ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

Факультет программной инженерии и компьютерной техники

ОТЧЕТ

НА ТЕМУ «Inverted index»

ПО ДИСЦИПЛИНЕ «Структуры и алгоритмы индексации данных»

Студент: Минина Дарья Николаевна
Группа: P4135
Преподаватель: к.т.н.
Платонов Алексей Владимирович

Санкт-Петербург
2024

СОДЕРЖАНИЕ

1. Инвертированный индекс

Инвертированный индекс — это структура данных, которая играет ключевую роль в информационном поиске и поисковых системах.

Инвертированный индекс — это структура данных, в которой для каждого слова в коллекции документов перечисляются все документы, в которых это слово встречается. Это позволяет быстро находить документы, содержащие определенные слова.

Обратный индекс обычно состоит из двух основных компонентов:

1. Список слов: содержит все уникальные слова из коллекции документов, расположенные в алфавитном порядке.
2. Документальные списки: для каждого слова указываются документы, в которых оно встречается.

Пример работы обратного индекса

Предположим, у нас есть три текста:

1. "It is what it is"
2. "What is it"
3. "It is a banana"

Обратный индекс для этих текстов может выглядеть так:

"a": {2}

"banana": {2}

"is": {0, 1, 2}

"it": {0, 1, 2}

"what": {0, 1}

Если мы хотим найти документы, содержащие слова "what", "is" и "it", мы просто берем пересечение списков для этих слов: {0, 1}.

Применение:

1. Поиск по одному слову: достаточно найти список, соответствующий этому слову.
2. Поиск по нескольким словам: берется пересечение списков для каждого слова.

3. Ранжирование документов: после поиска часто применяется ранжирование документов по их релевантности.

Особенности:

- Обратный индекс позволяет выполнять быстрый поиск по текстам.
- Он является одной из самых популярных структур данных в информационном поиске.
- Обратный индекс можно дополнить дополнительной информацией, такой как позиция слова в документе.

Заключение:

Обратный индекс — это мощный инструмент для оптимизации процесса поиска информации. Он лежит в основе многих современных поисковых систем и обеспечивает высокую производительность поиска по тексту.

2. Результаты

В лабораторной работе были реализован inverted index. Бенчмарки с тестированием на различных запросах приложены ниже.

Evaluating my queries...

Top 10 documents in rank list
Query: 1 Pr: 0.0 Re:0.0
Query: 2 Pr: 0.2 Re:0.13333333333333333
Query: 3 Pr: 0.2 Re:0.13333333333333333
Query: 4 Pr: 0.1 Re:0.05555555555555555
Query: 5 Pr: 0.1 Re:0.05263157894736842
Query: 6 Pr: 0.4 Re:0.2222222222222222
Query: 7 Pr: 0.6 Re:0.6666666666666666
Query: 8 Pr: 0.2 Re:0.5
Query: 9 Pr: 0.1 Re:0.125
Query: 10 Pr: 0.2 Re:0.08333333333333333
Avg precision: 0.21000000000000002
Avg recall: 0.19720760233918128

Top 50 documents in rank list
Query: 1 Pr: 0.0 Re:0.0
Query: 2 Pr: 0.12 Re:0.4
Query: 3 Pr: 0.14 Re:0.4666666666666667
Query: 4 Pr: 0.06 Re:0.16666666666666666
Query: 5 Pr: 0.14 Re:0.3684210526315789
Query: 6 Pr: 0.14 Re:0.3888888888888889
Query: 7 Pr: 0.16 Re:0.8888888888888888
Query: 8 Pr: 0.06 Re:0.75
Query: 9 Pr: 0.12 Re:0.75
Query: 10 Pr: 0.08 Re:0.16666666666666666
Avg precision: 0.10200000000000001
Avg recall: 0.4346198830409357

Top 100 documents in rank list
Query: 1 Pr: 0.0 Re:0.0
Query: 2 Pr: 0.09 Re:0.6
Query: 3 Pr: 0.09 Re:0.6
Query: 4 Pr: 0.06 Re:0.3333333333333333
Query: 5 Pr: 0.13 Re:0.6842105263157895
Query: 6 Pr: 0.09 Re:0.5
Query: 7 Pr: 0.09 Re:1.0
Query: 8 Pr: 0.03 Re:0.75
Query: 9 Pr: 0.06 Re:0.75

Query: 10 Pr: 0.04 Re:0.16666666666666666
Avg precision: 0.068
Avg recall: 0.5384210526315789

Top 500 documents in rank list
Query: 1 Pr: 0.002 Re:1.0
Query: 2 Pr: 0.03 Re:1.0
Query: 3 Pr: 0.03 Re:1.0
Query: 4 Pr: 0.032 Re:0.8888888888888888
Query: 5 Pr: 0.038 Re:1.0
Query: 6 Pr: 0.036 Re:1.0
Query: 7 Pr: 0.018 Re:1.0
Query: 8 Pr: 0.008 Re:1.0
Query: 9 Pr: 0.016 Re:1.0
Query: 10 Pr: 0.026 Re:0.5416666666666666
Avg precision: 0.0236
Avg recall: 0.9430555555555555

ЗАКЛЮЧЕНИЕ

В результате выполнения лабораторной работы был изучен принцип работы inverted index, проведено тестирование на различном количестве документов.