Боты в помощь контент-менеджеру BabyBrick.ru

Руслан Валеев

14.04.2016

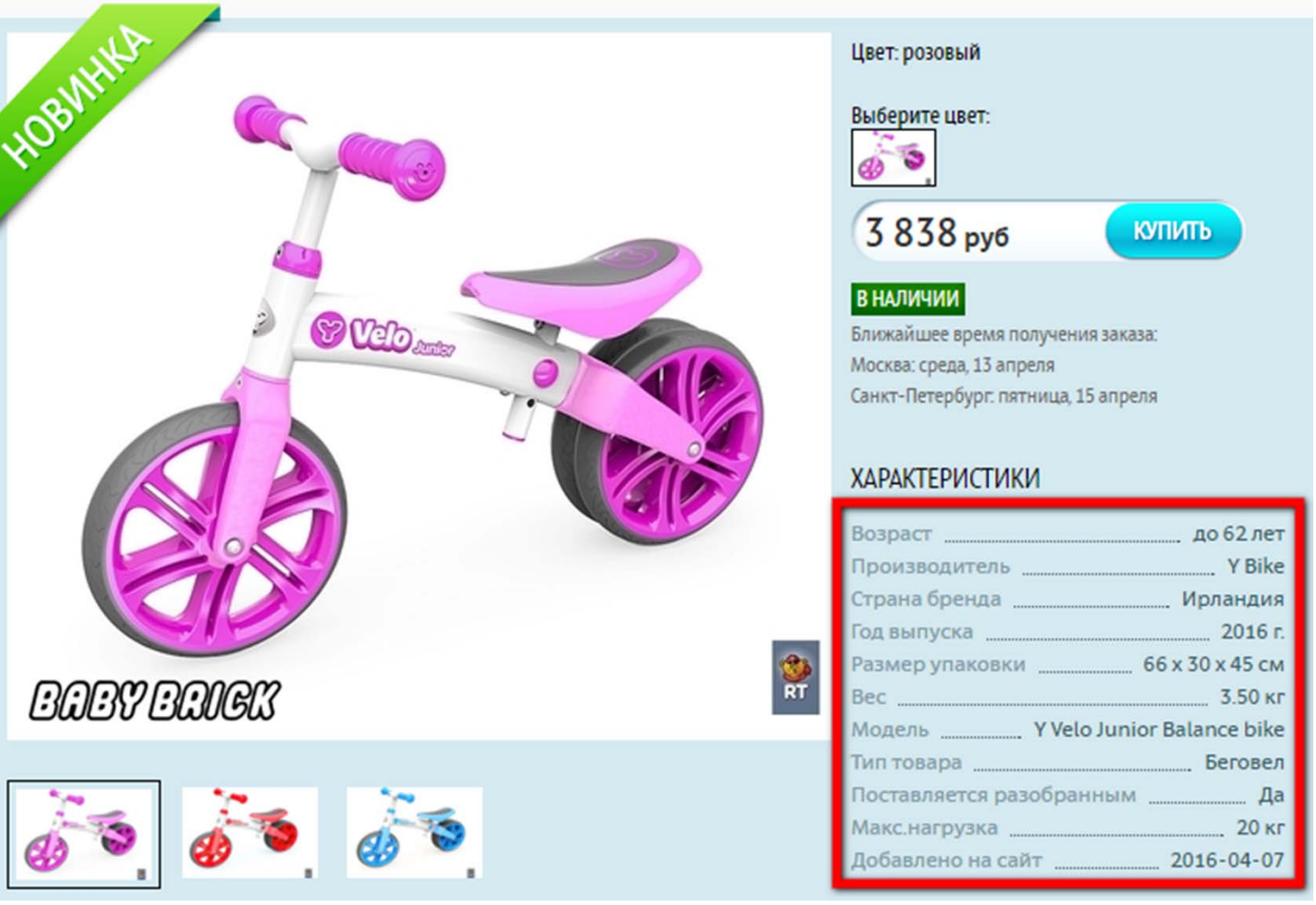
План

- Характеристики товара. Как возникла идея применить боты?
- Какие есть проблемы с автозагрузкой
- Этапы разработки (от php к solr)
- Как это работает (архитектура и технологии, включая прокси)
- Есть ли экономия?
- Как бороться с чужими ботами

Характеристики товара

Toy.BabyBrick.ru → Игрушки для улицы и дома → Беговелы

Велобалансир Y volution Y Velo Junior Balance bike



Цвет: розовый

Выберите цвет:

3 838 руб КУПИТЬ

В НАЛИЧИИ

Ближайшее время получения заказа:
Москва: среда, 13 апреля
Санкт-Петербург: пятница, 15 апреля

