Технология «Спектр» Яндекса и классификация веб-страниц

Трофименко Евгений

контакты info@promosite.ru

услуги http://promosite.ru/

seo-сервисы http://tools.promosite.ru/

Как угадать намерения пользователя?

раньше

Василий Иваныч: - Петька! Приборы!

Петька: Двадцать!

В.И.: - Что двадцать?!!

П.: А что приборы?!!

теперь

Василий Иваныч: - СПЕКТР! Приборы!

СПЕКТР.: Вы хотите:

- Купить приборы?
- Смотреть фото приборов?
- Почитать отзывы о приборах?
- Ударить приборы ногой?
- Снять показания приборов?

В. И.: - Два наряда вне очереди!

«спектр»: намерения пользователя

http://clubs.ya.ru/company/replies.xml?item_no=32028

a-plakhov:

Когда пользователи задают запросы к Яндексу, примерно в 20% случаев они **формулируют запрос неоднозначно.** Например, по запросу [наполеон] кто-то хочет найти полководца, а кто-то – рецепт торта. А задавая запрос [суши], человек может искать и ресторан с доставкой на дом, и рецепт блюда...

...мы внедрили новую поисковую технологию, которая умеет **учитывать множество неявных целей пользователей** и показывать соответствующие ответы...

В основе работы «Спектра» лежит статистика поисковых запросов. Система исследует запросы всех пользователей Яндекса и **выделяет в них различные объекты**...

Кроме того, «Спектр» умеет учитывать при поиске **различные потребности пользователей**. У каждой категории есть список возможных потребностей – тех намерений, с которыми пользователи ищут тот или иной объект. Например, когда люди ищут какой-нибудь товар, они, как правило, хотят купить его или почитать отзывы и обзоры. То есть для категории «товары» среди потребностей будут «купить», «отзывы» и «обзоры»...

«Спектр» анализирует поисковые запросы полностью **автоматически**...

как работал* «спектр» в начале

классификация всего найденного?

(*) подсвечивал?

Пример: запрос «**ноутбуки**». Ввод СПЕКТРа, зима 2010.

Виды тематик найденных результатов:

1. Тема страницы: «новые, купить»

Подсвечены в сниппете: продажа, цена, купить, каталог, новые, т.п.

2. Тема страницы: «б/у»

Подсвечены в сниппете: б/у, подержанные, т.п.

НЕ Подсвечены: продажа, цена, ремонт

3. Тема страницы: «ремонт»

Подсвечены в сниппете: ремонт, т.п.

НЕ Подсвечены: продажа, цена, б/у

Тематики раздельные, доп. слова не пересекаются.

как «спектр» работает теперь

подмешивание отдельных рез-тов

* запросы настраиваются отдельно!

