Универсальный парсер для вертикального поиска

Протасов С., Плошихин В.

Rambler Media

РИТ Высокие Нагрузки 2008

Обзор доклада

Обзор доклада

Что такое вертикальный поиск?

Примеры вертикалей

- Поиск по товарам
- Поиск по подержанным автомобилям
- Поиск по съему жилья

Особенности вертикального поиска

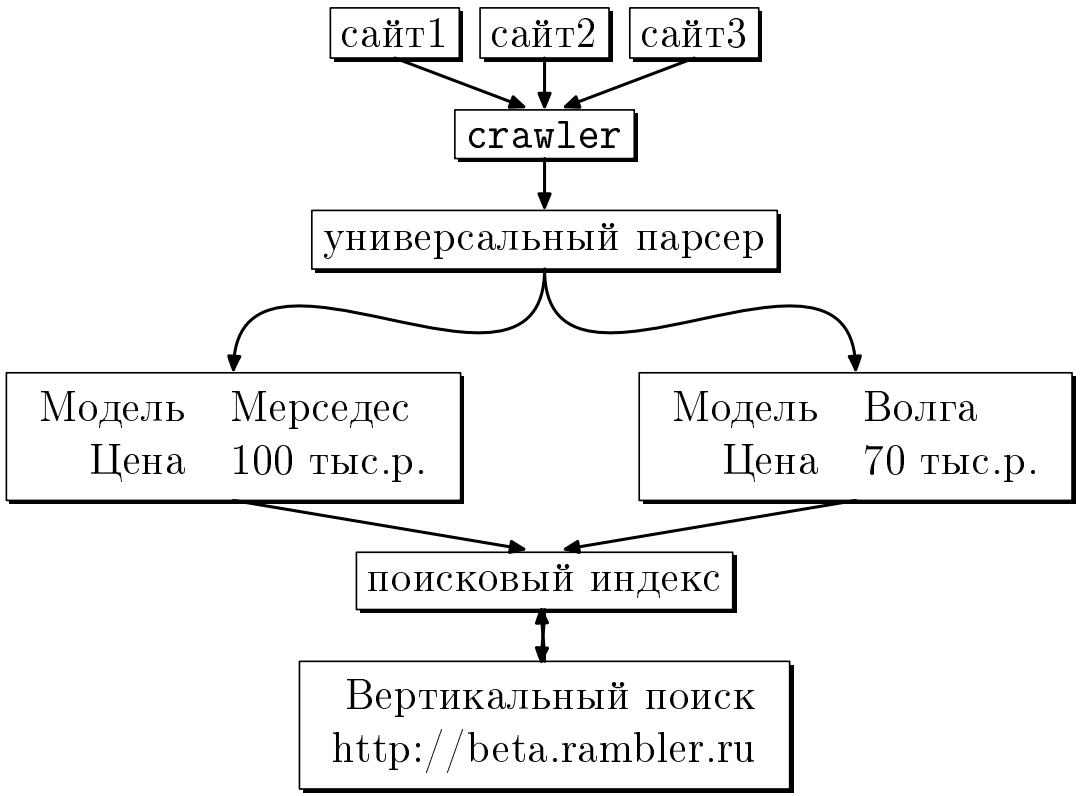
Высокая скорость обновления данных

Для некоторых вертикалей необходимо не реже чем раз в час.

- Обходить и парсить сайты
- Находить новые объявления (возможно среди 100 тыс объявлений)

Не забывать удалять неактуальные объявления (найти их среди 100 тыс объявлений)

Схема работы вертикального поиска



Rambler

Обзор доклада

Универсальный парсер

Универсальный парсер

в общем виде не существует

- Универсальный парсер - это несколько разных парсеров.
- Каждый парсер умеет определять свою успешность
- Если парсер не срабатывает, то следующий

Подходы

Подходы:

- Словарный метод
- Регулярные выражения (шаблоны)
- Комбинированные подходы
- ...

Сейчас нас интересуют только шаблоны

Задачи

Задачи, которые нужно решить

- Как автоматически создавать словари.
- Как автоматически создавать регулярные выражения

Обзор доклада

Классы эквивалентностей

Классы эквивалентностей - КЭ

это объекты, которые строятся без посторонней помощи через анализ массива страниц

КЭ позволяют автоматически создавать шаблоны

Пример класса эквивалентностей

Пример класса эквивалентностей

Рассматриваем сайт из 5 страниц

вектор	токен
1-1-1-1-1	<html>
1-1-1-1-1	<body>
1-1-1-1-1	Модель:
1-1-1-1-1	Цена:
1-1-1-1-1	Город:
1-1-1-1-1	</body>
1-1-1-1-1	</html>

Таблица: Вектор - встречаемость токена на каждой странице

Пример класса эквивалентностей

Токены, вне КЭ

вектор	токен
0-1-0-0-0	Мерседес
0-0-1-0-0	Волга
0-1-0-0-0	100
0-0-1-0-0	200
0-1-1-0-0	руб
1-0-0-1-1	usd
1-1-1-1-0	тыс.