ХАРАКТЕРИСТИКИ

Возраст	до 62 лет
Производитель	Y Bike
Страна бренда	Ирландия
Год выпуска	2016 г.
Размер упаковки	66 x 30 x 45 см
Вес	3.50 кг
Модель	Y Velo Junior Balance bike
Тип товара	Беговел
Поставляется разобранным	Да
Макс.нагрузка	20 кг
Добавлено на сайт	2016-04-07

Проблемы с автозагрузкой

- Каждый сайт имеет свою уникальную структуру
- Структура сайтов со временем меняется
- Сложность поиска товара по названию и артикулу (на других сайтах товар может иметь другой артикул)
- Различное написание названий производителей
- Цена и наличие зависит от региона (для некоторых магазинов)
- Нет списка атрибутов товаров, которые доступны на конкретном сайте
- Не всегда удается правильно вычленить нужную характеристику из текста
- Фотографии бывают с водяными знаками
- Блокировка доступа к сайтам при агрессивном скачивании
- Некоторые сайты практически не возможно выкачать (amazon.com)

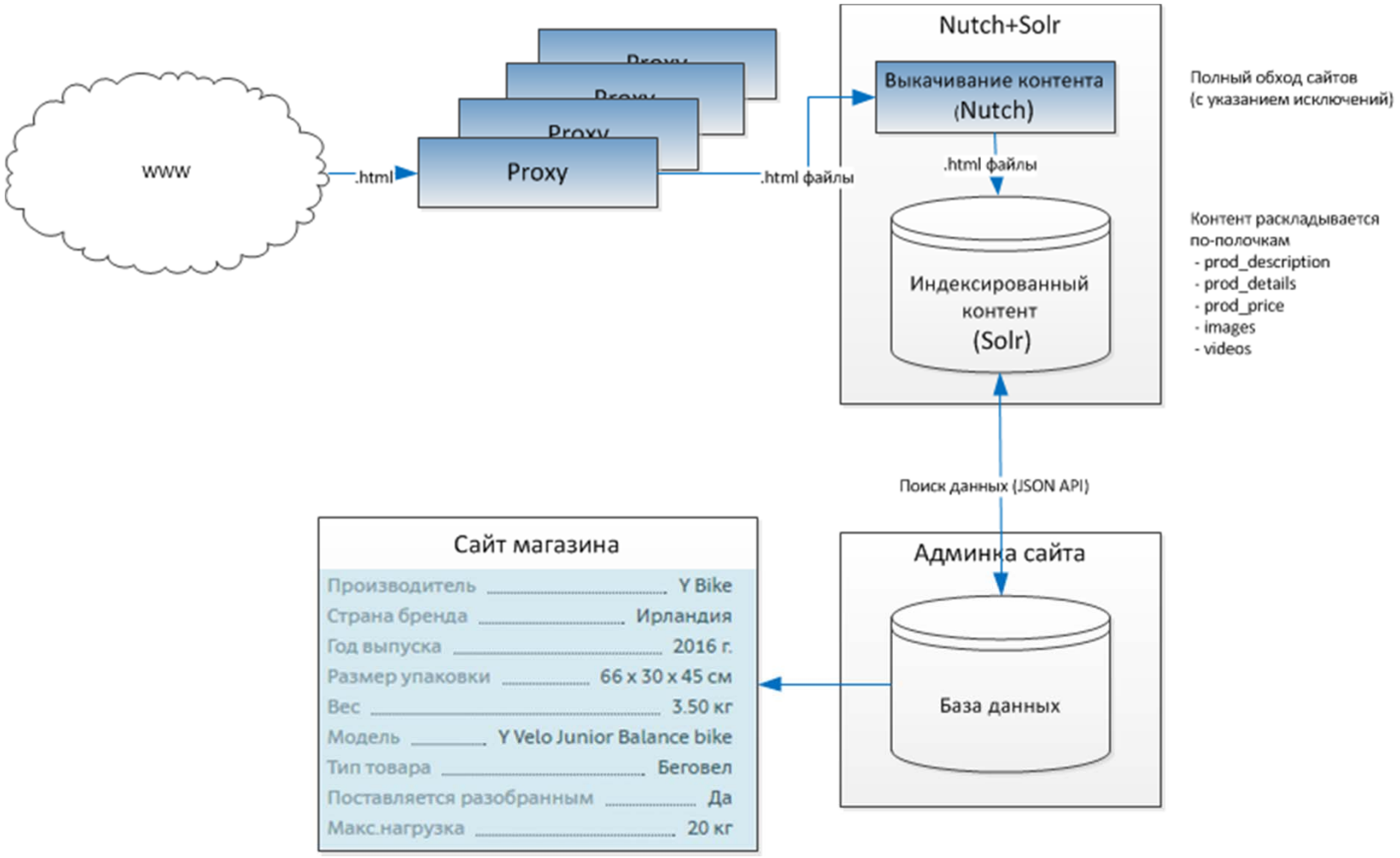
Допущения и ограничения

- Выкачиваем только сайты с “нашими” артикулами и “нашими” названиями производителей
- При таком подходе невозможно быть №1 на рынке, но можно быть №2 ☺
- Необходимо глазами просматривать и вручную подчищать контент

