Яндекс на РОМИП-2004

Илья Сегалович, Михаил Маслов
{iseg, maslov}@yandex-team.ru

Поиск и ранжирование

- Архитектурные особенности
- Фильтрация
- Ранжирование
- Результаты на РОМИП-2004

Архитектурные особенности

- Инвертированный массив полных словопозиций
- Обработка: одновременный проход (Document Ordered)
- Субиндекс: self-indexing inverted files
- Операции пересечения: полная словопозиция+контекстные ограничения+субиндекс

Фильтрация

- Контекстные ограничения
  - Многоместный оператор AND
  - Кворум: Коэффициент мягкости
    (несколько **op1** *слов* **op2** *запроса) //* **softness**
  - Контекст документа; предложения; назначенный пользователем; назначенные системой
  - Контекст, назначенный системой
    - Синтаксический разбор
    - Статистика слов в индексе

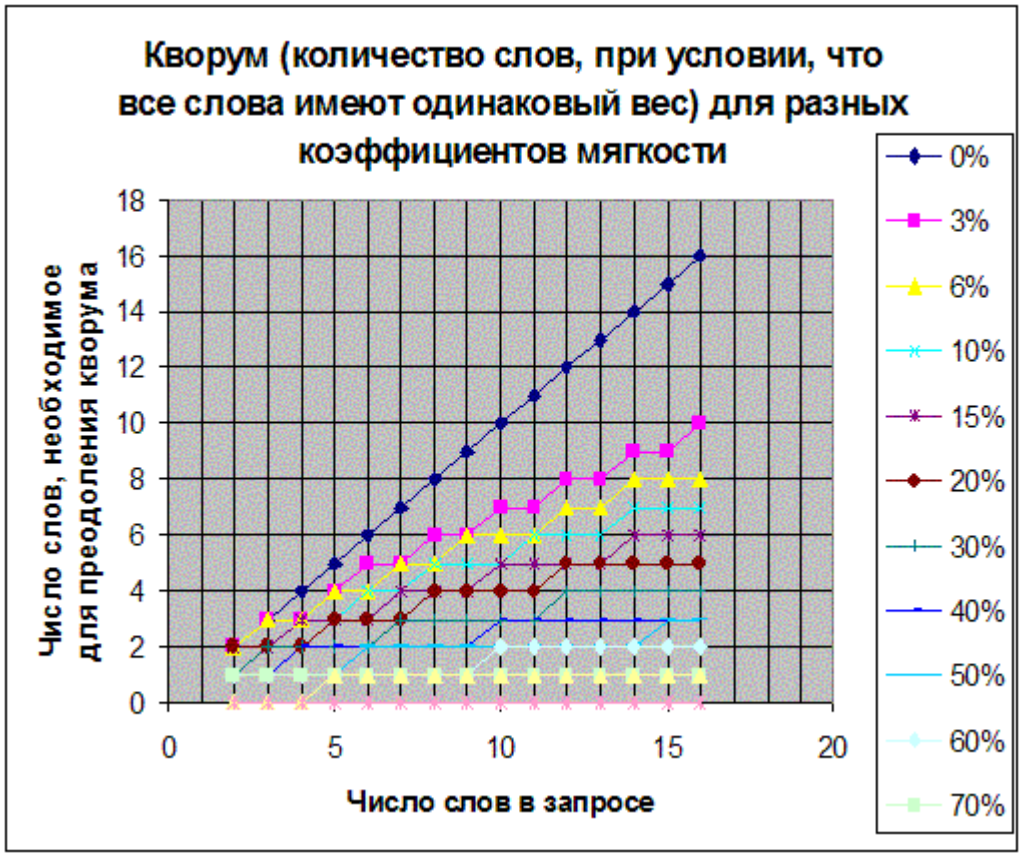
Кворум

- Формула для доли веса

$$QuorumWeight = (1 - Softness)^{\frac{1}{\sqrt{QL-1}}}$$

где *QL* – число слов в запросе
Softness – число от 0 до 1

Кворум


Кворум (количество слов, при условии, что все слова имеют одинаковый вес) для разных коэффициентов мягкости
Число слов, необходимое для преодоления кворума
0
2
4
6
8
10
12
14
16
18
0
5
10
15
20
Число слов в запросе
0%
3%
6%
10%
15%
20%
30%
40%
50%
60%
70%

Принципы ранжирования

- Внутридокументная частота (TF) по релевантным пассажам
  - Корректное нормирование на длину документа
- Вычисляемый вес каждой словопозиции
  - Пример, где второй пассаж ранжируется выше
    [*aa* ___ *BB* ___ *cc dd ee*]
    [*aa BB cc dd* ___ ___ *ee*]

Ранжирование: контекстуальное сходство

- Объемлющие пассажи игнорируются
- Считается вес для каждой опоры: вес пассажа равен сумме весов опор
- Ранг неполных пассажей строго меньше ранга полных
- Вес опоры убывает с ростом расстояния между опорами

Ранжирование: форматирование текста

- Вес слова учитывает его вхождение в заголовки
- Бонус для полных насыщенных пассажей
- Учет при индексировании

Результат по прогонам

- Выбранные параметры
  - Softness = 6
  - Без группировки по хостам (для веб-коллекции)
  - Документный контекст
- Результаты
  - (на словах)