Разработка алгоритмов обеспечения качества распределенного поискового робота для сети Интернет

Волков Сегрей Андреевич, 545 гр.
Научный руководитель: Выговский Л.С.

Индексация новостей

- Особенности:
  - дата;
  - регион;
  - тематика.
- Применение
  - Маркетинговые, PR агенства, пресслужбы...
  - Роскомнадзор

Цель работы

- Создание системы, способной качественно индексировать новости в русскоязычном интернете
  - ~5 000 сайтов
  - ~50 000 документов в сутки
  - 12 часов – разница между публикацией в rss и попадением в рабочий индекс
- Работа выполняется в рамках проекта Skai

Поисковые роботы

- DataparkSearch
- AspSeek
- **Nutch**
- Hounder

Подход

1. Запустить систему без модификаций
2. Найти самое слабое место
3. Принять меры по его устранению
4. Перейти к пункту 2

Ранжирование ссылок

- Scoring Filter
  - On-Line Page Importance Computation
    - Важность страницы определяется важностью страниц, на неё ссылающихся
    - 10% из скачиваемых web-страниц новости
  - SkaiScoring
    - Работа на основе регулярных выражений
    - 80% из скачиваемых web-страниц новости

Слияние индексов

- Стандартная реализация
  - Синхронно
  - Каждый раз перезаписывается весь индекс
  - 2M документов (70GB) – 11 часов на c1.medium
- Оптимизация
  - Не связано с основным циклом
  - Древовидная стратегия слияния

Раннее удаление дубликатов

- Key-Value Storage
  - (url,segmentid,taskid)
  - (md5,segmentid,taskid)
  - MongoDb
- Происходит в момент индексации
- Готовые индексы не редактируются

Автоматическое создание фильтров

- URLFilter
  - *.avi *.mp3
  - foonews.ru/forum/*
  - foonews.ru/out/*
  - ~~foonews.ru/news/*~~
  - ~~foonews.ru/archive/*~~
- Анализ базы URL
  - Crawldb >> Index

Результаты

- Сделан сравнительный анализ различных поисковых роботов
- Изменено поведение Nutch для работы с индексом большого объема
- Разработан и реализован плагин для раннего удаления дубликатов
- Разработан и реализован плагин для ранжирования ссылок
- Разработана и реализована система для автоматического создания фильтров
- Измененная система протестирована на реальных данных
- Все изменения включены в стабильную версию системы