

СБОР И ПОИСК СТРУКТУРИРОВАННОЙ ИНФОРМАЦИИ

ВЕРТИКАЛЬНЫЙ ПОИСК

- Ориентирован на определенную область
- Данные хранятся с учетом структуры предметной области
- Позволяет более точно задать запрос
- Возможность фильтрации, сортировки

ПРИМЕР

metafind -- поиск товаров и услуг

http://www.metafind.ru/ metafind

можно искать одной фразой

можно задать время работы

можно отфильтровать по типу аптеки. видна статистика, что в аптеках - 393 наименования, а в аптечных пунктах - 305

можно задать разные фильтры

виден разброс цены

название, расстояние, адрес, телефон, режим работы

указана доп. информация

все сразу видно на карте

все товары сгруппированы по месту продажи

23 июня - дата актуальности цены
24 июня - дата обработки

остальные результаты из данного места

Результатов: примерно 698 (0.90 сек.)

Аптека и лекарства

Уточнить	Цена	Время работы	Тип
Цена	от <input type="text"/> до <input type="text"/>	пн вт ср чт пт	Аптека 393
Время работы	от 17.9 р. до 1070 р.	сб вс праздники	Аптечный пункт 305
Радиус поиска	<input type="text"/>	00:00	
Тип	<input type="text"/>	00 04 08 12 16 20 24	
Скидки	<input type="text"/>	круглосуточно	

Сортировать по цене

- "АСНА Щелковская" ~ 436 метров** - 9-я Парковая ул., 68 корпус 1
с 08:00 до 22:00, тел.: 8-495-468-34-87, 8-499-713-37-93
Тип: Аптека Специализация: Аптека общего профиля, Оп...
[Колдрекс хотрем со вкусом лимона и меда пор д/напитка пак 6г N5x1 Смиткляйн Бичем ИСП - 114 р.](#) [30 Июня]
[Колдрекс хотрем со вкусом лимона пор д/напитка пак 6г N50x1 Смиткляйн Бичем ИСП - 865 р.](#) [30 Июня]
[Колдрекс МаксГрипп со вкусом лимона пор д/оралн р-ра пак N5x1 СКБ/ГСККХ ИСП - 146 р.](#) [30 Июня]
[еще 13 результатов >](#)
- "Линия Жизни на Щелковской" ~ 557 метров** - Уральская ул., 5
с 08:30 до 22:00, тел.: 8-495-462-06-05, 8-495-462-02-03
Тип: Аптека Специализация: Аптека общего профиля
[Колдрекс Юниор Хот Дринк пор д/р-р внутрь детский пак 3г N10x1 СмиткЛБ С.А. ИСП - 191 р.](#) [23 Июня]
[Колдрекс МаксГрипп со вкусом лимона пор д/оралн р-ра пак N5x1 СКБ/ГСККХ ИСП - 125 р.](#) [23 Июня]
[Колдрекс тб уп яч конт спар N12x1 ГлСмКл/ГлСмКлДан ИРД - 65 р.](#) [23 Июня]
[еще 5 результатов >](#)
- "АСНА - Уральская" ~ 886 метров** - Уральская ул., 11
круглосуточно, тел.: 8-499-940-03-73
Тип: Аптека
[Колдрекс хотрем пор д/внутри лимон 5г пак N5x1 ГлСмКл/ГлСмКлБич ИСП - 112 р.](#) [30 Июня]
[Колдрекс тб бл N12x1 Смиткляйн Бичем ВБР - 68 р.](#) [30 Июня]

ПРИМЕРЫ ЗАПРОСОВ

- черная икра в ресторане с караоке около кремля
- гостиница с бассейном около киевской
- комплексная мойка в выхино
- преображенская площадь бассейн 50 метров
- банкомат с долларами рядом с пушкинской

ПОДХОДЫ К СБОРУ ДАННЫХ

Полуавтоматически й

- (+) Универсальность
- (+) Гибкость
- (-) **Временные затраты**
- (-) **Участие человека**

Автоматический

- (-) Не для каждого сайта
- (+) Быстро
- (+) Дешево

ПОЛУАВТОМАТИЧЕСКИЙ СБОР

- Свести к минимуму человеческое участие
- Легкость реализации
- Простота поддержки
- ~~Никаких RegExp, XPath~~

ИДЕЯ

Класс
(предметная область)

+

Шаблон
(специфика сайта)

КЛАСС

- Описывает структуру предметной области
- Похож на ООП класс
- Набор правил, как эту структуру обрабатывать и валидировать, нормализовывать

ПРИМЕР КЛАССА

Товар интернет магазина

Наименование	string
Цена	price
Категории	categories
Наш артикул	reference

ПРИМЕРЫ ТИПОВ

price

1 000,10р. -> 1000.00

1,000,000 рублей -> 1000000.00

address

Пушкинская -> Москва, Пушкинская
площадь; lat: 44.333, lon: 33.112,

ПРИМЕР ШАБЛОНА

```
1  ${url:"shop.ru"}, function ($) {
2      ${link:'каталог'}, function ($) {
3          ${autoPager:true}, function ($) {
4              $.newProduct({
5                  name: 'h3',
6                  price: '::(цена)',
7                  categories: '.breadcrumbs'
8              });
9          });
10     });
11 }
```

Средние временные затраты на 1 шаблон: 10-15 минут

РЕЗУЛЬТАТ РАБОТЫ ШАБЛОНА

- Данные структурированы ,
провалидированы, нормализованы
- Удалены дубли
- Есть diff по сравнению с
предыдущими данными
- Мониторинг «отвалившихся»
шаблонов

ВОЗМОЖНОСТИ ПОИСКА

- Разбор запроса
- Поддержка морфологии и транслита
- Неверная раскладка клавиатуры
- «Возможно, вы имели в виду»
- Фасеты
- Фильтрация, сортировка, группировка
- Гео-поиск

РЕАЛИЗАЦИЯ

- Работает на базе Apache Solr
- Разная структура документа в зависимости от предметной области
- Можно гибко настраивать правила индексации через метаданные класса
- Отдельный индекс для анализа запроса

СПАСИБО!

Минченков Павел
pavel@metahouse.ru
Метахаус