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

Факультет прикладной математики и физики  
Кафедра вычислительной математики и программирования

**Лабораторная работа №8  
по курсу «Информационный поиск»**

Ранжирование TF-IDF

|  |  |  |
| --- | --- | --- |
|  | **Студент:** | Рожлейс И.А. |
| **Группа:** | М8О-108М-18 |
| **Преподаватель:** | Калинин А.Л. |
| **Дата:** | 09.06.2019 |

Москва, 2019

**Цель работы**

Необходимо сделать ранжированный поиск на основании схемы ранжирования TF-IDF. Теперь, если запрос содержит в себе только термины через пробелы, то его надо трактовать как нечёткий запрос, т.е. допускать неполное соответствие документа терминам запроса и т.п. Примеры запросов:

• [ роза цветок ]

• [ московский авиационный институт ]

Если запрос содержит в себе операторы булева поиска, то запрос надо трактовать как булев, т.е. соответствие должно быть строгим, но порядок выдачи должен быть определён ранжированием TF-IDF. Например:

• [ роза && цветок ]

• [ московский && авиационный && институт ]

В отчёте нужно привести несколько примеров выполнения запросов, как удачных, так и не удачных.

**Ход работы**

Для выполнения этой работы потребовалась модернизация лабораторной работы №3 «Булев индекс». Теперь помимо прямого и координатного индексов программа собирает статистику корпуса документов и сохраняет её в файл. Далее этот файл используется для ранжирования по формуле TF-IDF.

Статистический файл, как и индексные файлы поделён на три секции: заголовок файла, словарь и секция данных. Структурная схема файла представлена на **табл. 1**. Заголовок файла имеет такие поля данных как количество уникальных токенов (порядка 1,5 млн.) и количество документов в корпусе (порядка 65 тыс.) Далее следует словарь файла.

Словарь состоит из записей. Одна запись описывает один токен. В записи указан идентификатор токена, количество документов корпуса в которых встречается токен и смещение в байтах относительно начала файла в секцию данных где расположен список документов, содержащих этот токен. Идентификаторы токенов представляют собой контрольную сумму байт UTF-8 строки токена по алгоритму CRC-32.

Секция данных представляет собой список документов и информацию, связывающую токен с документом: количество вхождений токена в документ и количество всего токенов в документе. Информация с секции данных избыточна. Так, например, количество всего токенов в документе не зависит от количества вхождений конкретного токена, тем не менее эта информация дублируется для каждого документа. Файл возможно хранить более компактно, однако в таком случае возрастает сложность работы с файлом.



**Табл. 1.** Структурная схема статистического файла.

Вес документа высчитывается как сумма меры TF-IDF токенов входящих в состав документа, которая представляет собой произведение множителей TF и IDF:

Мера TF (term frequency) определяет частоту вхождения токена в документ и представляет собой отношение количества вхождений токена к общему количеству токенов в документе:

где n количество вхождений токена, а N количество всего токенов в документе. Чем чаще употребляется искомый токен в документе, тем больший вес имеет этот документ.

Мера IDF (inverse document frequency) в свою очередь определяет частоту использования токена среди документов всего корпуса и по своей сути перераспределяет значимость документов в сторону редкоупотребляемых токенов. Так чем в меньшем количестве документов упоминается искомый токен, тем больший вес имеют эти документы:

где N общее количество документов в корпусе, а d количество документов в состав которых входит токен.

Ранжирование списка документов происходит в функции range посредством сортировки документов по их весу.



Сначала функция формирует вектор пар *<документ, токен>* для которых считывает информацию из файла статистики. Для каждой пары высчитывается мера TF-IDF, суммируется и присваивается нужному документу. Далее список пар *<документ, вес>* сортируется функцией стандартной библиотеки std::sort по ключу *«вес»*. Способ сортировки описан предикатом pred который передаётся третьим аргументом.

