Эволюция алгоритмов Яндекса и методов исследований: новые возможности анализа

Трофименко Евгений

сЭо-эксперт

info@promosite.ru

http://tools.promosite.ru/

Краткое содержание

- **Индексация рунета** (и так уже много страниц – пора выкидывать)
- **Апдейты** (текстовые, ссылочные, гео, т.д. – пересчет параметров)
- **Взвешивание ссылок** (влияют ли ссылки с АГС и баненных сайтов?)
- **Переформулировки поисковых запросов Яндексом**
  (расстояния, веса, словоформы, новые операторы, зоны и термы)
- **Контрастности слов, «веса»** (НЕ IDF, три разных базы – три веса)
- **Численные значения релевантности** - предварительные итоги
  (учет ДМОЗ, классы запросов, рел-ть группы и элемента разные)

Я: Что интересного за год:

•Почти ровно год назад, сентябрь 2009 – выстрелил **фильтр АГС-17**

•Осень 2009 – алгоритм обучения формулы ранжирования **Matrixnet**

– **оператор «минус»** стал применяться к текстам ссылок, теперь **не находит НПС**

- **отмена неранжирующего И (<<)**, изменения в языке запросов

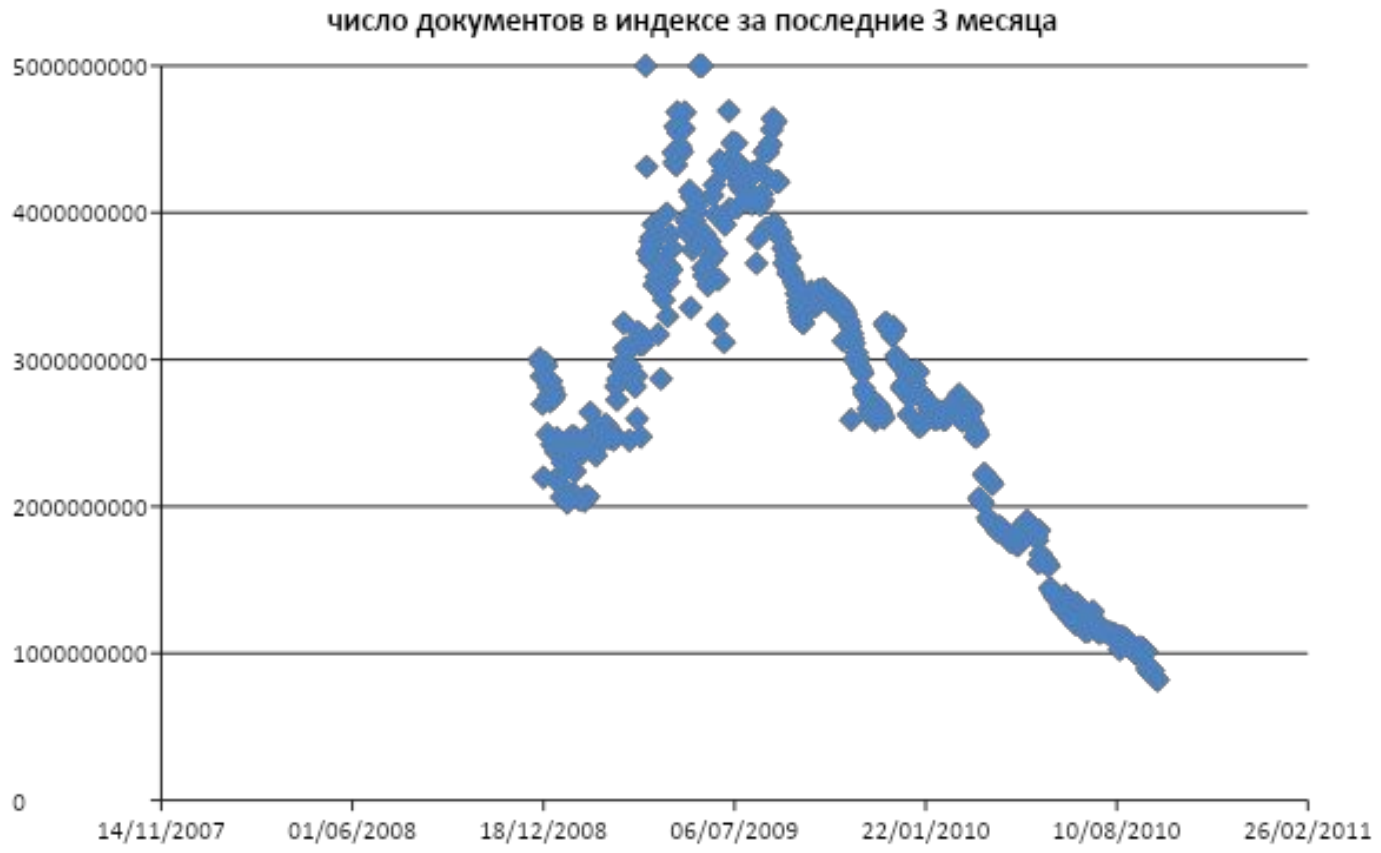
*** не особо известное:

•Лето-Осень 2009 – отмена показа числа страниц «еще с сайта». Число страниц в «еще с сайта» сейчас отличается от общего. Имхо = **введение «яндекс-соплей»**, летом была перезагрузка.