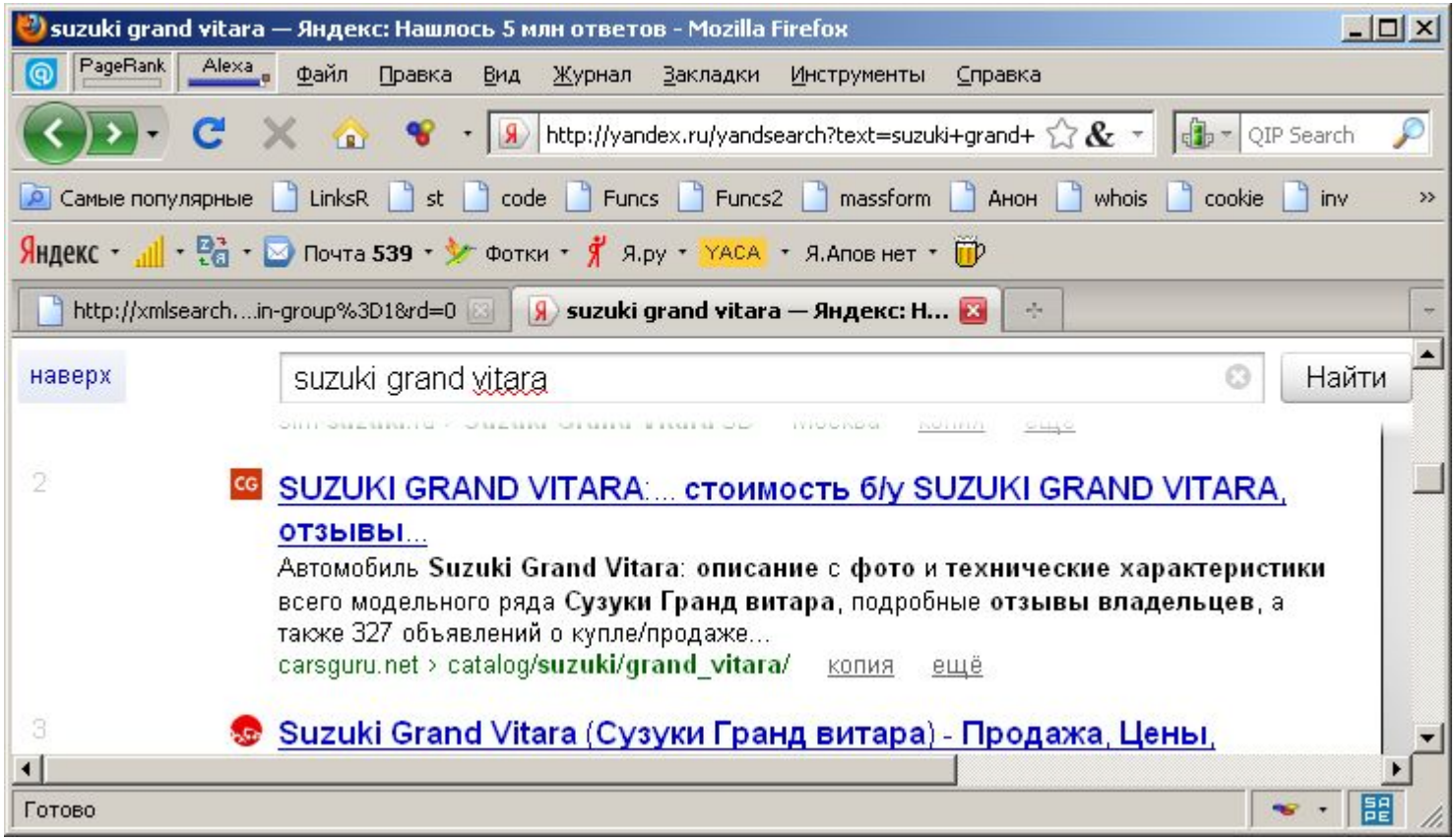
Для запроса выделяется список доп. интентов (~намерений, тематик), в выдачу добавляются результаты из отдельной базы классифицированных страниц (по соответствию тематике).

- Подсветка «спектровых» слов в «обычных» результатах
- Подсветка всех «спектровых» слов независимо от интентов (тем)
- Подмешивание – это костыль. «Автофургоны» забивают выдачу.
- Бывает несколько примесных результатов по одному интенту.

(*) еще один признак костыля

Евгений Трофименко info@promosite.ru

Теперь – подсветка всего:



6 **Отзывы о Suzuki Grand Vitara (Сузуки Гранд Витара) | Отзывы...**
NewMPV.ru – это самая полная база отзывов владельцев **Suzuki Grand Vitara**. ... **Сузуки Гранд Витара** - хорошая машина. Это основной вывод. Брал новую.
newmpv.ru › **Suzuki Grand Vitara** копия

«сузуки гранд витара» — 1 936 видеороликов


Suzuki Grand Vitara V6 - Сузуки 14:59


2011 **Suzuki Grand Vitara** / Тест 17:42


2011 **Suzuki Grand Vitara** / Тест 18:50

video.yandex.ru

7 **Suzuki Grand Vitara** -... и **подержанные** автомобили **Сузуки Гранд Витара**...
Контакты, сервис и обслуживание **Сузуки** у официальных дилеров. **Продажа** автомобилей **Suzuki Grand Vitara (Сузуки Гранд Витара)**.
acars.ru › **Suzuki** › **Grand Vitara** копия ещё

8 **Форум Suzuki Vitara**, Escudo, **Grand Vitara**, **Grand** Escudo - Powered...
Обзор прессы о **Suzuki Grand Vitara**. **Гранд Витара** в прессе. от Benjamin. ... (**отзывы**, качество обслуживания, да вообще любые комменты) Сим-**сузуки**. от Boeing.
club-**vitara**.ru копия

9 **Отзывы владельцев Suzuki Grand Vitara (Сузуки Гранд Витара)**...
Дополнительные **мнения** о **Suzuki Grand Vitara (Сузуки Гранд Витара)** → Режим просмотра

как отличить спектровую примесь?