Была произведена оценка качества поиска по 30 запросам (**табл. 2**) и произведена сравнительная оценка с поисковой машиной Google. Была выбрана четырёхзначная шкала релевантности документов, где 0 означает совершенно нерелевантный документ, а оценка 3 – сильно релевантный документ.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Запрос** | **Google** | **Собств.** |
| **1** | «фотосинтезирующие бактерии» | 3 1 1 1 3 | 0 0 1 0 0 |
| **2** | «хладнокровные рептилии» | 3 3 3 3 0 | 3 3 3 3 2 |
| **3** | «палочка Коха» | 3 2 3 0 1 | 0 1 1 0 0 |
| **4** | «аэробные бактерии» | 0 3 3 0 3 | 2 0 0 1 2 |
| **5** | «Дарвин» | 3 0 0 0 0 | 1 1 1 1 1 |
| **6** | «дезоксирибонуклеиновая кислота» | 3 1 2 3 0 | 1 0 0 0 0 |
| **7** | «бубонная чума» | 3 1 2 1 1 | 1 0 3 0 1 |
| **8** | «бывает ли так что рыбы летают» | 3 0 0 0 0 | 0 0 1 1 1 |
| **9** | «фотосинтез» | 3 2 2 3 3 | 2 1 2 2 2 |
| **10** | «рецессивный признак» | 3 0 3 2 0 | 0 0 3 1 0 |
| **11** | «кроссинговер хромосом» | 3 1 2 2 1 | 1 2 0 0 3 |
| **12** | «собака Павлова» | 1 3 0 1 0 | 0 0 0 1 0 |
| **13** | «страусиное яйцо польза» | 1 0 0 0 0 | 0 0 0 0 0 |
| **14** | «где живут пингвины» | 3 2 2 2 2 | 3 2 2 2 3 |
| **15** | «внеземная жизнь» | 2 1 2 1 1 | 2 0 1 1 0 |
| **16** | «как работает клетка» | 3 0 2 1 3 | 1 1 0 0 0 |
| **17** | «комплекс Гольджи» | 3 2 1 1 0 | 0 1 3 0 0 |
| **18** | «гипертрихоз» | 3 2 2 2 0 | 0 0 0 0 0 |
| **19** | «мутагенные вещества» | 3 1 2 2 0 | 0 0 0 0 0 |
| **20** | «кислородное голодание» | 3 3 3 0 3 | 2 1 1 1 0 |
| **21** | «кто первый вышел на сушу» | 3 2 0 2 3 | 0 0 0 0 0 |
| **22** | «знаменитые биологи» | 3 3 0 3 1 | 0 0 0 0 0 |
| **23** | «теория зарождения жизни» | 3 3 3 0 2 | 0 2 0 0 2 |
| **24** | «нейроны человека» | 2 3 2 1 1 | 2 2 2 0 1 |
| **25** | «самое редкое заболевание» | 3 3 3 3 0 | 1 0 3 0 0 |
| **26** | «как долго живут слоны» | 3 2 2 1 0 | 3 3 0 1 1 |
| **27** | «генная инженерия» | 3 2 2 2 1 | 1 2 2 2 1 |
| **28** | «зачем нужны аминокислоты» | 2 2 1 2 1 | 1 1 1 1 2 |
| **29** | «осмос в клетке» | 1 3 2 3 3 | 2 3 0 1 1 |
| **30** | «нейробиологическое программирование» | 3 1 0 1 0 | 0 2 0 1 0 |