•Яндекс занимается экстракцией фактов в большом поиске - **новые поисковые зоны документа и термы**, которые соответствуют ФИО

Я: число документов в индексе

Число документов (сумма по дням индексации за последние 3 месяца)
уменьшилось за год в 4-5 раз (было 4500M, стало 800M)

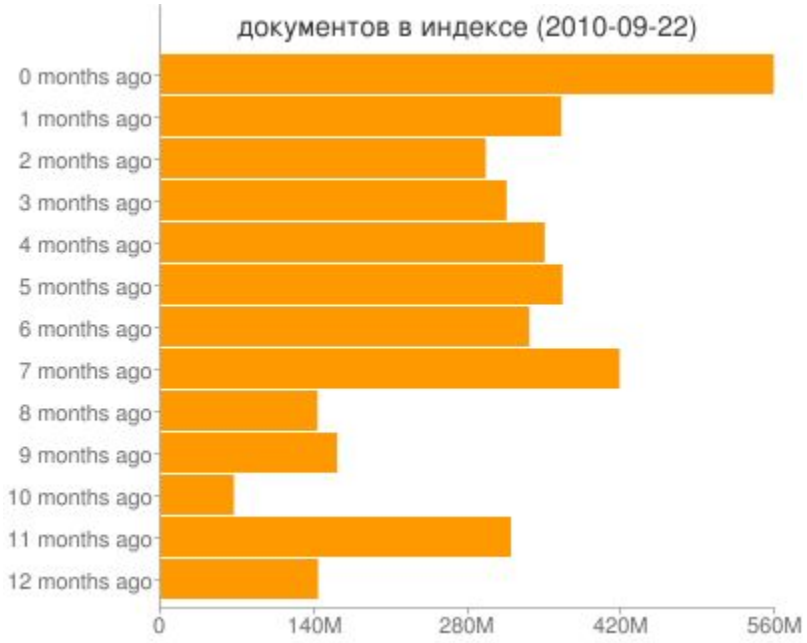


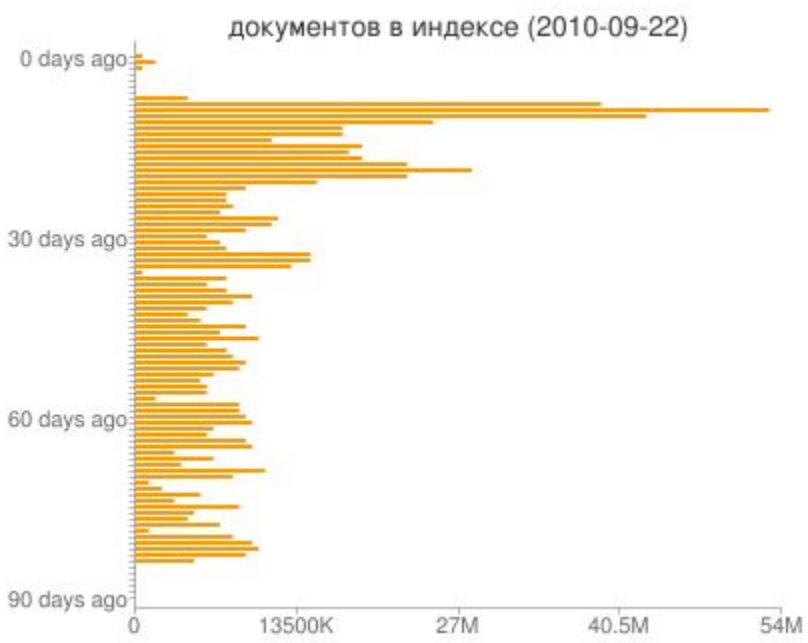
Евгений Трофименко info@promosite.ru

Я: скорость индексации Рунета

Скорость переиндексации рунета **уменьшилась за год в три раза**:
Было ~50-60 дней, стало ~150-170 дней

Метод: операторами дат ищем число документов за интервалы и взвешиваем число документов по дате (ищем «центр масс»)





Евгений Трофименко info@promosite.ru

2. Апдейты Яндекса:

русский и западный индексы

Выкладывание новых проиндексированных страниц, появление в поиске текстов страниц.

Метод: увеличение числа страниц и сайтов поиском **date:YYYYMMDD**

Как отличаем русский от западного: **lang:ru** (uk) и **lang:en** (de, fr)
Обычно западный индекс апдейтится раньше.

Примерное время: около часа ночи, раз в 3 дня

Апометр Дениса Иванова – упорядочение по дате.
Быстроробот мешает.

Важно:
новые страницы и обновления старых не видны раньше апдейта

Апдейты Яндекса:

сохраненной копии (метод komdir)

Выкладывание новой сохраненной копии происходит чуть раньше.

На ~10-15 минут раньше.

Проблемы: иногда сохраненная копия берется «на лету» и кажется слишком свежей.

Редко, но бывает: обновляется на день позже.

Важно:
Проверка ссылок в сохраненных копиях страниц не даст эффекта, если ее обновления не произошло.

Апдейты Яндекса: ссылочные

новые ссылки в анкор-файле

Обновление анкор-файла, учет новых проиндексированных ссылок.

Метод:
поиск свежих не-быстророботных-НПС «найденных по ссылке»

Экспериментальные сайты и быстроиндексируемые ссылки, время взятия ссылки роботом написано сразу в НПС

Время: около 4-5 утра
Частота – в последние дни каждый текстовый апдейт, раньше – раз в три недели.

***** важно:**
пересчет «веса» ссылок может происходить независимо от обновления анкор-файла.

Евгений Трофименко info@promosite.ru

Апдейты Яндекса: гео

Изменение числа геопривязанных сайтов

Число сайтов и страниц, найденных с ограничением по региону меняется резко и не всегда по всем регионам.

