



Российские  
интернет-технологии  
2012

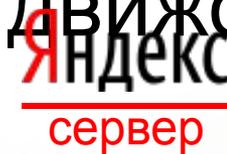
# Поиск на своем сайте, обзор open source решений

Олег Бунин

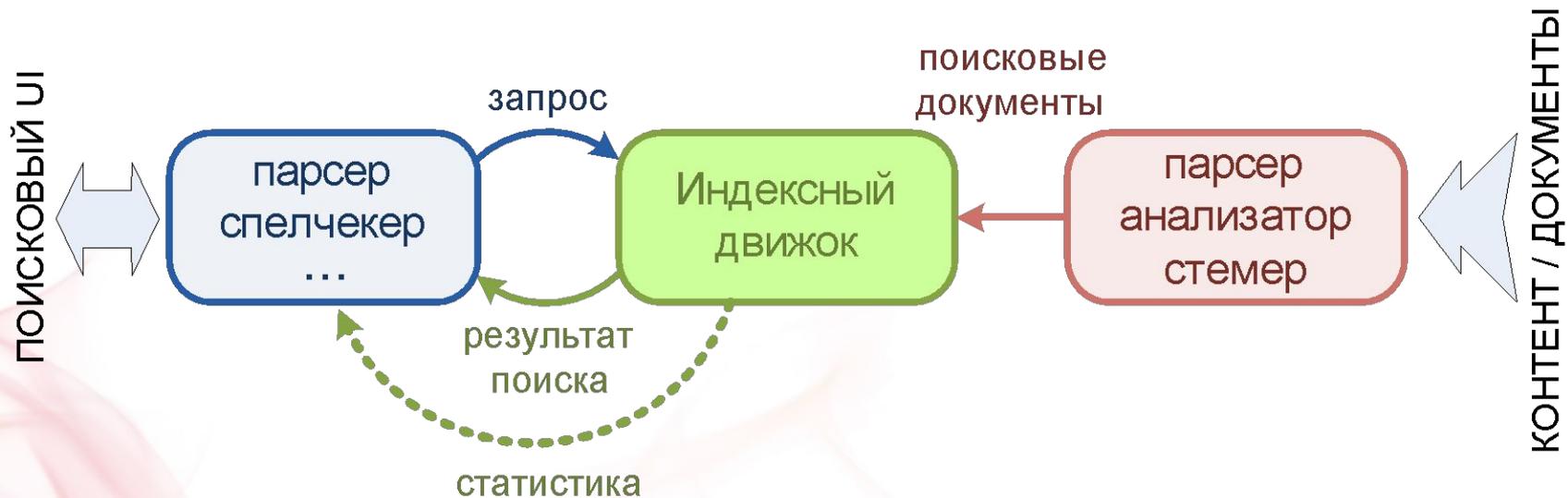
# Поиск на своём сайте

- 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



# Ингредиенты поиска



# Поисковый документ



## Термы

id:184647753295609857  
author:pledbook tag:grails  
tag:elasticsearch text:look  
text:use text:plugin  
text:grails text:website  
url:grails.org/plugin/...  
date:2012.03.27



## Анализ документа:

токенизация, стеминг,  
морфология, стоп слова



# 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`е
- 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

- сложные 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



Российские  
интернет-технологии

2012

- Подсказки (автодополнение)

# Фейсеты

## Digital cameras

Динамический набор категорий

Активные фильтры

Число найденных документов по запросу и фильтру

Refine your results

Manufacturer	Flash memory	Resolution	More
Sony (39)	SD Memory Card (134)	Less than 3 megapixels (4)	Maximum ISO
Nikon (20)	SDHC Memory Card (103)	4 megapixels (4)	Weight
Panasonic (19)	MultiMediaCard (70)	5 megapixels (17)	Optical sensor type
Canon (37)	SDXC Memory Card (47)	6 megapixels (8)	Zoom range
Olympus (19)	Memory Stick (36)	7 megapixels (22)	
Fujifilm (11)	MultiMediaCardplus (24)	8 megapixels (62)	

See all

You selected: \$300 - \$400 Compact remove all

187 results

Show 10 results per page Sort by: Review date Compare Selected

 **Canon PowerShot SX260 HS (Black)** \$336 to \$374 at 11 stores

Editors' rating: ★★★★★  
Reviewed on 03/24/2012

**The Bottom Line:** The Canon PowerShot

\$336 - Amazon.com  
\$349 - Abes Of Maine

# Кластеризация

Кластеры  
создаются на  
основе текста  
документов

The screenshot displays a search interface with a folder tree on the left and search results on the right. The folder tree is titled 'Folders' and includes tabs for 'Circles' and 'FoamTree'. The 'All Topics (99)' folder is expanded, showing sub-folders such as 'Java Lucene (25)', 'Open Source (22)', 'Lucene Solr (14)', 'Apache Software (11)', 'API (9)', 'Java Search Engine (9)', 'Lucene Java Documentation (8)', 'Developed (6)', 'Lucene in Action (6)', and 'Lucene.Net (5)'. A 'more | show all' link is visible at the bottom of the folder list. The search results on the right are titled 'Top 99 results of about 1120000 fo' and list three items: 1. 'Apache Lucene - Apache Lucene' with a description of the technology and a URL; 2. 'Apache Lucene - Welcome to Ap' with a description of the search engine library and a URL; 3. 'Apache Lucene.Net' with a description of the port of the Java Lucene search engine and a URL. At the bottom of the interface, the query is 'lucene' and the source is 'Web (99 results, 1886 ms)'. The clusterer used is 'Lingo (165 ms)'.