**Табл. 2.** Релевантность документов.

На основании поисковой выдачи были рассчитаны метрики (**табл. 3**) P, DCG, NDCG и ERR на уровнях @1, @3 и @5 для собственной поисковой машины и поисковой машины Google.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Google** | | | | **Собств.** | | | |
| **P** | **DCG** | **NDCG** | **ERR** | **P** | **DCG** | **NDCG** | **ERR** |
| **1** | 0.2, 0.2, 0.4 | 3.00, 4.13, 5.72 | 1.00, 1.00, 0.92 | 0.88, 0.89, 0.91 | 0.00, 0.00, 0.00 | 0.00, 1.00, 1.00 | 0.00, 0.50, 0.50 | 0.00, 0.50, 0.50 |
| **2** | 0.2, 0.6, 0.8 | 3.00, 6.39, 7.68 | 1.00, 1.00, 1.00 | 0.88, 0.93, 0.93 | 0.20, 0.60, 1.00 | 3.00, 9.00, 14.00 | 3.00, 6.39, 8.46 | 1.00, 1.00, 1.00 |
| **3** | 0.2, 0.6, 0.6 | 3.00, 5.76, 6.15 | 1.00, 0.98, 0.97 | 0.88, 0.92, 0.92 | 0.00, 0.00, 0.00 | 0.00, 2.00, 2.00 | 0.00, 1.13, 1.13 | 0.00, 0.69, 0.69 |
| **4** | 0, 0.4, 0.6 | 0.00, 3.39, 4.55 | 0.00, 0.69, 0.71 | 0.00, 0.47, 0.48 | 0.20, 0.20, 0.40 | 2.00, 2.00, 5.00 | 2.00, 2.00, 3.20 | 1.00, 1.00, 0.85 |
| **5** | 0.2, 0.2, 0.2 | 3.00, 3.00, 3.00 | 1.00, 1.00, 1.00 | 0.88, 0.88, 0.88 | 0.00, 0.00, 0.00 | 1.00, 3.00, 5.00 | 1.00, 2.13, 2.95 | 1.00, 1.00, 1.00 |
| **6** | 0.2, 0.4, 0.6 | 3.00, 4.63, 5.92 | 1.00, 0.97, 0.94 | 0.88, 0.90, 0.91 | 0.00, 0.00, 0.00 | 1.00, 1.00, 1.00 | 1.00, 1.00, 1.00 | 1.00, 1.00, 1.00 |
| **7** | 0.2, 0.4, 0.4 | 3.00, 4.63, 5.45 | 1.00, 0.97, 0.98 | 0.88, 0.90, 0.90 | 0.00, 0.20, 0.20 | 1.00, 4.00, 5.00 | 1.00, 2.50, 2.89 | 1.00, 0.69, 0.70 |
| **8** | 0.2, 0.2, 0.2 | 3.00, 3.00, 3.00 | 1.00, 1.00, 1.00 | 0.88, 0.88, 0.88 | 0.00, 0.00, 0.00 | 0.00, 1.00, 3.00 | 0.00, 0.50, 1.32 | 0.00, 0.50, 0.62 |
| **9** | 0.2, 0.6, 1 | 3.00, 5.26, 7.71 | 1.00, 1.00, 0.96 | 0.88, 0.91, 0.92 | 0.20, 0.40, 0.80 | 2.00, 5.00, 9.00 | 2.00, 3.63, 5.27 | 1.00, 0.97, 0.96 |
| **10** | 0.2, 0.4, 0.6 | 3.00, 4.50, 5.36 | 1.00, 0.92, 0.91 | 0.88, 0.91, 0.91 | 0.00, 0.20, 0.20 | 0.00, 3.00, 4.00 | 0.00, 1.50, 1.93 | 0.00, 0.50, 0.53 |
| **11** | 0.2, 0.4, 0.6 | 3.00, 4.63, 5.88 | 1.00, 0.97, 0.97 | 0.88, 0.90, 0.90 | 0.00, 0.20, 0.40 | 1.00, 3.00, 6.00 | 1.00, 2.26, 3.42 | 1.00, 0.86, 0.72 |
| **12** | 0, 0.2, 0.2 | 1.00, 2.89, 3.32 | 1.00, 0.80, 0.80 | 0.13, 0.51, 0.51 | 0.00, 0.00, 0.00 | 0.00, 0.00, 1.00 | 0.00, 0.00, 0.43 | 0.00, 0.00, 0.43 |
| **13** | 0, 0, 0 | 1.00, 1.00, 1.00 | 1.00, 1.00, 1.00 | 0.13, 0.13, 0.13 | 0.00, 0.00, 0.00 | 0.00, 0.00, 0.00 | 0.00, 0.00, 0.00 | 0.00, 0.00, 0.00 |
| **14** | 0.2, 0.6, 1 | 3.00, 5.26, 6.90 | 1.00, 1.00, 1.00 | 0.88, 0.91, 0.92 | 0.20, 0.60, 1.00 | 3.00, 7.00, 12.00 | 3.00, 5.26, 7.28 | 1.00, 1.00, 0.97 |
