Влияние селективности на исполнение запросов в структурах многомерного индексирования

Федотовский Павел Валерьевич, гр. 545

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра системного программирования

Научный руководитель: ассистент кафедры информатики, Чернышев Г.А.

Санкт-Петербург
2014 г.

Введение

- Индексирование
  - B^+-дерево
- Многомерное индексирование
  - Геоинформационные системы, мультимедиа СУБД, ...
  - Запросы на диапазон
- Селективность запроса q: $\frac{m}{n}$, где n - общее количество записей, m записей удовлетворяют запросу
- Транзакционные системы
  - In-memory

Мотивация

- Выбор оптимальной структуры данных
  - Селективность
  - Размерность данных
  - Распределение данных
  - Выборка первых k записей
- Применение: оптимизация производительности
  - Автоматическая настройка БД

Постановка задачи

- Провести обзор существующих алгоритмов для многомерного индексирования
- Реализовать тестовую систему для проведения измерений на существующем прототипе многомерного индекса
- Исследовать влияние селективности на производительность различных подходов к многомерному индексированию при различных сценариях

Обзор алгоритмов

- Сведение к одномерному случаю
- Иерархические структуры:
  - B^+-дерево
  - R-дерево
  - X-дерево
  - K-D-дерево
  - ...
- Структуры, основанные на хэшировании
  - LSH (Locality-Sensitive Hashing)
  - GRID-файл
  - ...