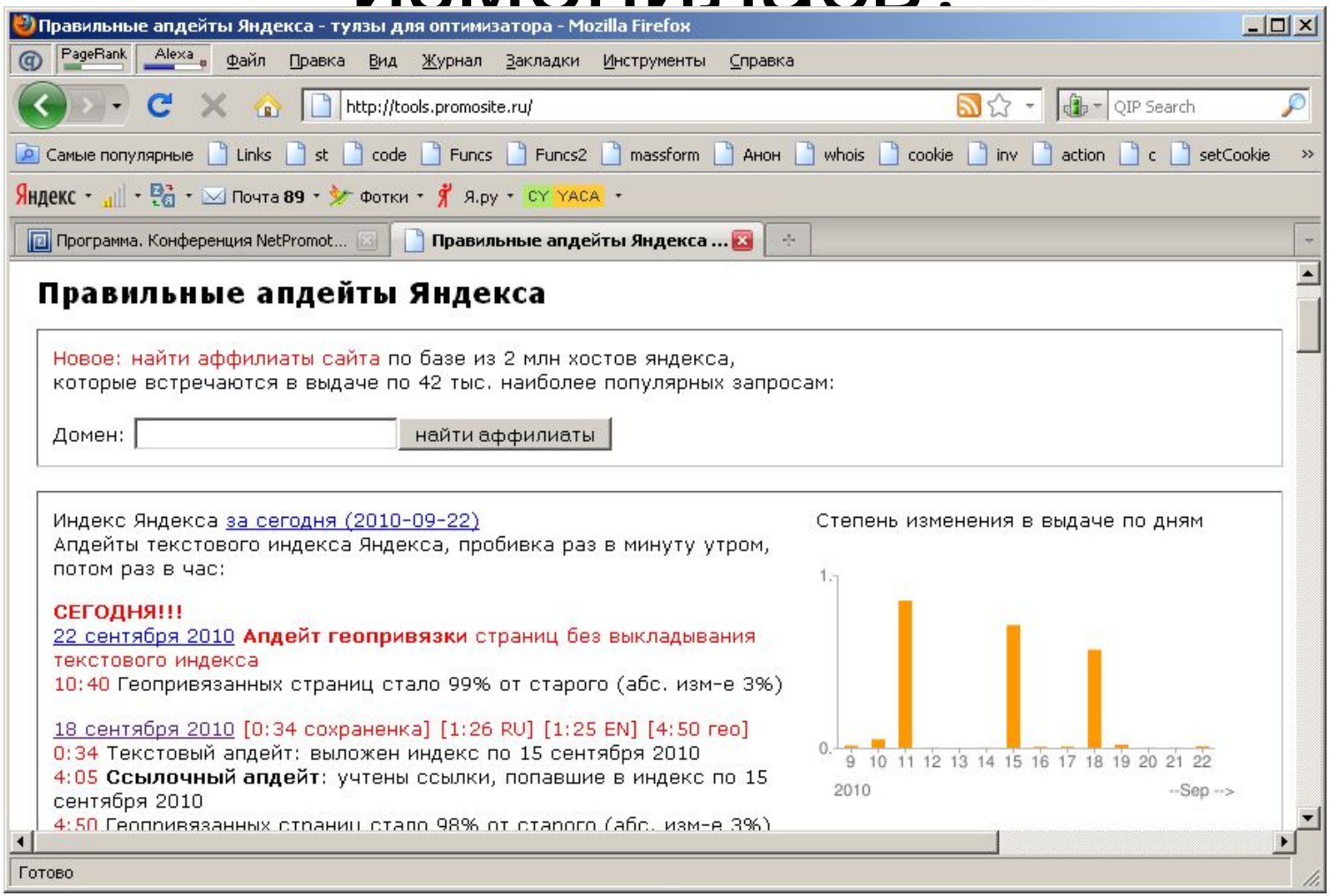
Метод: поиск с ограничением по регионам **&rstr=-213** (11 регионов) и с ограничением по хостам для уменьшения числа найденных сайтов.

Число сайтов и страниц не только увеличивается, но и уменьшается.

Время: около 5 утра, но иногда и в середине и в конце дня.

За половину изменений выдачи без выкладывания текстового индекса - отвечают в том числе и они.

Апов нет, а выдача изменилась?



Евгений Трофименко info@promosite.ru

= обновление формулы?

…или многое другое:
например,

пересчет весов ссылок?

1. Есть запросы, где в результатах много НПС
2. Положение НПС относительно друг друга меняется около 4 утра и без ссылочных апдейтов.
3. Видимо, это пересчет вИЦ и релевантности ссылок

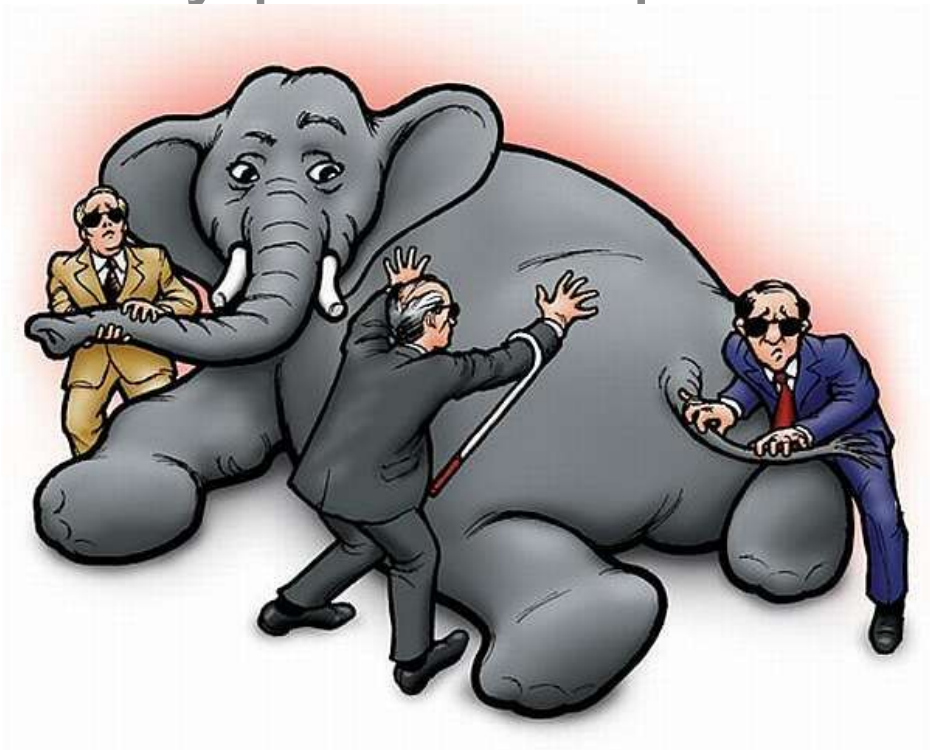
… это был анонс по сервису http://tools.promosite.ru/

А значат ли позиции НПС хоть что-нибудь?