| **15** | 0.2, 0.4, 0.4 | 2.00, 3.63, 4.45 | 1.00, 0.97, 0.97 | 0.38, 0.48, 0.50 | 0.20, 0.20, 0.20 | 2.00, 3.00, 4.00 | 2.00, 2.50, 2.93 | 1.00, 0.95, 0.94 |
| **16** | 0.2, 0.4, 0.6 | 3.00, 4.00, 5.59 | 1.00, 0.94, 0.88 | 0.88, 0.89, 0.91 | 0.00, 0.00, 0.00 | 1.00, 2.00, 2.00 | 1.00, 1.63, 1.63 | 1.00, 1.00, 1.00 |
| **17** | 0.2, 0.4, 0.4 | 3.00, 4.76, 5.19 | 1.00, 1.00, 1.00 | 0.88, 0.90, 0.90 | 0.00, 0.20, 0.20 | 0.00, 4.00, 4.00 | 0.00, 2.13, 2.13 | 0.00, 0.59, 0.59 |
| **18** | 0.2, 0.6, 0.8 | 3.00, 5.26, 6.12 | 1.00, 1.00, 1.00 | 0.88, 0.91, 0.91 | 0.00, 0.00, 0.00 | 0.00, 0.00, 0.00 | 0.00, 0.00, 0.00 | 0.00, 0.00, 0.00 |
| **19** | 0.2, 0.4, 0.6 | 3.00, 4.63, 5.49 | 1.00, 0.97, 0.96 | 0.88, 0.90, 0.90 | 0.00, 0.00, 0.00 | 0.00, 0.00, 0.00 | 0.00, 0.00, 0.00 | 0.00, 0.00, 0.00 |
| **20** | 0.2, 0.6, 0.8 | 3.00, 6.39, 7.55 | 1.00, 1.00, 0.98 | 0.88, 0.93, 0.93 | 0.20, 0.20, 0.20 | 2.00, 4.00, 5.00 | 2.00, 3.13, 3.56 | 1.00, 1.00, 1.00 |
| **21** | 0.2, 0.4, 0.8 | 3.00, 4.26, 6.28 | 1.00, 1.00, 0.93 | 0.88, 0.90, 0.91 | 0.00, 0.00, 0.00 | 0.00, 0.00, 0.00 | 0.00, 0.00, 0.00 | 0.00, 0.00, 0.00 |
| **22** | 0.2, 0.4, 0.6 | 3.00, 4.89, 6.57 | 1.00, 1.00, 0.96 | 0.88, 0.93, 0.93 | 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** | 0.2, 0.6, 0.8 | 3.00, 6.39, 7.17 | 1.00, 1.00, 0.99 | 0.88, 0.93, 0.93 | 0.00, 0.20, 0.40 | 0.00, 2.00, 4.00 | 0.00, 1.26, 2.04 | 0.00, 0.63, 0.62 |
| **24** | 0.2, 0.6, 0.6 | 2.00, 4.89, 5.71 | 1.00, 0.93, 0.94 | 0.38, 0.66, 0.66 | 0.20, 0.60, 0.60 | 2.00, 6.00, 7.00 | 2.00, 4.26, 4.65 | 1.00, 1.00, 0.99 |
| **25** | 0.2, 0.6, 0.8 | 3.00, 6.39, 7.68 | 1.00, 1.00, 1.00 | 0.88, 0.93, 0.93 | 0.00, 0.20, 0.20 | 1.00, 4.00, 4.00 | 1.00, 2.50, 2.50 | 1.00, 0.69, 0.69 |
| **26** | 0.2, 0.6, 0.8 | 2.00, 4.26, 5.12 | 1.00, 1.00, 1.00 | 0.38, 0.54, 0.56 | 0.20, 0.40, 0.40 | 3.00, 6.00, 8.00 | 3.00, 4.89, 5.71 | 1.00, 1.00, 0.98 |
| **27** | 0.2, 0.6, 0.8 | 3.00, 5.26, 6.51 | 1.00, 1.00, 1.00 | 0.88, 0.91, 0.91 | 0.00, 0.40, 0.60 | 1.00, 5.00, 8.00 | 1.00, 3.26, 4.51 | 1.00, 0.87, 0.89 |
| **28** | 0.2, 0.4, 0.6 | 2.00, 3.76, 5.01 | 1.00, 1.00, 0.99 | 0.38, 0.51, 0.55 | 0.00, 0.00, 0.20 | 1.00, 3.00, 6.00 | 1.00, 2.13, 3.34 | 1.00, 1.00, 0.84 |
| **29** | 0, 0.4, 0.8 | 1.00, 3.89, 6.35 | 1.00, 0.82, 0.83 | 0.13, 0.52, 0.54 | 0.20, 0.40, 0.40 | 2.00, 5.00, 7.00 | 2.00, 3.89, 4.71 | 1.00, 0.91, 0.91 |
| **30** | 0.2, 0.2, 0.2 | 3.00, 3.63, 4.06 | 1.00, 1.00, 0.98 | 0.88, 0.88, 0.89 | 0.00, 0.20, 0.20 | 0.00, 2.00, 3.00 | 0.00, 1.26, 1.69 | 0.00, 0.63, 0.64 |