Эволюция разработки

- Версия 1
  - Выкачивание контента сайта конкурентов с помощью PHP и анализ содержимого
  - Для каждого товара в базе прописывался URL на товар конкурента, чтобы можно было быстро узнать цену
  - Помогает только выгрузить характеристики и цену
  - Обновление контента вручную
- Версия 2
  - Система ботов на Java с ипользованием системы Nutch+Solr с открытым кодом
  - Выгрузка характеристик, фото, цен, описаний и пр.
  - Автоматическое добавление товаров на сайт
  - Автоматическое обновление характеристик товаров
- Версия 3
  - Тоже что и версия 2, но перед добавлением и обновлением товаров просматриваем контент глазами

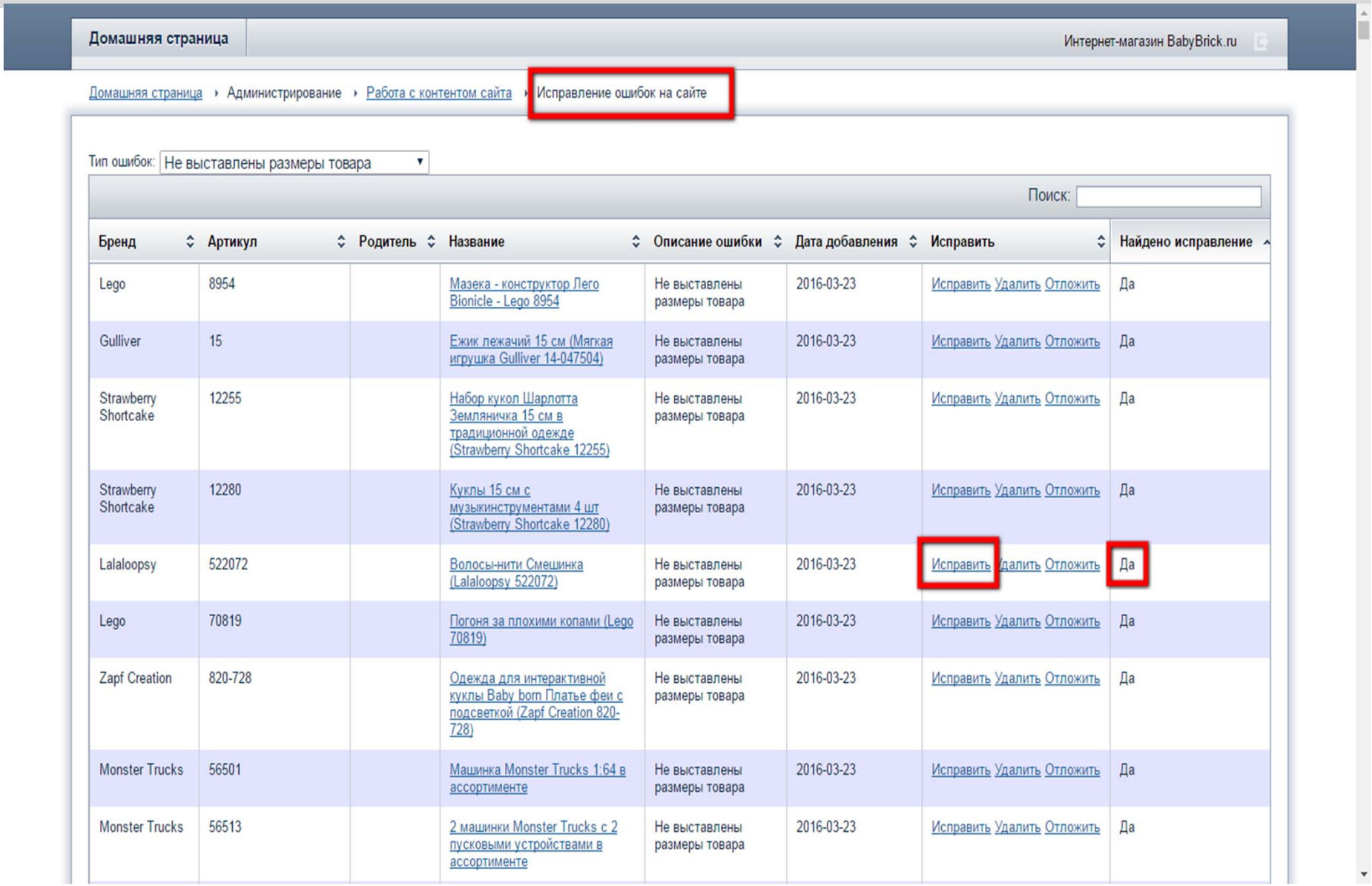
Архитектура



Как бороться с чужими ботами

- 99% используемых ботов не содержат движок JavaScript
- Решение:
  - Через Javascript выставлять шифрованное cookie
  - На стороне сервера проверять, есть cookie или нет
  - Не показывать контент при отсутствии cookie
- Что делать с хорошими ботами поисковых систем?
  - Разрешить доступ на уровне диапазонов IP адресов
  - Вести базу ботов помеченных флагами “разрешено” или “запрещено”
  - Если при обращении к контенту бота в базе нет, то по IP адресу сделать reverse DNS и посмотреть доменное имя. Разрешить доступ в случае “правильного” доменного имени и сохранить в базу.
- Что делать с оставшимся 1%? Варианты
  - Ничего не делать
  - Анализировать как запрашивается контент и если он запрашивается не так как это делает обычный пользователь – в бан

Демо: отчет об отсутствующих атрибутах



Бренд	Артикул	Родитель	Название	Описание ошибки	Дата добавления	Исправить	Найдено исправление
Lego	8954		Мазека - конструктор Лего Bionicle - Lego 8954	Не выставлены размеры товара	2016-03-23	Исправить Удалить Отложить	Да
Gulliver	15		Ежик лежачий 15 см (Мягкая игрушка Gulliver 14-047504)	Не выставлены размеры товара	2016-03-23	Исправить Удалить Отложить	Да