Евгений Трофименко info@promosite.ru

3. НПС и взвешивание ссылок

…кстати - суфийская притча о слоне

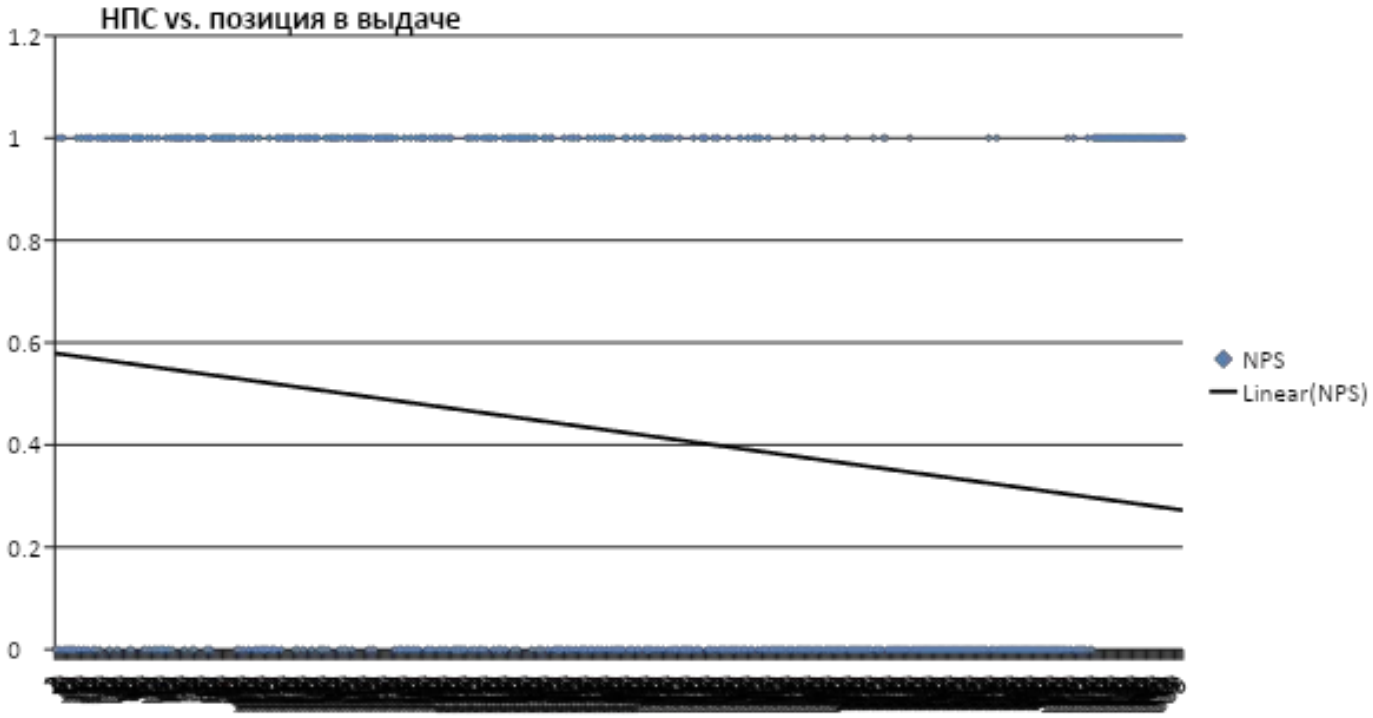


Анализ не полной выдачи, а топа = ощупывание слона слепцом.
Ищем маленькие выдачи, которые можно ощупать полностью.
У меня таких 250К (500-1000 сайтов) из которых 30К (900-1000 сайтов)

Сначала было слово…

…а потом – ошибка кодировки

Приведена полная выдача (X) и отмечены НПС-результаты (Y)
Запрос – секретный, но картина стандартная. ☺
В конце выдачи подряд идут «плохие» НПС.



Евгений Трофименко info@promosite.ru

Как применить?

1 вывод: **баненные и/или АГСные доноры не работают.**
…а если и летают, то очень-очень низко. В хвосте.

Не можем найти ссылающиеся сайты-доноры для НПС из хвоста выдачи – они все под АГС или в бане.

2 вывод: **построение сетки для сравнения качества доноров.**

Берем НПС чуть выше «плохих» - вот маркер для поиска мусора.
Разбиваем выдачу выше на фрагменты – вот сетка для перехода к количественному измерению.

… дело за малым – найти НПС по ссылкодонорам.

4. Переформулировки поисковых запросов

Лето 2008 – введены переформулировки запросов:
Теперь поиск Яндекса (версия "Магадан") еще учитывает следующие отношения:
а) некоторые типы переходов из одной части речи в другую ("гамбург" -> "гамбургский");
б) транслитерация ("mazda" -> "мазда");
в) аббревиатуры (МГУ -> Московский государственный университет).

Примерно в то же время отменен показ «переколдовки» и существенное увеличение граничных расстояний (поиск соседних слов в пределах документа)

С помощью добавления некоторого вида слов и операторов в XML **можно получить информацию о переформулированном запросе**, который, вероятно, отрабатывает вместо введенного.

в архиве их много.

Все так, как и обещалось – большие расстояния или ограничения расстояний, новые части речи, новые операторы и термы.