По идентификатору документа в XML-выдаче.

Обычный документ: 4 фрагмента
<doc id="**49**-0-16-ZA21FA0474B79859A">

СПЕКТР: 3 фрагмента
<doc id="**52**-115-Z7725D3069AAE1668">

Быстроробот: 3 фрагмента
<doc id="**53**-66-Z6AF572834514019F">

Ультраробот (сейчас нет): 2 фрагмента
<doc id="55-Z7725D3069AAE1668“>

Евгений Трофименко info@promosite.ru

подсветка спектровых слов, ограничения и как их обойти

Теперь все дополнительные слова подсвечиваются в выдаче независимо от классификации.

Подсветка спектровых слов идет и на спектровой примеси, и на обычных документах. Однако для продвижения …

Ограничение – подсветка (как и примесь) идет только по первой десятке (не обходится увеличением numdoc)

Ограничение обходится поиском внутри сайта (ограничением параметрами serverurl, surl и перебором большого количества путей внутри сайта). Так можно взять подсвеченные спектром в сниппетах слова по всем страницам сайта.

Евгений Трофименко info@promosite.ru

база пробивки «спектр»а: 6.5M запросов, лето 2011

Для 5% запросов есть спектровые примеси (325K из 6.5M)
А один ли результат в примеси? ☺

примесных результатов	запросов	%
1	231383	3.56%
2	71855	1.11%
3	18301	0.28%
4	3486	0.05%
5	573	0.01%
6	69	
7	20	
8	5	
9	7	
10	5	

Евгений Трофименко info@promosite.ru

Большое число спектровых результатов в десятке для…

Особенно непонятных запросов.
«Петька! Приборы! -двадцать!»

запрос	спектровых результатов в топ10
я	10
о войне 1941 1945	10
перми	9
казани	9
новосибирска	9
одессы	9
харькова	9
виктора цоя	9
про собак	8
феодосии	8
из фильмов	8
волгограда	8

доп. тематики «спектр»а запрос [**казани**] – 7 из 10

1 **Карта Казани**
2 **Достопримечательности Казани** -... Татарстана. Туристу о **Казани**, Елабуге...
3 **Новости** | Казанский Портал
4 **Гостиницы в Казани**
5 Город **Казань**
6 2ГИС — **карта Казани** с улицами и домами, справочник организаций
7 **Карта Казани**
8 **Недвижимость в Казани** из рук в руки: объявления о продаже...
9 Журнал КАЗАНСКАЯ **НЕДВИЖИМОСТЬ**:... офисов в городе **Казани**...
10 GISMETEO.RU: **Погода в Казани** на сегодня, завтра. Прогноз погоды...

Евгений Трофименко info@promosite.ru

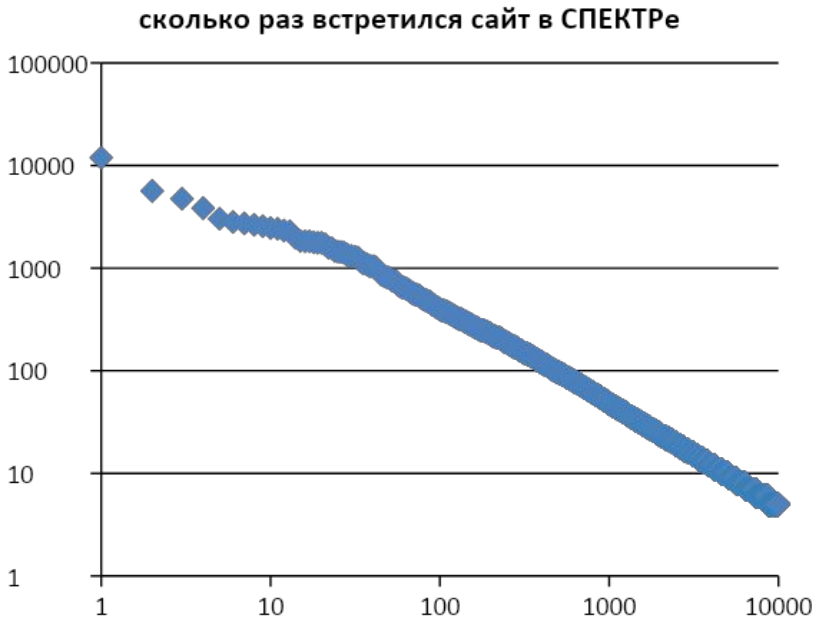
А вот запрос [**казань**] «спектр»а нет вообще…