# Автокоррекция

The screenshot shows the blekko search engine interface. The search bar contains the text "lucne". Below the search bar, a light blue banner displays the suggestion "Did you mean **lucene**?", where "lucene" is highlighted in blue. To the left of the main content area, there are navigation links for "WEB", "IMAGES", "VIDEOS", and "LOCAL". Below the suggestion banner, the search results section shows "1 to 10 of 71 results" and a "sort by: date · relevance" option. The first result is titled "Facebook - Johan Lucne" and includes a brief description of the user's profile on Facebook, along with a link to the profile page.

n-gram индекс

- Отдельный индекс для коррекции

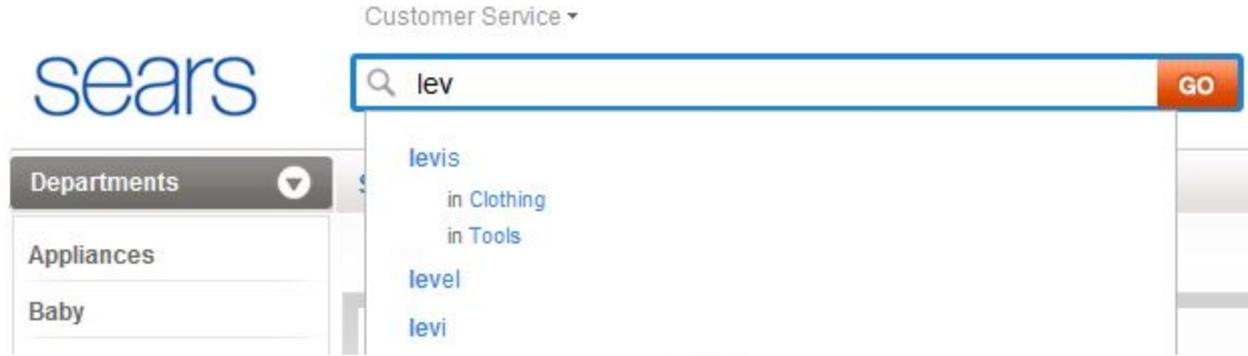


Российские  
интернет-технологии

2012

Нечёткий поиск

# Подсказки



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



# 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

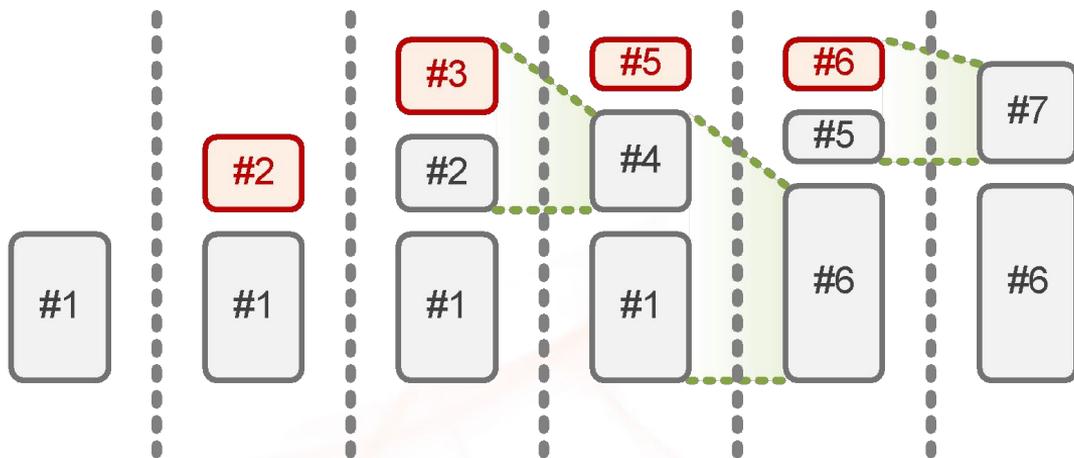
- Позволяет изменять набор документов без перестройки всего индекса

- Требуется регулярной “оптимизации” индекса



# Индексные сегменты в Lucene

## Логарифмическое слияние сегментов



*Самые молодые сегменты можно не спешить писать*



# 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 – партиционирование индекса
  - Каждая партиция выполняет скоринг независимо
  - Результаты нужно агрегировать



# Резюме: Lucene

- Фреймворк/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: “enterprise” приложения
- Обширная экосистема



# Резюме: Sphinx

- 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 crowdsourcing
- Скоринг по динамическим атрибутам (голоса и т.п.)
- Управление поисковым “облаком”



# Спасибо

Алексей Рагозин  
[alexey.ragozin@gmail.com](mailto:alexey.ragozin@gmail.com)



Российские  
интернет-технологии  
2012