Пример переформулировки:

продвижение сайтов

=> становится:

(продвижение::19047
^ ((про::2793-движение::8030))
^ продвигать::40288
^ продвигаться::199208)
&&/(-32768 32768) сайтов::410

- Новые части речи, транслит, аббревиатуры
- Большие расстояния
- «двоеточечные» веса
- Оператор ^ (терм не обязан присутствовать, но если есть, это плюс)
- Точные фразы и ограничения расстояний
- Почему-то возвратные глаголы тоже отдельно

ограничения расстояний

очень «короткие» и очень «длинные»

анализ финансового состояния предприятия
анализ::8714 &/(-1 1) финансового::6288 &/(-1 1) состояния::5054 &&/(-7 7) предприятия::3492

дизайны комнат нижний новгород
дизайны::4379 &&/(-7 7) комнат::6878 &&/(-7 7) нижний::8101 &&/(-7 7) новгород::10583

5800 nokia
5800::248895 &/(-3 3) nokia::12493

партия единая россия
(партия::10385 &&/(-32768 32768) ((единая::10481 &/(-1 3) россия::827) ^ **ер**::234393) ^ !!**едро**::2480323) ^ !!**педирос**::492344160

разбиение на фрагменты

И склейка фрагментов

downloadmanager
downloadmanager::27273214 ^ ((**download**::1501**-manager**::7788))

z11xrn *(модель ноутбука)*
z11xrn::709103565 ^ (!(**z**::3403 &/(1 1) **11**::672 &/(1 1) **xrn**::39394642)) ^ ((**z11**::1975218 &/(1 1) !**xrn**::39394642))

ps 3
(**ps**::19277 &/(-1 1) **3**::229) | **ps3**::56914

переводчик онлайн
(переводчик::30986 ^ перевод::7100) &&/(-32768 32768) (**онлайн**::2124 ^ **online**::3661 ^ ((**он**::301-**лайн**::28714)))

«двоеточечные» веса

Веса слов разные по трем коллекциям

По каждому слову есть двоеточечный вес, и слова в запросах часто повторяются.

У одного и того же слова может быть несколько разных весов для разных запросах!

Есть три коллекции документов, по каждой считается свой вес.

Русская (запрос с русскими словами)
Англоязычная (запрос весь из цифр и английских букв)
Украинская (пример: музика скачати безкоштовно)

Одно и то же слово может обладать разной контрастностью для разных баз. Разное число документов, разная популярность слов.

ФИО – новые зоны и термы

Ахтунг!!! Экстракция сущностей в большом поиске!

Для запросов, содержащих имена в виде 2+ слов

иосиф бродский

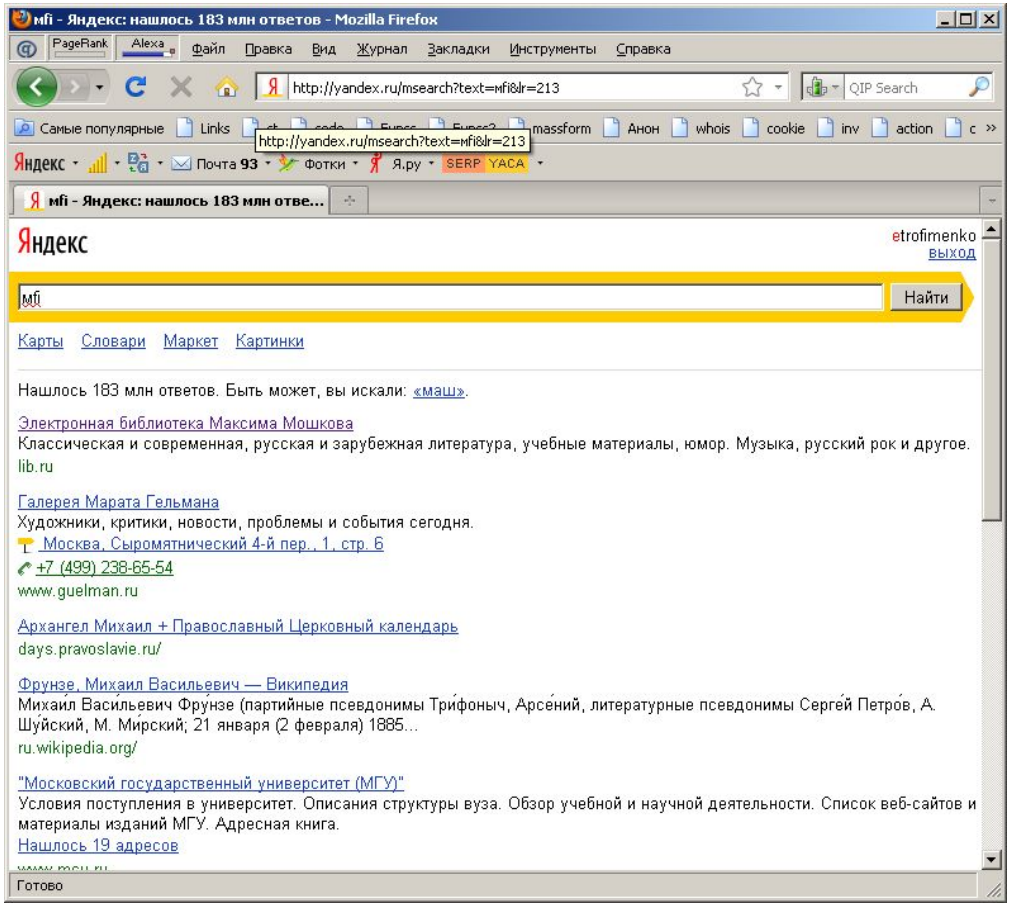
Переформулируется с фрагментом
*** (
fioname[((**иосифfi**::178320 &&/(-32768 32768) !!бродский::358329))] |
fiinname[((**ifi**::3277 &&/(-32768 32768) !!бродский::358329))] |
fiinoinname[((**ifi**::3277 &&/(-32768 32768) !!бродский::358329))] |
finame[((**иосифfi**::178320 &&/(-32768 32768) !!бродский::358329))]
)

Новые операторы (новые зоны?) соответствующие поиску по имени
Новые термы (ifi) – поиск всех имен на букву «И» и сокращений

Евгений Трофименко info@promosite.ru

mfi – все имена на букву M

экстракция объектов из текста…



Евгений Трофименко info@promosite.ru

4.5 Какая польза?