1 Город **Казань**
2 **Казань** — Википедия
3 Туристический портал г.**Казань** - города с тысячелетней историей
4 VIP **Казань — Казань** для достойных людей
5 Город **Казань** - Портал **Казань** 24
6 **Казань**: ruKazan - сайт города **Казани**, клубы, **работа, вакансии**, магазины...
7 **Казань**
8 **Казань** по-новому. **Новости**, квартиры, **работа, погода**, объявления...
9 **Казань**. Информационный портал города - Главная
10 Моя **Казань** - информационно-развлекательный портал города

самые частые сайты, которые попадают в примесь

Некоторым везет больше.
Некоторым (сова72, автофургоны) везет временно.

сколько раз встретился сайт в СПЕКТРе	сайт
11941	www.torrentino.com
5657	www.zaycev.net
4750	www.fast-torrent.ru
3845	nnm-club.ru
3024	www.slovopedia.com
2811	bigtorrent.org
2725	tfile.ru
2652	musicmp3.spb.ru
2561	best-mp3.ru
2460	www.rutor.org



«Узнать или купить?

Классификатор страниц обзоров и интернет-магазинов»

http://www.dialog-21.ru/dialog2011/materials/pdf/17.pdf

Braslavski P. I., Yandex, Kiselev Yu. A., Ural Federal University

Решается похожая проблема – выяснение намерения пользователя.

Shop classifier

Term features. We identifi ed the most informative term-features based on mutual Information … As expected, **the most contrasting terms were *магазин, рубль, каталог, цена, прайс, and корзина*** *…The full list of terms used* for classifi cation **consisted of about one hundred terms**.

Lexical features. We used the list of trademarks and brands

Review classifier

Term features. .. lexical variety of reviews is much higher than that of shop pages, the list of contrasting words was **much longer and exceeded 7,000 words**.

*Lexical features**. The list of 165 manually collected appraisal adjectives —**хороший, прекрасный, великолепный, плохой, отвратительный, ужасный, etc.* (*good, excellent, magnifi cent, bad, disgusting, awful, etc.)*

«Тематические» слова

доп. слово	встретилось
скачать	133398
mp3	40354
2011	27195
онлайн	20833
отзывы	20781
торрент	16200
смотреть	15964
фото	15598
2	14535
игры	14087
перевод	12859
бесплатно	11321
карта	10222
ооо	10136
инструкция	8773
фильмы	6894
аккорды	6859