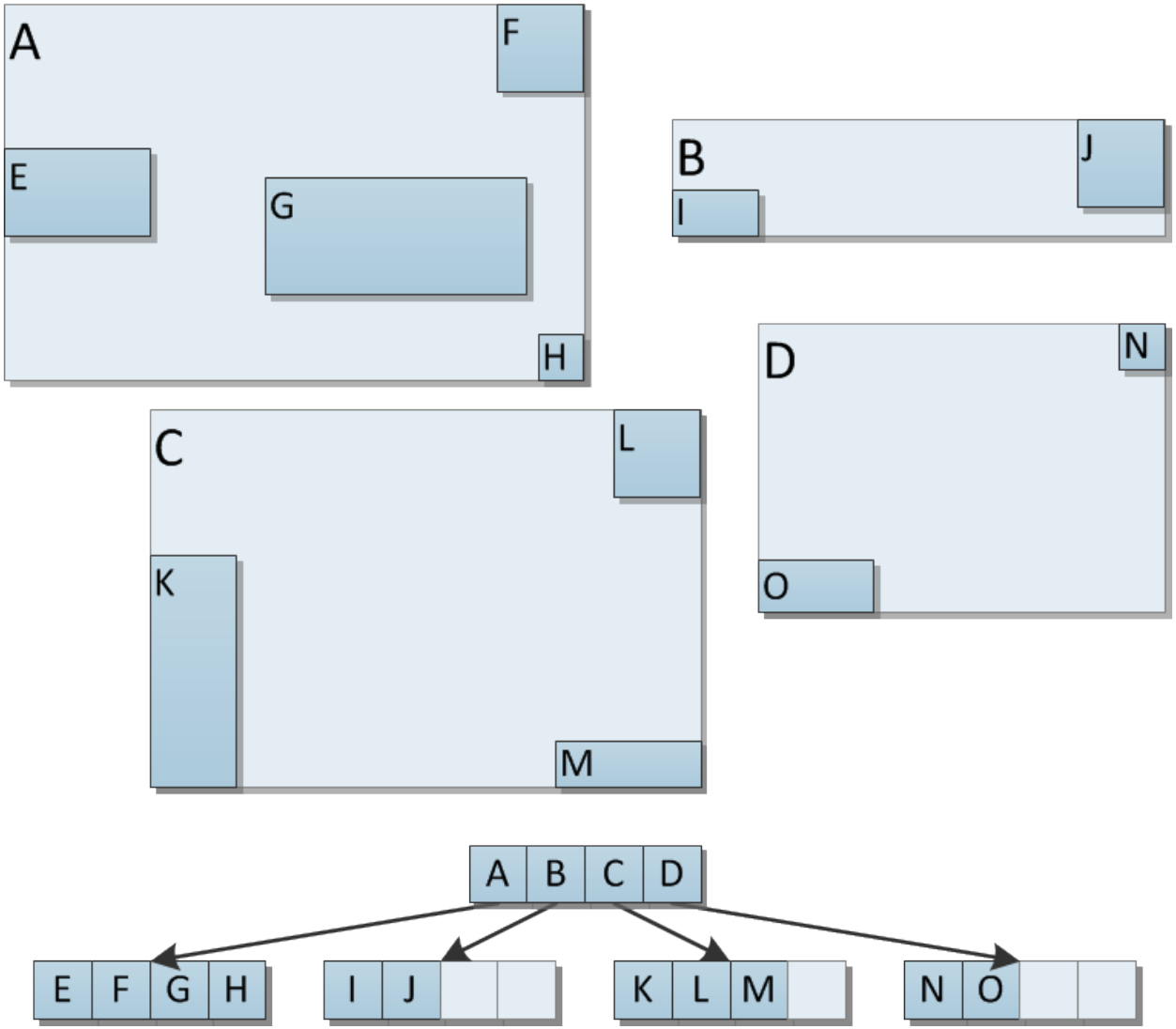
B^+-дерево

- Данные хранятся только в листьях (в отличие от B-дерева)
- Преимущества
  - Множество алгоритмов обеспечения корректности параллельного доступа
  - Данные хранятся в упорядоченном виде
- Недостатки
  - Низкая производительность исполнения запросов на диапазон в случае высокой селективности

R-дерево

- Обобщение B^+-дерева на многомерный случай
  - Данные также хранятся только в листьях
  - Обе структуры сбалансированы
  - Вершины хранят ограничивающий прямоугольник, что можно считать обобщением интервалов в B^+-дереве
- Реализовано в PostgreSQL, Oracle, MySQL, Informix, SQLite, ...

R-дерево: пример



Исследование влияния селективности

- Выбор платформы
  - Elki, Amdb, ...
- Выбранный прототип многомерного индекса:
  - Разрабатывался на матмехе СПбГУ в рамках гранта РФФИ №12-07-31050
  - Реализация R-дерева, B^+-дерева
  - In-memory индекс
  - Уровень изоляции транзакций - read committed

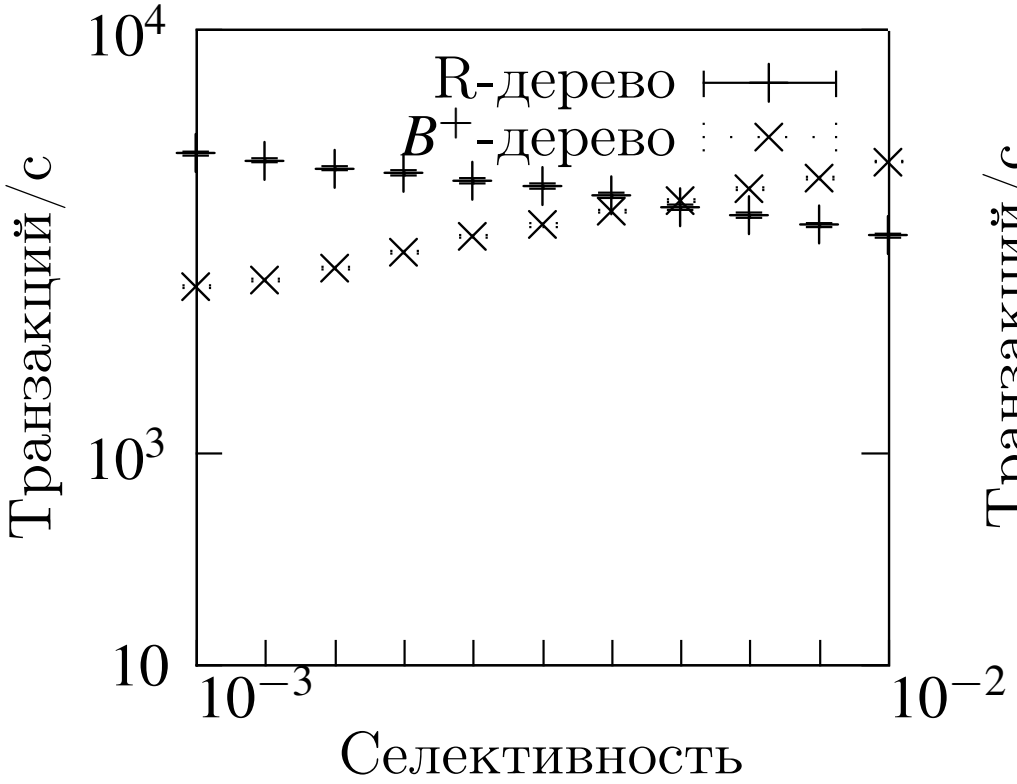
Тестовая система

- Модуль для генерации начальных данных и запросов
  - Размер индекса
  - Размерность данных
  - Распределение данных
  - Селективность запросов
- Модуль для проведения измерений
  - Время заполнения индекса
  - Количество транзакций в секунду