Раньше мы знали про переформулировки, но теперь очевидно, что **переформулировка производится на уровне исходного запроса**
Поэтому «дополнительные» слова обязаны давать вклад в релевантность, это не просто подсветка.

- **Новые операторы** (^, fio* и другие)
- Использование доп. слов **при оптимизации и в ссылках**
- Знания об **ограничении расстояний** в переколдовке – необходимы!
- **Веса слов** тоже полезны

Возможно, это будет внедрено в сервис http://tools.promosite.ru/

А экстракция сущностей в большом поиске - ***это мощные изменения…***

И ведь без микроформатов и разметки…

Евгений Трофименко info@promosite.ru

5. Контрастности (веса) слов

::вес – это НЕ IDF (классический)

IDF (*inverse document frequency* — обратная частота документа)

$$\mathrm{IDF} = \log \frac{|D|}{|(d_i \supset t_i)|}$$

А как выглядят набор **::вес**ов – дискретный набор, являются целочисленными дробями от максимального веса.
По куску коллекции ---

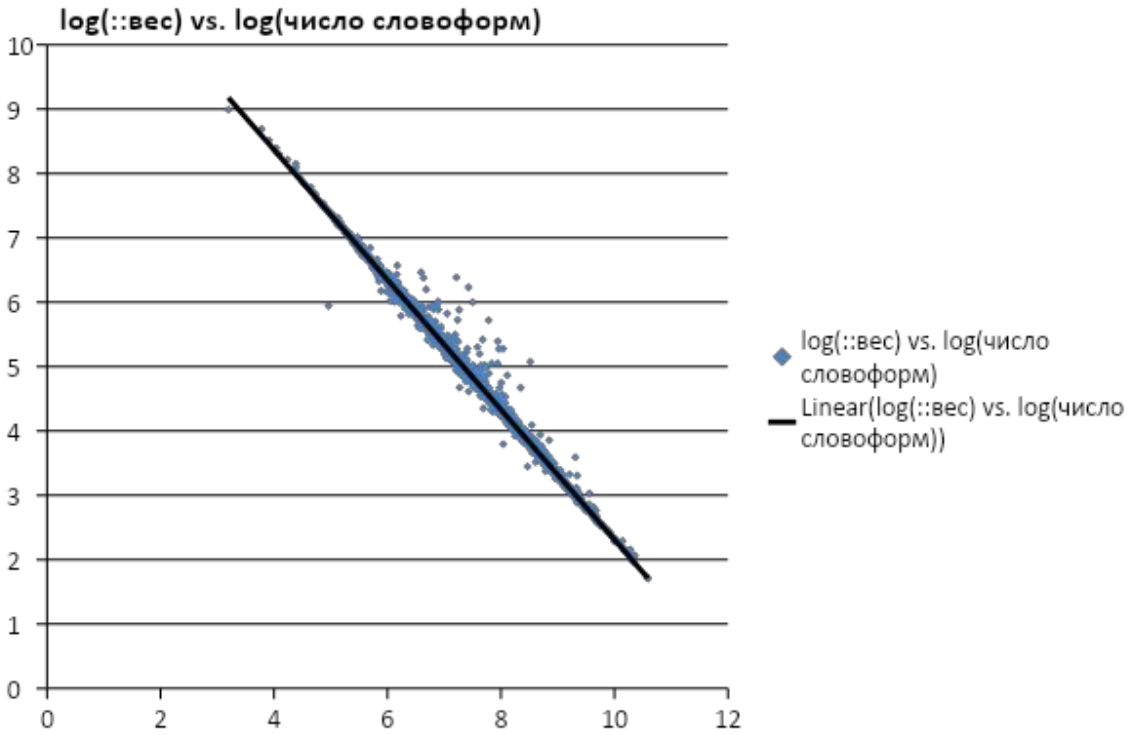
::вес	слов	отличие, раз
984688320	2080	1
492344160	302	2
328229440	206	3
246172080	197	4
196937664	148	5

Догадываемся - **::вес=D/Di**
Это отношение числа документов.
Чтобы получить IDF, берем логарифм:

=> IDF=log(::вес)

::веса -не документные?

А от словоформ? Не IDF, а ICF?



6. Тестовый XML

Лето 2010: http://xml.yandex.ru/test_query.xml

```
<group>
  <categ attr="d" name="detskaya-poliklinika.ru" />
  <relevance priority="phrase">106678464</relevance>
  <doc id="13-6-15-ZEBD96DEF8527C4F3">
    <relevance priority="phrase">106678464</relevance>
    <url>http://www.detskaya-poliklinika.ru/</url>
    <domain>www.detskaya-poliklinika.ru</domain>
    <properties>
      <BaseType>rus</BaseType>
      <_Factor_DocLen>0.047059</_Factor_DocLen>
      <_HeadlineSrc>dmoz</_HeadlineSrc>
      <_HilitedUrl>www.<hlword>detskaya</hlword>-<hlword>poliklinika</hlword>.ru</_HilitedUrl>
      <_UrlMenu>[["medinfa.ru/", "medinfa.ru"], ["medinfa.ru/polyclinic/", "Поликлиники Москвы"], ["",
"detskye_policliniky"]]</_UrlMenu>
      <catalog>title="Детская поликлиника Медэп" ;desc=Медицинские программы. Онлайн-консультации
специалистов. Контакты.
;screenshot=http%3A%2F%2Fcards2.yandex.net%2Fcat-get%2F4103%2F6e099396ce8211de996dc1afc15d2a
cd.png ;</catalog>
      <clon>47784</clon>
      <geo>213</geo>
      <geoa>213</geoa>
      <lang>ru</lang>
    </properties>
```

Что показывалось:

<relevance priority="phrase">106678464</relevance>
-числовые значения релевантности группы в целом и элемента
Релевантность группы и первого эл-та не всегда совпадают!!!