Самые частотные тематики – фильмы, торренты, mp3

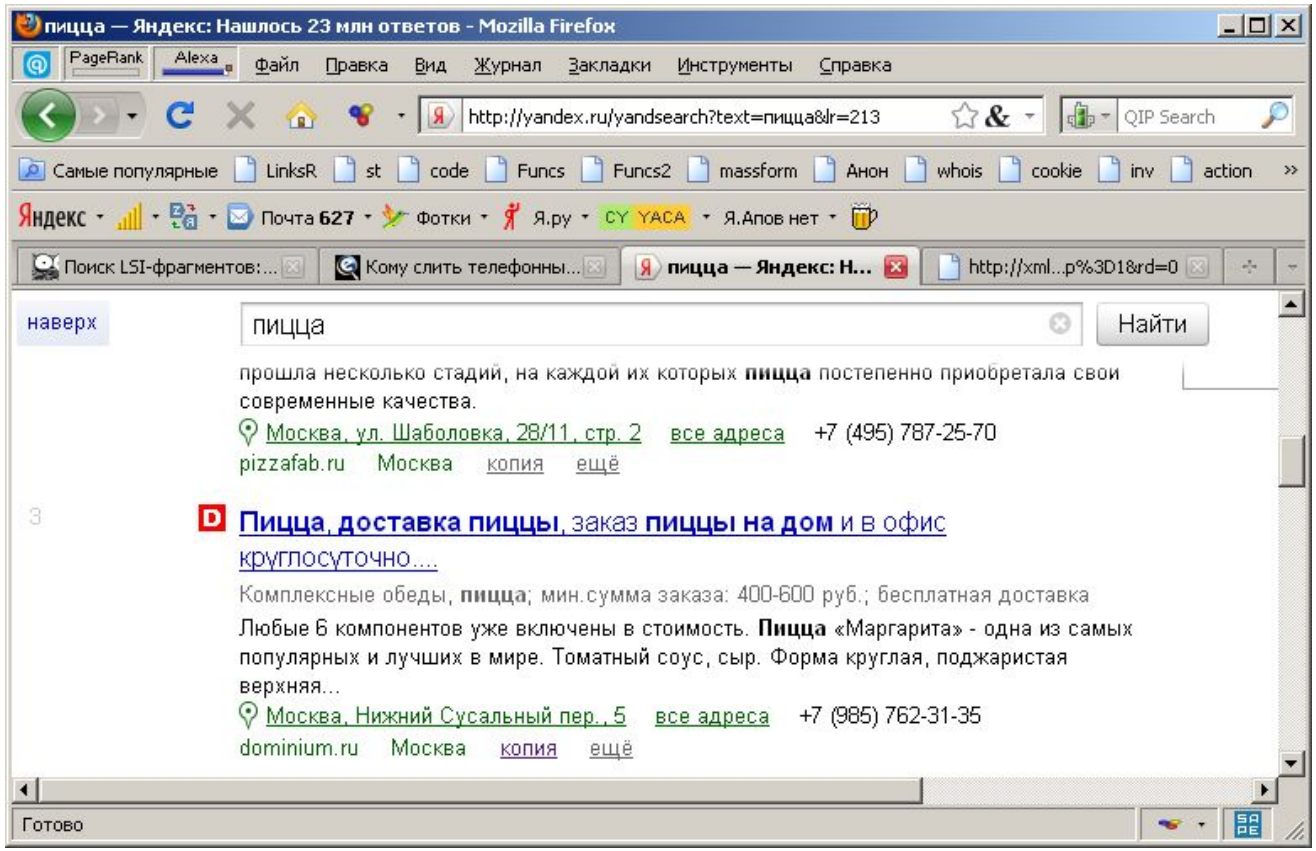
«Новизна» (2011), отзывы, карты, т.д.

… Это верхушка.
Но для классификации используются даже такие служебные слова, как **сайт** и **меню.**

… и многословные фрагменты

Выделение фрагментов из нескольких слов целиком: **на дом**

Только классификация СПЕКТРом или учет в ранжировании?



многословные фрагменты в «спектр»е

Встречались в 20-25% спектровых примесей (общим числом 83К)

А уникальных: 127 штук!

Очевидно, они сделаны вручную.

многословный фрагмент	встречался	%
что такое	21773	26.1%
смотреть онлайн	17034	20.4%
текст песни	10970	13.1%
своими руками	9809	11.7%
в домашних условиях	4062	4.9%
прогноз погоды	2639	3.2%
отзывы владельцев	2324	2.8%
слова песни	2049	2.5%
тексты песен	1862	2.2%
скачать драйвера	1001	1.2%
на карте	992	1.2%
технические характеристики	970	1.2%
онлайн смотреть	899	1.1%
краткое содержание	741	0.9%
карта города	681	0.8%
скачать драйвер	634	0.8%

Евгений Трофименко info@promosite.ru

Что делать?

1. Качественный сайт. (*) Некоторые сайты подмешиваются чаще.

(*) это слишком сложно!!1

2. Выяснять классификационные слова и многословные фрагменты для ваших запросов (запросы м.б. на разные тематики)

3. Не стесняться их употреблять в тексте.

4. Польза не при ранжировании, а при классификации и подмешивании.

Евгений Трофименко info@promosite.ru

FIN.

Технология «Спектр» Яндекса и классификация веб-страниц

Трофименко Евгений

контакты info@promosite.ru

услуги http://promosite.ru/

seo-сервисы http://tools.promosite.ru/