Таблица: Вектор - встречаемость токена на каждой странице

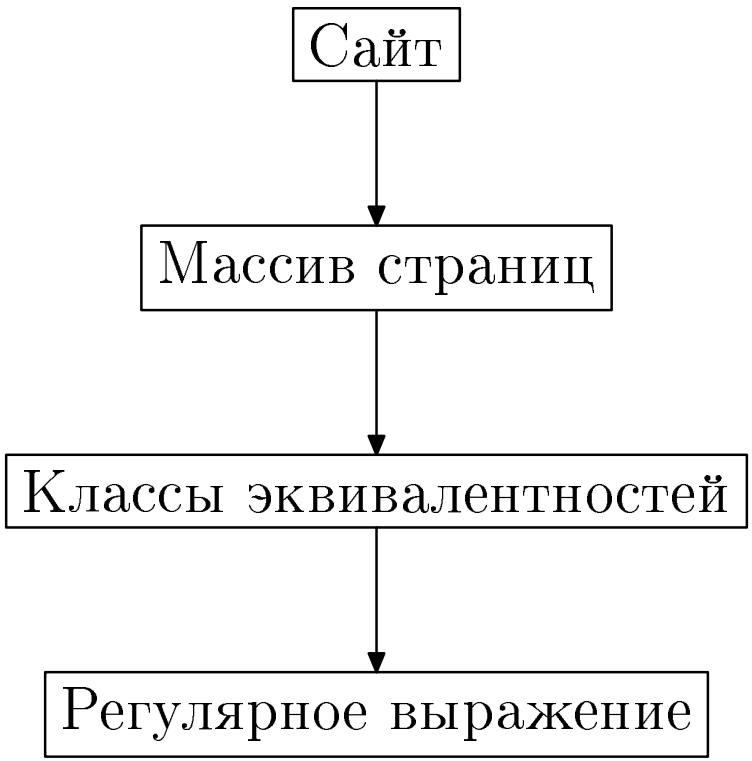
Пример регулярного выражения

Пример регулярного выражения (шаблона)

