Оптимизация процесса сборки документов системой nutch

Волков С. А. Научный руководитель: Выговский. Л. С.

14 мая 2010 г.

O nutch

- Основное назначение — загрузка и обработка документов
- Набор работ для hadoop
- Разработан apache
- Инкрементальная сборка

Дополнительная информация

- Ограничение области поиска
- $http://lenta\.ru/news/\d\{4\}/\d\{2\}/\d\{2\}/\w+/$
- $url \longrightarrow$ “полезность“

Цель

- Повышение эффективности сборки, учитывая дополнительную информацию об url
- Оценка эффективности: n_i / t_i
  - n_i — число полученных “полезных” url за итерацию
  - t_i — время итерации
- Оценка времени: $t_i = n_i \cdot c_1 + n_g \cdot c_2$
  - n_i — размер базы ссылок
  - n_g — размер выборки
  - $c_1 / c_2 \approx 0.0013$

Подходы

- ранжирование выборки
  - rss
  - сперва “полезные”
- фильтрация
  - технические
  - остальные
  - важно сохранить достижимость “полезных” url
    $http://lenta\.ru/\d{4}/\d{2}/\d{2}/$

Результаты

- Ранжирование
  - Повышение эффективности порядка 8 раз
  - Ранний этап сборки
- Автоматическая генерация фильтров
  - Удачно отсекаются технические разделы
  - Незначительные накладные расходы
  - Поздний этап сборки

Итоги

- Изучен Nutch
- Реализованы алгоритмы
- Повышение эффективности на всем протяжении жизни