<geo>213</geo> - ID регионов
<geoa>213</geoa> - ID регионов (автоматическое?)
<clon>47784</clon> - яндексовый ID группы аффилиатов

<_HeadlineSrc>dmoz</_HeadlineSrc> - описание взято из DMOZ (?)
<_UrlMenu> или **<snippets><sitelinks>** - «быстрые ссылки» (?)
<catalog> - параметры описания сайта из Я.Каталога
<_Factor_DocLen> - нормированная длина документа вида N/255

Что удалось взять:

топ-1000 по **42К** запросов (seorate+частотные)

В сумме:

2.5М разных хостов (отдельно **2М** из выдач+**650К** ссылкодоноров)

записей про клоны 234К, хостов с клонами 185К (**7.4% хостов**)
записей про geo 396К, хостов с приписанным geo 360К (**14.4% хостов с geo**)
записей про автоматическое гео 1.6М, хостов с приписанным geoa 1.54М (**62% хостов с geoa**)

Взята география по всем ссылкодонорам.

Видимо, определение клонов автоматическое.

Аффилировалка - автомат?

http://tools.promosite.ru/use/clones.php

Что наводит на мысль об автоматическом определении клонов:

1. Очень много хостов с клонами (7.4%) – вручную не осилить.
2. Очень крупные группы клонов на субдоменах ucoz.ru, co.cc, …
3. Частенько аффилируются сайты с полностью разным контентом.

Методы подтверждения аффилиатов:

1. В XML с группировкой по хосту: **host:site1.ru | host:site2.ru**
2. В выдаче проверяем 1-2 места: **site1.ru | site2.ru**

В сказки про то, что Яндекс не борется, мы уже не верим…
Метод определения клонов через оператор domain перестал быть удобным после того, как работу оператора специально искривили.

Евгений Трофименко info@promosite.ru

Цифры релевантности

400111552

Очень похожи на моделирование оценок асессоров (0-4)*100M

4xx M – витальные результаты
3xx M – почти витальный ответ на поиск domain.ru
2xx M – по отдельным странным запросам (некоммерческим?)
1xx M – самая массовая группа
[1-9]x M – не очень релевантные документы

4xx M – витальные результаты

Есть по 4.5К запросам из 42К (11%)

Много действительно витальных:

пик недвижимость	400108192	www.pik-estate.ru
ferrari	400107584	www.ferrari.ru
мир лимузинов	400109440	www.limo-world.ru
цб рф	400109952	www.cbr.ru
asus eee pc	400109600	eeepc.asus.com
mozilla firefox скачать бесплатно	400110048	www.mozilla-russia.org

Но есть и сомнительные:

горячие туры	400108224	www.hott.ru
лучшие интерьеры	400108448	www.lui.ru
интерьеры махараджей	400108352	vostok-art.ru
днс	400109472	www.dns-shop.ru

Без ручной правки не обошлось, хотя запросов много…

3xx M – поиск домена

Некоторым очень везет…

reklama lv	300109408	www.reklama.lv
macbook pro	300107072	macbook.pro
demotivation.ru	300110304	www.demotivation.ru
www.picnik.com	300111360	www.picnik.com