Strawberry Shortcake	12255		Набор кукол Шарлотта Земляничка 15 см в традиционной одежде (Strawberry Shortcake 12255)	Не выставлены размеры товара	2016-03-23	Исправить Удалить Отложить	Да
Strawberry Shortcake	12280		Куклы 15 см с музыкинструментами 4 шт (Strawberry Shortcake 12280)	Не выставлены размеры товара	2016-03-23	Исправить Удалить Отложить	Да
Lalaloopsy	522072		Волосы-нити Смешинка (Lalaloopsy 522072)	Не выставлены размеры товара	2016-03-23	Исправить Удалить Отложить	Да
Lego	70819		Погоня за плохими копами (Lego 70819)	Не выставлены размеры товара	2016-03-23	Исправить Удалить Отложить	Да
Zapf Creation	820-728		Одежда для интерактивной куклы Baby born Платье феи с подсветкой (Zapf Creation 820-728)	Не выставлены размеры товара	2016-03-23	Исправить Удалить Отложить	Да
Monster Trucks	56501		Машинка Monster Trucks 1:64 в ассортименте	Не выставлены размеры товара	2016-03-23	Исправить Удалить Отложить	Да
Monster Trucks	56513		2 машинки Monster Trucks с 2 пусковыми устройствами в ассортименте	Не выставлены размеры товара	2016-03-23	Исправить Удалить Отложить	Да

Размеры товара

www.toy.ru
dimensions[0] 360 dimensions[1] 130 dimensions[2] 380 Сохранить

Бренд Lalaloopsy Артикул 522072 522072 Поиск Добавить вручную

Найденные ошибки в товарах

Поиск:

Сайт	H1	Title	Артикул	Описание	Характеристики	Картинки	Перейти на сайт
http://www.toy.ru	*Lalaloopsy 522072 Лалалупси* Волосы-нити, Смешинка Добавить отзыв	Купить *Lalaloopsy 522072 Лалалупси* Волосы-нити, Смешинка в интернет-магазине Toy.ru	*522072*		Характеристики Рекомендуемый возраст – 4-5 лет, 6-8 лет Пол: для девочек Бренд: *Lalaloopsy* Артикул: *522072* Размер упаковки: 0.36 x 0.13 x 0.38 м Вес: 1.21 кг Посмотреть все товары раздела Куклы и аксессуары Посмотреть сопутствующие товары	 	На сайт

Записи с 1 до 1 из 1 записей

Monster Trucks	56501		Машинка Monster Trucks 1:64 в ассортименте	Не выставлены размеры товара	2016-03-23	Исправить Удалить Отложить	Да
Monster Trucks	56513		2 машинки Monster Trucks с 2	Не выставлены	2016-03-23	Исправить Удалить Отложить	Да

Вопросы

- ?