Российские
интернет-технологии
2012

Поиск на своем сайте, обзор open source решений

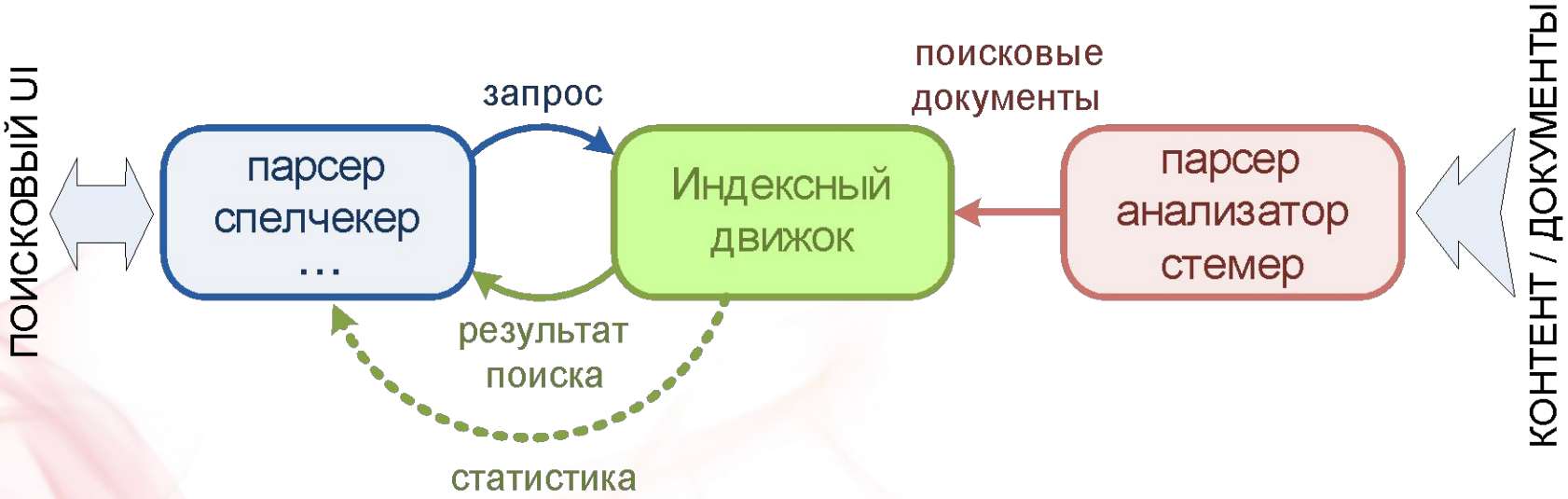
Олег Бунин

Поиск на своём сайте

- Искать SQLem по своей базе
  - Вы наверное не слышали об альтернативах
- Использовать поиск с Google Яндекс
  - Отличный вариант для небольших статических сайтов
- Установить у себя поисковый движок

Apache Solr Sphinx elasticsearch Яндекс сервер

Ингредиенты поиска



Поисковый документ

Анализ документа:

токенизаиця , стеминг, морфология, стоп слова

Термы

```
id:18464775329560985 7
author:pledbook tag:grails
tag:elasticserach text:look
text:use text:plugin
text:grails text:website
url:grails.org/plugin/...
date:2012.03.27
```

Поисковый индекс

Типичная RDBMS

- Инвертированный индекс
- Индекс по отдельному полю или композитному полю
- Одно значение на поле (full text – много значений)
- *Обычно* только один из индексов используется в select`e
- Btree – эффективные апдейты

Поисковый индекс

Типичный поисковый движок

- Инвертированный индекс
- Все поля в одном индексе
- Произвольное количество значений поля на документ
- Поиск происходит по всем полям запроса одним махом
- Интегрированный скоринг
- Плотный бинарный формат индекса – read only

Поисковый запрос vs SQL

SQL

- сложный запросы / joins
- детерминированный запрос
- возвращает данные

Поисковый запрос

- “плоский” select
- сложные комбинации атрибутов

Open Source

Lucene – библиотека / фреймворк - Java

- Solr – всё в одном, прицел на “enterprise”
- elasticsearch – простота
- IndexTank – crowdsourcing

Sphinx – скорость - C++

Что **нужно** хотеть от поиска?

