких запросов – «огромные динозавры». Запрос не предполагает наличия промежуточных состояний. Некоторые запросы в свою очередь были специально сформулированы неясно либо в некоторой степени даже противоречиво. Например, ответом на запрос «биологический взрыв» предполагалось получить документы на тему увеличения количества биологических видов, в связи с чем некоторые документы оказались нерелевантными данному запросу. Другие запросы формировались по принципу нахождения любой информации максимально близкой к теме запроса.

В **табл. 2** представлены вычисленные метрики отдельно для каждого запроса на трёх уровнял @1, @3 и @5.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **P** | | | **DCG** | | | **NDCG** | | | **ERR** | | |
|  | @1 | @3 | @5 | @1 | @3 | @5 | @1 | @3 | @5 | @1 | @3 | @5 |
| **1** | 0,00 | 0,20 | 0,40 | 1,00 | 2,76 | 3,97 | 1,00 | 0,88 | 0,87 | 0,13 | 0,31 | 0,36 |
| **2** | 0,20 | 0,60 | 1,00 | 3,00 | 6,39 | 8,85 | 1,00 | 1,00 | 1,00 | 0,88 | 0,93 | 0,93 |
| **3** | 0,20 | 0,40 | 0,60 | 3,00 | 4,63 | 5,49 | 1,00 | 0,97 | 0,96 | 0,88 | 0,90 | 0,90 |
| **4** | 0,20 | 0,60 | 0,80 | 3,00 | 6,39 | 7,64 | 1,00 | 1,00 | 1,00 | 0,88 | 0,93 | 0,93 |
| **5** | 0,20 | 0,20 | 0,60 | 2,00 | 2,50 | 4,14 | 1,00 | 0,95 | 0,88 | 0,38 | 0,40 | 0,48 |
| **6** | 0,20 | 0,60 | 0,60 | 2,00 | 4,26 | 4,26 | 1,00 | 1,00 | 1,00 | 0,38 | 0,54 | 0,54 |
| **7** | 0,00 | 0,00 | 0,40 | 1,00 | 1,00 | 2,64 | 1,00 | 1,00 | 0,70 | 0,13 | 0,13 | 0,25 |
| **8** | 0,20 | 0,40 | 0,60 | 2,00 | 3,26 | 4,55 | 1,00 | 1,00 | 0,87 | 0,38 | 0,49 | 0,58 |
| **9** | 0,20 | 0,60 | 0,80 | 2,00 | 5,39 | 6,68 | 1,00 | 0,92 | 0,92 | 0,38 | 0,67 | 0,67 |
| **10** | 0,20 | 0,60 | 0,80 | 2,00 | 4,76 | 5,92 | 1,00 | 0,90 | 0,88 | 0,38 | 0,61 | 0,61 |
| **11** | 0,00 | 0,00 | 0,00 | 1,00 | 2,13 | 2,95 | 1,00 | 1,00 | 1,00 | 0,13 | 0,21 | 0,25 |
| **12** | 0,20 | 0,60 | 1,00 | 3,00 | 6,39 | 8,85 | 1,00 | 1,00 | 1,00 | 0,88 | 0,93 | 0,93 |
| **13** | 0,20 | 0,40 | 0,40 | 2,00 | 3,50 | 3,93 | 1,00 | 0,82 | 0,83 | 0,38 | 0,56 | 0,56 |
| **14** | 0,20 | 0,60 | 1,00 | 3,00 | 6,39 | 8,85 | 1,00 | 1,00 | 1,00 | 0,88 | 0,93 | 0,93 |
| **15** | 0,20 | 0,40 | 0,40 | 3,00 | 4,63 | 4,63 | 1,00 | 0,97 | 0,97 | 0,88 | 0,90 | 0,90 |
| **16** | 0,20 | 0,60 | 0,80 | 2,00 | 4,89 | 5,75 | 1,00 | 0,93 | 0,94 | 0,38 | 0,66 | 0,66 |
| **17** | 0,20 | 0,20 | 0,20 | 3,00 | 3,50 | 3,89 | 1,00 | 0,96 | 0,94 | 0,88 | 0,88 | 0,88 |
| **18** | 0,00 | 0,40 | 0,80 | 0,00 | 3,39 | 5,46 | 0,00 | 0,69 | 0,75 | 0,00 | 0,47 | 0,48 |
| **19** | 0,00 | 0,20 | 0,40 | 1,00 | 2,00 | 3,25 | 1,00 | 0,76 | 0,77 | 0,13 | 0,23 | 0,29 |
| **20** | 0,00 | 0,00 | 0,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 0,13 | 0,13 | 0,13 |
| **21** | 0,20 | 0,60 | 0,80 | 3,00 | 5,76 | 7,35 | 1,00 | 0,98 | 0,96 | 0,88 | 0,92 | 0,92 |
| **22** | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| **23** | 0,20 | 0,60 | 1,00 | 3,00 | 5,26 | 7,33 | 1,00 | 1,00 | 0,97 | 0,88 | 0,91 | 0,92 |
| **24** | 0,00 | 0,00 | 0,20 | 0,00 | 0,00 | 1,16 | 0,00 | 0,00 | 0,39 | 0,00 | 0,00 | 0,18 |
| **25** | 0,00 | 0,00 | 0,20 | 0,00 | 0,50 | 1,27 | 0,00 | 0,50 | 0,48 | 0,00 | 0,04 | 0,11 |
| **26** | 0,20 | 0,60 | 0,80 | 3,00 | 6,39 | 7,60 | 1,00 | 1,00 | 0,99 | 0,88 | 0,93 | 0,93 |
| **27** | 0,00 | 0,00 | 0,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 0,13 | 0,13 | 0,13 |
| **28** | 0,00 | 0,00 | 0,20 | 1,00 | 1,00 | 2,20 | 1,00 | 1,00 | 0,70 | 0,13 | 0,13 | 0,21 |
| **29** | 0,00 | 0,00 | 0,40 | 0,00 | 0,00 | 2,02 | 0,00 | 0,00 | 0,47 | 0,00 | 0,00 | 0,20 |
| **30** | 0,00 | 0,00 | 0,20 | 0,00 | 0,00 | 1,16 | 0,00 | 0,00 | 0,39 | 0,00 | 0,00 | 0,18 |

