# Реализация нечеткого поиска в условиях низкого порога схожести в системе обработки документов

Чередник Кирилл, 445 гр.

Научные руководители:
Чернышев Г.А.
Смирнов К.К.

Кафедра системного программирования, СПбГУ, 2013

# Введение

- Нечеткий поиск - нахождение всех строк из набора, близких к данной по заданной метрике на пространстве строк.
  - Информационный поиск
  - Биоинформатика
  - Проверка орфографии

- Соревнования, Работы, Прототипы
  - SIGIR, SIGMOD…
- ACM SIGMOD Programming Contest 2013

# Постановка задачи

Цель работы – разработка методов, реализующих нечеткий поиск в системе, обладающей следующей спецификой:

- Малый порог разницы
- Короткие слова
- Поддержка метрик (Дискретная, Хэмминга, Левенштейна)

Задачи:

- Обзор общих решений
  - Фильтрации
  - Сравнения
- Разработка  структур данных и алгоритмов, учитывающих специфику системы

# Обзор существующих решений (фильтрация)

- Дискретная
  - Hash table
- Хэмминга и Левенштейна
  - M-tree
  - BK-tree
  - Trie
  - FQA
  - VP-tree

# Обзор существующих решений (сравнение)

- Дискретная метрика
- Метрика Хэмминга
- Метрика Левенштейна
  - Прямой подход
  - Отсечение Укконена
  - Автоматы

# Предложенные решения

- Фильтрация
  - **Метрика Хэмминга**
  - Метрика Левенштейна
- Сравнение
  - Метрика Хэмминга
  - Метрика Левенштейна

# Предложенные решения: Фильтрация, Хэмминг

- Хэш таблицы
  - Слово “abcdefg” разбивается на части


ε =3

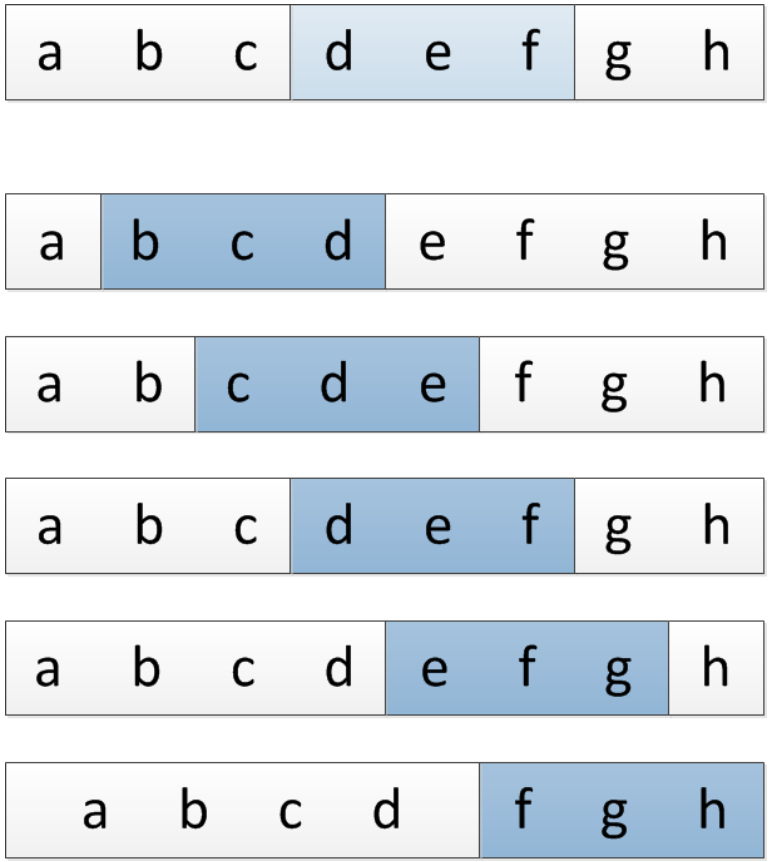
  - Таблицы работают с частями слова

# Предложенные решения

- Фильтрация
  - Метрика Хэмминга
  - **Метрика Левенштейна**
- Подсчёт расстояния
  - Метрика Хэмминга
  - Метрика Левенштейна

# Предложенные решения: Фильтрация, Левенштейн

- Хэш таблицы
  - Схожи с хэш-таблицами для Хэмминга, но учитываются сдвиги.




ε =2

# Предложенные решения: Фильтрация, Левенштейн

- Битовые маски

| | a | b | c | x | y | f | g |
|---|---|---|---|---|---|---|---|
| 0 | **1** | **1** | **1** | 0 | 0 | 0 | 0 |
| 1 | **1** | **1** | **1** | **1** | 0 | 0 | 0 |
| 2 | **1** | **1** | **1** | **1** | **1** | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | **1** | **1** | **1** | **1** |
| 6 | 0 | 0 | 0 | 0 | **1** | **1** | **1** |
| … | … | … | … | … | … | … | … |
| 23 | 0 | **1** | **1** | **1** | **1** | **1** | **1** |
| 24 | 0 | 0 | **1** | **1** | **1** | **1** | 0 |
| 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

ε =2

# Предложенные решения

- Фильтрация
  - Метрика Хэмминга
  - Метрика Левенштейна
- Подсчёт расстояния
  - **Метрика Хэмминга**
  - Метрика Левенштейна

# Предложенные решения: Сравнение, Хэмминг

- SSE4.2

- Тип __m128i, операции:
  - _mm_cmpestrm(a, b)
  - _mm_popcnt_u64(a)

# Предложенные решения

- Фильтрация
  - Метрика Хэмминга
  - Метрика Левенштейна
- Подсчёт расстояния
  - Метрика Хэмминга
  - **Метрика Левенштейна**

# Предложенные решения: Сравнение, Левенштейн

- Автоматы

- Отдельные ДКА для проверяемых слов разной длины (всего $2 \times \varepsilon + 1$ типов автоматов).

# Результаты

- Проделан обзор существующих подходов
- Разработаны алгоритмы и структуры данных для решения задачи, учитывающие специфику системы
- Результаты будут доложены на конференции ACM SIGMOD 2013