<body>Модель(.+?)Цена:(.+?)Город:(.+?)</body>

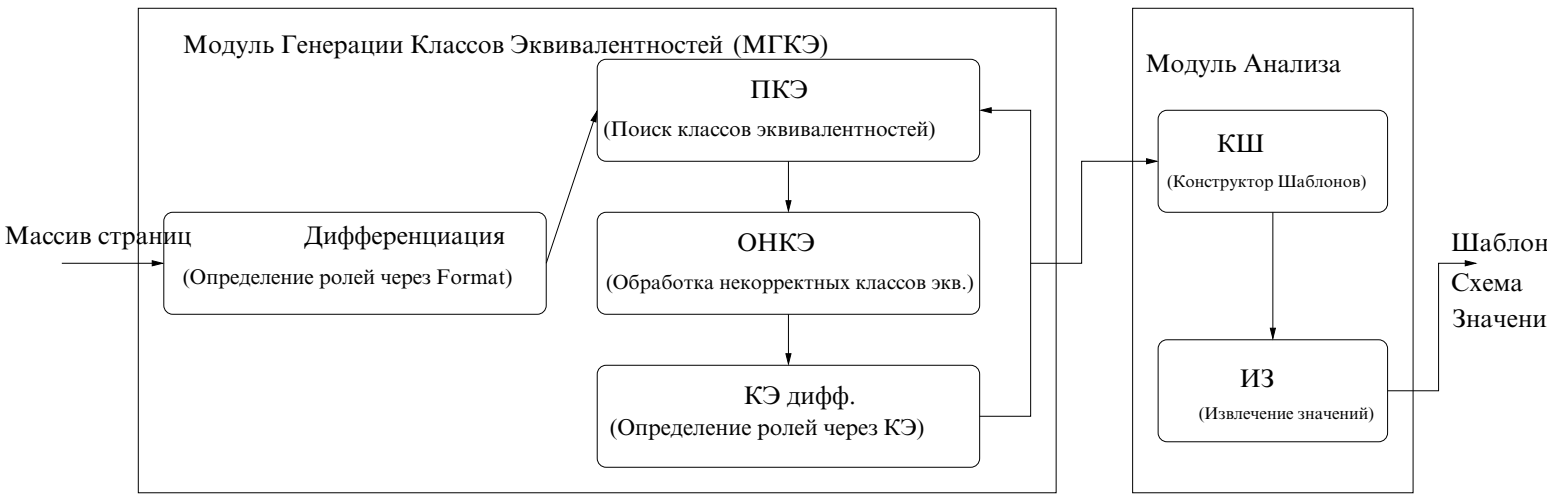
Обзор доклада

Общая схема работы конструктора шаблонов



Конструктор шаблонов

Схема работы КЭ



Что не парсится

Пример сайта, который не парсится

- Стр1: Мерседес * 100 000 руб * Москва
- Стр2: Волга * Новосибирск
- Стр3: Санкт-Петербург * Москвич * Бензин

не за что зацепится

Пример словаря

Примеры словарей

- (Москва|Санкт-Петербург|Новосибирск|Новокузнецк)
- (дизель|Дизтопливо|Газ|Бензин|Бензиновый|бензин)

словари городов и типов двигателей

Словарный метод

Словарный метод

Пример сайта, который парсится словарным методом

- Стр1: Мерседес * 100 000 руб * Москва
- Стр2: Волга * Новосибирск
- Стр3: Санкт-Петербург * Москвич * Бензин

он же не парсится с помощью КЭ

Обзор доклада

Интеллектуальный Маппинг

Маппинг подвыражений

это поиск соответствий между областями значимых данных (подвыражений) к названию атрибута.

- Москва,Санкт-Петербург - это город
- Мерседес,Волга - это марка
- 100 000 руб - это цена

<body>Модель(.+?)Цена:(.+?)Город:(.+?)</body>

Алгоритмы: наивный Байес, SVM

Список сайтов авто-вертикали и успешность КЭ

сайт	атрибутов	КЭ	частично	bad
autochelru	13	9	4	0
autofirmru	9	8	1	0
autonavigatorru	19	15	4	0
autonetru	21	15	6	0
autoriaua	12	10	2	0
autoru	21	16	5	0
autosakhcom	12	7	5	0
autositecomua	15	7	8	0
avby	17	7	10	0
avto-avtoru	16	15	1	0

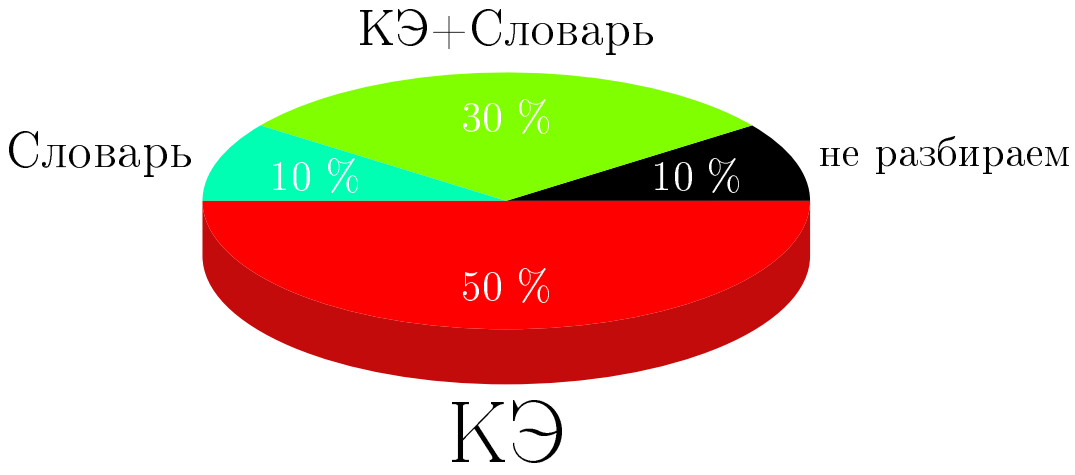
Таблица: Эффективность алгоритма

Оценка

Критерии оценки качества парсеров

- Полнота - сколько % сайтов разбирает
- Полнота - сколько % атрибутов вытягивает
- Точность - сколько % документов разбирает правильно
- За какое время разбирает 100 тыс документов.

Эффективность разбора - % сайтов



Полнота и точность парсеров

Оценки полноты и точности парсеров

метод	полнота %	точность
КЭ	50%	95%
Словарный метод	10%	20-80%
КЭ + Словарный метод	30%	50-80%

Таблица: Общая эффективность: 80-90%

Словарный метод плохо срабатывает на числовых данных

Спасибо! Вопросы?

Спасибо за внимание!

- Протасов Сергей s.protasov.rambler-co.ru
- Плошихин Виктор v.ploshikhin.rambler-co.ru

Вопросы?