в дополнение к качественным
и релевантным результатам

- Фейсетная навигация
  - авто-таксономия
  - кластеризация
- Автокоррекция

- Подсказки (автодополнение)

Фейсеты

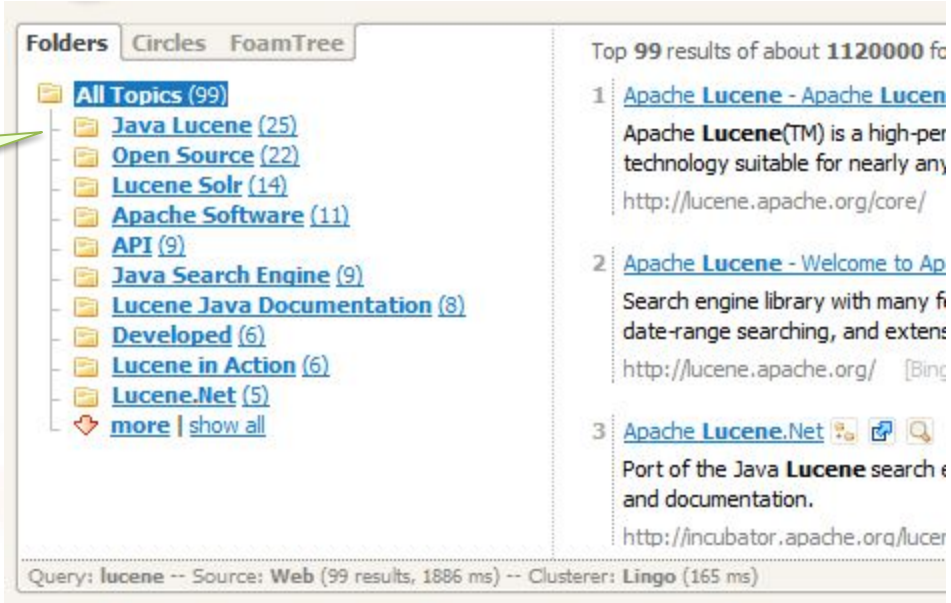
Динамический набор категорий

Активный фильтры

Число найденых документов по запросу и фильтру

Кластеризация

Кластеры создаются на основе текста документов

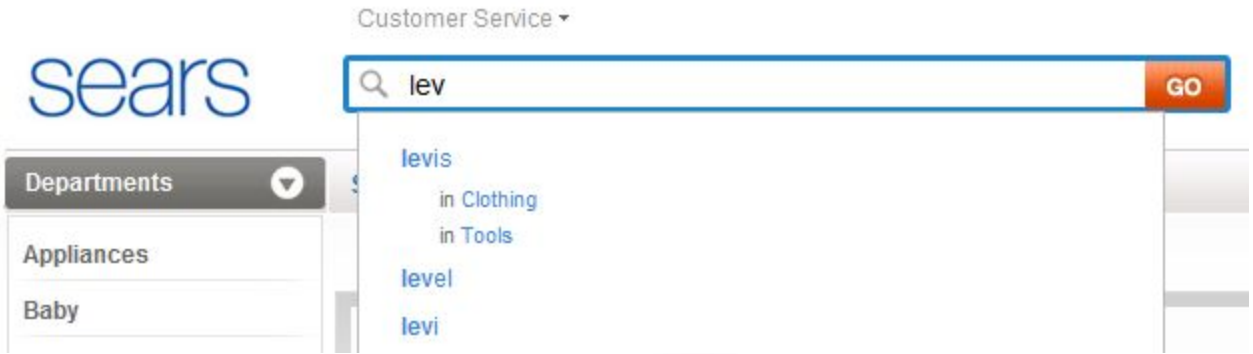


Автокоррекция

n-gram индекс

- Отдельный индекс для коррекции

Подсказки



Похоже на автокоррекцию. Запрос “дописывается” на основе словаря в индексе и дополняется ключевыми словами на основе статистики.

