Оптимизация выполнения запросов по диапазонам для библиотек индексации, построенных на основе обратного индекса

Научный руководитель: Пименов А.А.
Рецензент: доктор физ.-мат. наук, профессор
Новиков Б.А.
Выполнил: Евстифеев С.В.

Математико-механический факультет СПбГУ
2010 год

Поисковые системы

- Очень распространены
- Поиск по слабоструктурированным архивам данных
- Гибкая схема
- Сортировка документов по релевантности
- Информация о документе хранится в полях. К примеру, «имя документа», «содержимое», «дата публикации»

Библиотека индексации Lucene

- Полнофункциональная библиотека. Присутствуют средства для обновления индекса.
- Высокая производительность
- Масштабируемость
- Открытый исходный код
- Единственная в своем роде система
- Активно используется

Структура библиотек индексации

- Пользователи выполняют запрос по набору термов (bag-of-word)
- Обратный индекс: отношение терм → документы
- Требуется предварительная индексация

Запросы по диапазонам

- Легко реализуются в базах данных
- Структура обратного индекса слабо приспособлена для таких запросов
- Возможен только поиск по набору термов – необходима предварительная обработка запроса: [1,5] → {1,2,3,4,5}
- Что если большие диапазоны или вещественные значения?
- Нужно оптимизировать

Существующие решения

- Кэширование
- Можно заменить на простую фильтрацию без сортировки по релевантности
- Индексирование с помощью нарастающих префиксов:
  `C CC CCY CCYY CCYYMM CCYYMMD CCYYMMDD`
  Пример: все года после 1990го, поиск по «199»
- Используется двоичное представление чисел

Существующий алгоритм

- Разбивает интервал поиска на диапазоны, по которым можно быстро искать с использованием префиксного индексирования
- Неоптимально работает для диапазонов [0, 2^k - 2] и близких к ним
- Не принимается во внимание возможность вычитания диапазонов

Предложение по оптимизации

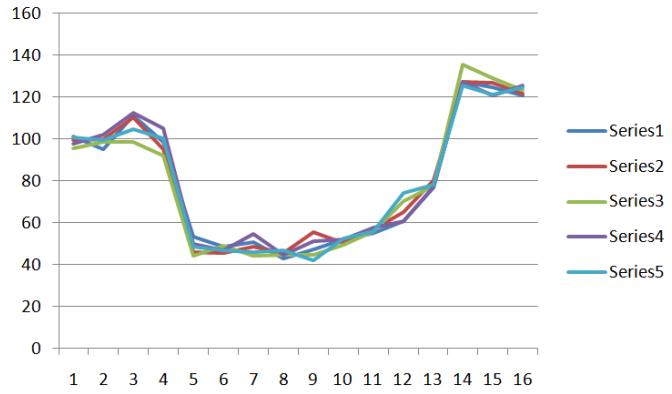
- Использовать вычитание диапазонов для ускорения работы, к примеру:
  [0, 2^k–2] = [0, 2^k-1] / [2^k-1, 2^k-1]
  k операций vs 2 операции
- Вычисление количества поисковых операций для исходного и альтернативного запросов, определение оптимального
- Результат выполнения запроса не изменяется, требуется некоторая последующая обработка

Выполнено

- Создан и описан алгоритм
- Реализован алгоритм с учетом задаваемых параметров
- Проведено тестирование и анализ результатов на наборе из 500тыс документов

Статистика

- Алгоритм уменьшает время работы для специфических диапазонов [0, 2^k-2-A], где A >= 0, A – не велико
- До 50% выигрыш в скорости: отношение времени работы в % опт. алгоритма к времени работы исходного для интервалов [0, 2^k-2]



Заключение

- Возможность применения для любой библиотеки индексации, использующей обратный индекс (не только Lucene)
- Оптимизация в худшем случае – важно для поисковых систем
- Не требуется изменять сборку Lucene, останавливать систему. Модификация запросов пользователя перед подачей на вход системы.