**Табл. 3.** Метрики качества поиска.

Было вычислено среднее арифметическое каждой метрики на каждом уровне отдельно для обоих поисковых машин (**табл. 4**).

|  |  |  |
| --- | --- | --- |
|  | **Google** | **Собств.** |
| **P** | 0.17, 0.43, 0.59 | 0.06, 0.18, 0.25 |
| **DCG** | 2.57, 4.49, 5.55 | 0.97, 2.06, 2.64 |
| **NDCG** | 0.97, 0.96, 0.95 | 0.56, 0.67, 0.67 |
| **ERR** | 0.70, 0.79, 0.80 | 0.20, 0.30, 0.32 |

**Табл. 4.** Средние значения метрик.

Показатели метрик Google ожидаемо превосходят показатели метрик собственной поисковой машины. Уровень качества собственного поисковика можно соотнести с качеством собственного поиска Википедии несмотря на то что всё равно уступает последнему.

**Заключение**

Был рассмотрен способ ранжирования документов по формуле TF-IDF. Был разработан формат файла для хранения статистической информации корпуса документов. Размер статистического файла равен 189 Мб. Средняя скорость выполнения запроса составила порядка 2 секунд, что считается довольно долгим временем выполнения поискового запроса.

Было оценено качество поисковой выдачи по показателям P, DCG, NDCG и ERR. Аналогичные характеристики были вычислены в лабораторной работе №2 «Оценка качества поиска» для поисковика Википедии, что даёт возможность сравнить собственное качество поиска с поиском Википедии.

Из таблицы запросов видно, что односложные запросы выполняются очень хорошо. Двух и более сложные запросы отрабатываются сравнительно неплохо, однако в начало списка попадают документы имеющие мало общего с изначальным запросом. Запросы по типу вопросов обрабатываются очень плохо, так как обработка такого рода запросов требует наличия специальных алгоритмов.

В процессе выполнения запросов в работе поисковой программы были замечены следующие неточности:

* на некоторые запросы в начало списка попадают служебные документы Википедии не представляющие никакой ценности для пользователя;
* на запрос «гипертрихоз» программа вернула всего 4 документа;
* на запрос «известные биологи» программа вернула исключительно служебные документы, не содержащие искомой информации (были просмотрены все 50 первых документов из списка);
* документы из запроса «нейроны человека» оказались перемешаны между служебными документами.

**Источники**

[1] kshiian. Алгоритм BM25 [Электронный ресурс] // URL: <https://habr.com/ru/post/162937/>