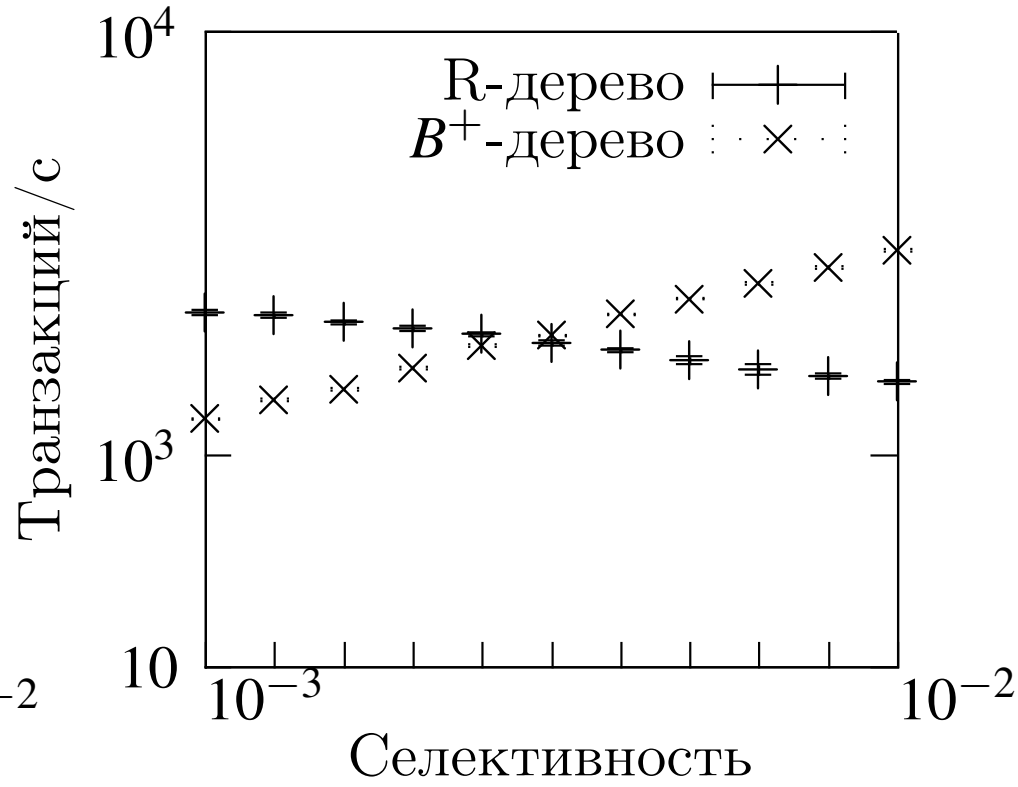
Тестовый стенд

- Конфигурация 1:
  - Аппаратное обеспечение: Intel Core i7-2630QM, 2.00 GHz, 6GB RAM
  - Программное обеспечение: x86_64 GNU/Linux, ядро 3.11.0-12-generic.
- Конфигурация 2:
  - Аппаратное обеспечение: Intel Core i7-3770, 3.40GHz, 8GB RAM
  - Программное обеспечение: x86_64 GNU/Linux, ядро 3.8.0-26-generic.

Измерения



Равномерное распределение, размерность 2



Равномерное распределение, размерность 4

Анализ

$P = a * S^b$

- P пропускная способность
- a и b параметры
- S селективность запроса

Распределение	2	4	6	8
Равномерное	0.60 ± 0.03	0.80 ± 0.03	0.85 ± 0.02	0.87 ± 0.02
Нормальное	0.80 ± 0.06	0.89 ± 0.03	0.94 ± 0.03	0.94 ± 0.02
Закон Ципфа	0.57 ± 0.03	0.79 ± 0.02	0.85 ± 0.02	0.87 ± 0.02

Распределение	2	4	6	8
Равномерное	0.62 ± 0.02	0.79 ± 0.03	0.86 ± 0.03	0.88 ± 0.01
Нормальное	0.81 ± 0.06	0.92 ± 0.04	0.99 ± 0.04	1.00 ± 0.03
Закон Ципфа	0.57 ± 0.04	0.79 ± 0.02	0.86 ± 0.02	0.86 ± 0.03

Значения параметра b для B^+-дерева

Результаты

- Проведен обзор существующих алгоритмов для многомерного индексирования
- Реализована тестовая система для проведения измерений на существующем прототипе многомерного индекса
- Исследовано влияние селективности на производительность B^+-дерева, R-дерева
- По теме работы была опубликована статья на конференции SYRCoDIS'13