**Табл. 2.** Метрики качества поиска

На **табл. 3** представлены усреднённые значения вычисленных ранее метрик.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **P** | | | **DCG** | | | **NDCG** | | | **ERR** | | |
| @1 | @3 | @5 | @1 | @3 | @5 | @1 | @3 | @5 | @1 | @3 | @5 |
| 0,11 | 0,31 | 0,51 | 1,70 | 3,30 | 4,46 | 0,80 | 0,81 | 0,82 | 0,41 | 0,50 | 0,54 |

**Табл. 3.** Средние значения метрик

**Заключение**

Размер итогового зонного индекса оказался равен 407 Мб. Сжатие данных для хранения зонного индекса не использовалось. В качестве возможных зон были выбраны три типа зон: заголовок документа, первая и вторая половина документа. В качестве весов для этих зон были подобраны значения 0.5, 0.4 и 0.1 соответственно.

Метрики собственного поиска стали гораздо лучше по сравнению с результатами оценки качества собственного поиска без использования зонного индексирования. Положительный прирост качества поиска наблюдается по всем метрикам относительно лабораторной работы №3 «Лемматизация» и составляет: 175% для P, 162% для DCG, 132% для NDGC и 150% для ERR.

Ухудшения качества поисковой выдачи по отдельным запросам не было обнаружено, однако некоторые запросы были обработаны крайне неудовлетворительно. Примерами могут служить запросы «орудие труда обезьян» или «представители рода пасленовых» в первых позициях списка которых не было практически ни одного релевантного документа. Для улучшения качества таких запросов возможно стоит учитывать частоту совместного использования слов из поискового запроса.

**Источники**

[1] Маннинг, К. Введение в информационный поиск // Вильямс, 2014. – 528 с.