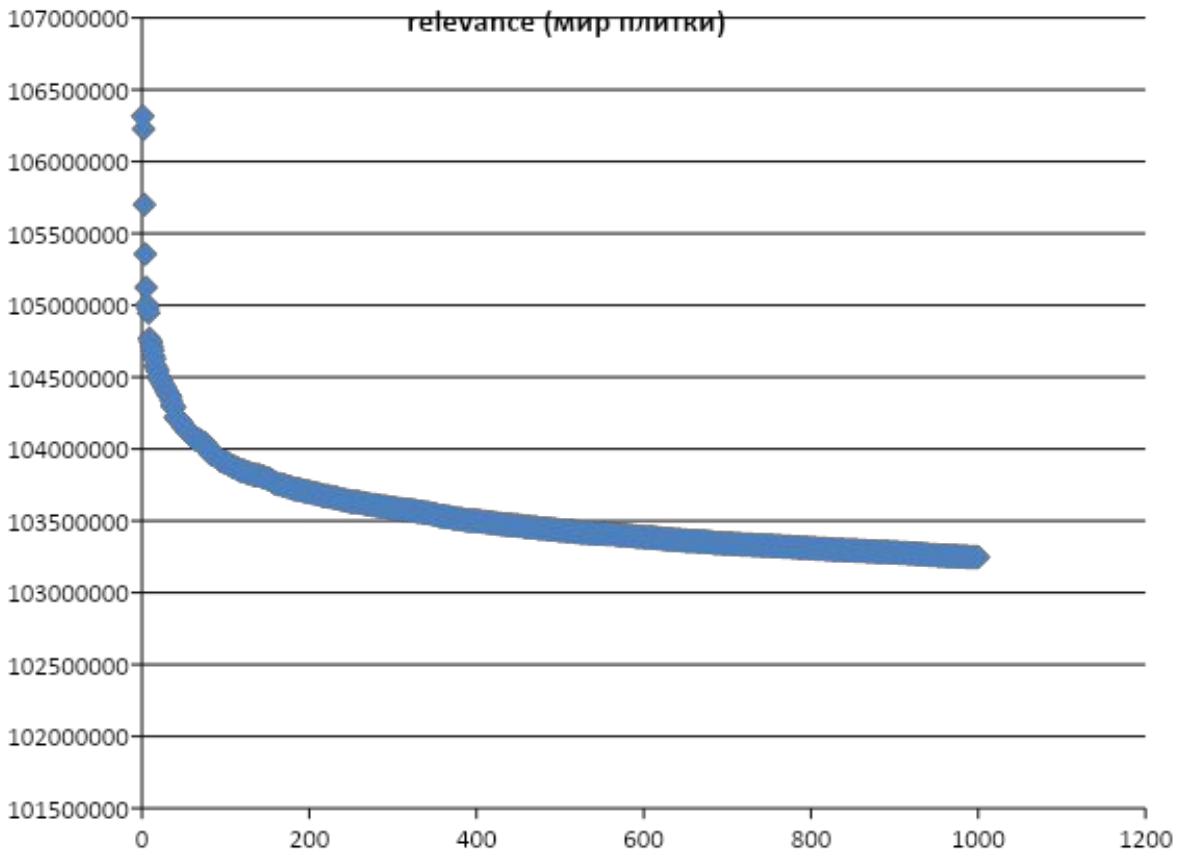
2xx M –странные запросы

1.2К запросов вся выдача из 2xx

глисты
трихомонада
язва
диатез
погрузчик
инвестиционный проект это
прямые инвестиции это
андеррайтинг это
газель
индекс доу джонса
беременные
александр головин
джинсы
клещи
кира пластинина

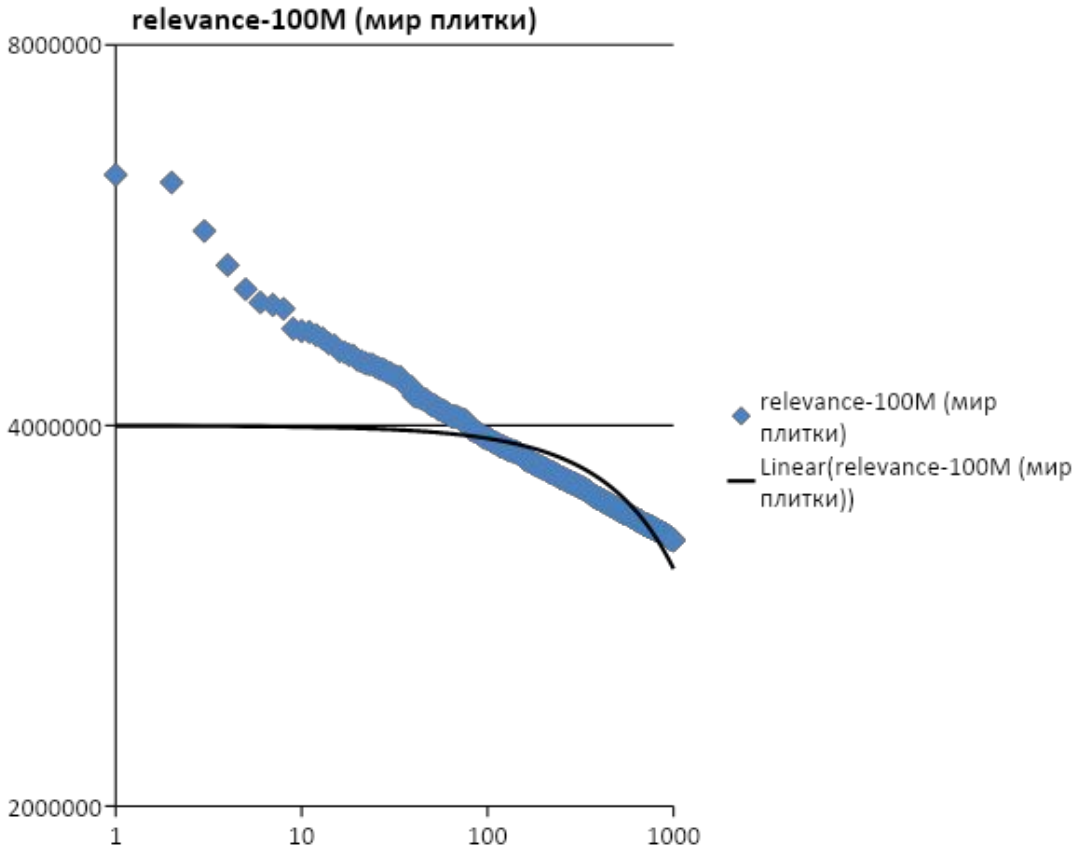
Евгений Трофименко info@promosite.ru

1xx M – все остальное



1xx M - поиграемся

Степенной закон, чо…



Евгений Трофименко info@promosite.ru

[1-9]x M - поражены в правах?

Дублирование контента?

Как правило, заспамленные тематики, но могут быть приличные сайты
Позиции: очень глубоко. Предположительно, дублирование контента.

запрос	***место***	***relevance***	***сайт***
ролики бесплатно смотреть онлайн	371	99999952	vkontakte.ru
индийские фильмы онлайн смотреть бесплатно	964	99999784	kolotibablo.com
фото приколы	946	99963176	www.pravda.com.ua
медицина для вас справочная	656	99920152	www.aptekari.com
санаторий приокские дали	615	99875184	hh.by
отзывы об отелях	596	**10444148**	tourout.ru
аудиокниги скачать бесплатно	961	**10423272**	zapomni.org.ua
встраиваемые духовые шкафы	921	**10400875**	shopv.ru
каталог отелей	853	**10361856**	www.elio-tour.com
фильмы онлайн бесплатно	932	**10319647**	binmovie.org
предстательная железа	973	**10281967**	www.3630363.ru
венерические болезни	968	**10268491**	base.consultant.ru
анатомия человека в картинках	962	**10237372**	i-60.livejournal.com
лимфогранулематоз	985	**10157077**	www.ma-ma.ru

Евгений Трофименко info@promosite.ru

ВСЁ.

Эволюция алгоритмов Яндекса и методов исследований: новые возможности анализа

Трофименко Евгений

сЭо-эксперт

info@promosite.ru

http://tools.promosite.ru/