Что индексировать?

Каталог продуктов

- Джинсы Levis #559, индиго
- Джинсы Levis #559, индиго, размер 32x32

Что считать документом?

Когда индексировать?

Поисковые индексы нужно перестраивать

Периодическая переиндексация всех документов

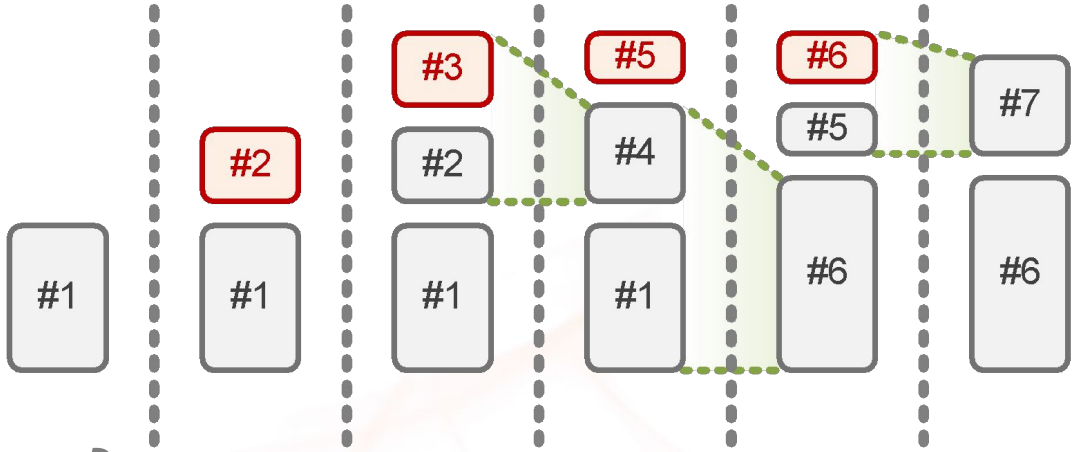
- Динамические атрибуты (пример наличие на складе)

Сегментированный индекс

- Позволяет изменять набор документов без перестройки всего индекса
- Требует регулярной “оптимизации” индекса

Индексные сегменты в Lucene

Логарифмическое слияние сегментов



Самые молодые сегменты можно не спешить писать на диск

Масштабирование

Производительность

- упирается в CPU
- сложные запросы могут быть очень CPU-ёмкими
- решение – несколько реплик индекса
  - read-only файлы облегчают задачу

Масштабирование

Объём индекса

- Индекс должен помещаться в память
- Решение – партицирование индекса
  - Каждая партиция выполняет скоринг независимо
  - Результаты нужно агрегировать

Резюме: Lucene

- Фреймоворк/библиотека
- Java API (нет сетевого интерфейса)
- Фундамент построения поисковой системы

Резюме: Solr

- Feature reach поиск из коробки
- Эффективная поддержка фейсетов
- Кросс платформенные клиент (HTTP)
- Интеграция со многими CMS
- Управление распределённым индексом и репликацией
- Фокус: “entrprise” приложения
- Обширная экосистема

Резюме: Sphix

- Простой и быстрый
- Интеграция с MySQL
- Интеграция со многими CMS
- Базовый поисковый функционал
  - Нет фейсетов, подсказок и т.п.
- Распределённый поиск (партицирование)

Резюме: elasticsearch

- Управление распределённым индексом
- Простой HTTP API
- Использует Lucene
- Фейсеты
- Проще в настройке чем Solr
- Фокус: простота и масштабируемость

Резюме: Index tank

Index tank появился как поисковый SaaS. После покупки компании, код продукта был опубликован как open source.

- Фокус: социальный контент и crowdsorcing
- Скоринг по динамическим атрибутам (голоса и т.п.)
- Управление поисковым “облаком”

Спасибо

Алексей Рагозин
alexey.ragozin@gmail.com

Российские
интернет-технологии
2012