



Использование особенностей языка запросов поиска Яндекса для исследований

Трофименко Е.А. Корпорация РБС, начальник отдела исследований и аналитики
trofimenko.evgeny@rbscorp.ru, <http://www.bdbd.ru>



Краткое содержание

Особенности работы операторов исключения

- Оператор «минус» не применяется к текстам ссылок
- Оператор «~~» вычищает НПС-результаты из выдачи
- Можно узнать, как Яндекс расширяет запрос пользователя
- Особенности контекстных ограничений и использование Яндекс.XML

Как используют операторы поиска по результатам прямого эфира

- Поиск дублей текстов
- Ошибки и «чужие» операторы
- Пробивка «тыпоследнего» и сборы баз

Возможности мониторинга особенностей выдачи

- Изменение макро-параметров (НПС, главные, ЯК) во времени
- Изменение макро-параметров по отдельным классам запросов



Операторы Яндекса

Присутствие: **+** (плюс) - слово обязано находиться

Исключение:

~ ~ (тильда) - исключение в пределах документа (~ предложения)

- (минус) - недокументированный: был исключением в контексте

Возможные контексты поиска:

- Документ (текст), Предложение (текст)
- Ссылки (анкор-файл)
- ...расстояние в несколько слов или предложений, указанное явно операторами $/(-N +N)$ или $\&\&/(-N +N)$



Как SE должны исключать результаты при отсутствии информации о документе?

Поиск точно знает, что есть и чего нет в тексте страницы

Поиск не уверен, что знает абсолютно все тексты ссылок

Поиск работает и по текстам, и по ссылкам.

Поэтому при исключении слов у поиска два варианта:

Исключать то, что есть в тексте страницы, и:

1. искать по тексту ссылок («не знать» о его существовании, оператор «минус») - оставляем НПС
2. не искать по тексту ссылок (делать вид, что ссылок не существует, оператор «~~») - исключаем НПС



Запрос [слово -слово]: оставляем НПС-результаты

слово - документы, содержащие слово в текстах или во входящих ссылках
-слово - исключаются документы, содержащие слово в текстах
Итог: найденные по ссылкам документы (сниппет м.б. из Я.Каталога)

The screenshot shows a Yandex search results page for the query "слово -слово". The browser address bar shows the URL: <http://yandex.ru/yandsearch?text=слово+-слово&stpar2=%2Fh1%2Ftm8%2F>. The search bar contains the text "слово -слово" and a "Найти" button. Below the search bar, there are options for "в найденном" and "в Москве", and a "расширенный поиск" link. The search results show 150 thousand pages found. The first result is from mignews.com.ua, titled "Новое русское слово" (Европа-СНГ)..., with a file size of 20 KB. The second result is from <http://wordstat.yandex.ru/>. The page also includes navigation links like "Почта", "Мои находки", "Настроить поиск", and "Войти...".



~ и ~~ : форсируем поиск по текстам и удаляем НПС-результаты

Добавляя в любой запрос исключение ~~абракадабры, удаляем НПС:

Для доля найденных по ссылке результатов относительно высока:

автомобили	42% НПС
продажа автомобилей	12% НПС
аренда автомобиля с водителем	7% НПС



Расширение пользовательских запросов

Яндекс и раньше мог добавлять в запрос новые слова («что такое X»), но делал это редко, индивидуально.

Сейчас - расширение запроса поставлено на поток.

- Переходы из одной части речи в другую
(*гостиницы в Москве -> московские гостиницы*)
- Транслитерация («*mazda*» -> «*мазда*»)
- Аббревиатуры (*МГУ -> Московский государственный университет*)

Как узнать слова, которыми расширяется запрос?

Используем операторы исключения.



Исключаем точную форму слова: оставляем переформулировки

При исключении слова из запроса - в выдаче остаются и подсвечиваются переформулировки (+найденное в URL):

слова запроса –слово	Работает, но оставляет смесь переформулировок и НПС
слова запроса ~~!!(слово)	работает. Разное: гостиницы в москве гостиницы москвы



Особенности «колдунщика»: существует ли ограничение расстояний?

Колдунщик = расстановка неявных для пользователя ограничений на расстояние между словами, известен с 2004.

Например, для запроса «новый год» находились документы, содержащие от «год новый» до «новый [*] [*] год»

Эти ограничения можно было посмотреть. Но это закончилось. Однако вручную введенные ограничения отработывали.

Как узнать реальные ограничения на расстояние между словами?

Попробуем подобрать...



Пытаемся подобрать: перебор 7 частотных операторов

По статистике запросов Корпорации РБС, наиболее часто использовались:

&	Относительно: 100%	в пределах одного предложения
&/(-2 4)	9%	-2 +4 слов
&/(-1 3)	10%	-1 +3 слов
&/(1 1)	2%	строго по порядку
&&/(-7 7)	15%	в пределах 7 предложений
&&/(-3 3)	15%	3 предложений
&&	7%	в пределах документа

Перебор вариантов НЕ ДАЕТ РЕЗУЛЬТАТОВ...



Как себя ведут в Яндексе контекстные ограничения?

Из релиза Яндекса, Магадан:

«Мы смягчили фильтрацию отбора документов для ранжирования, что привело к улучшению ранжирования по запросам, для которых релевантные документы содержат слова запроса далеко друг от друга»

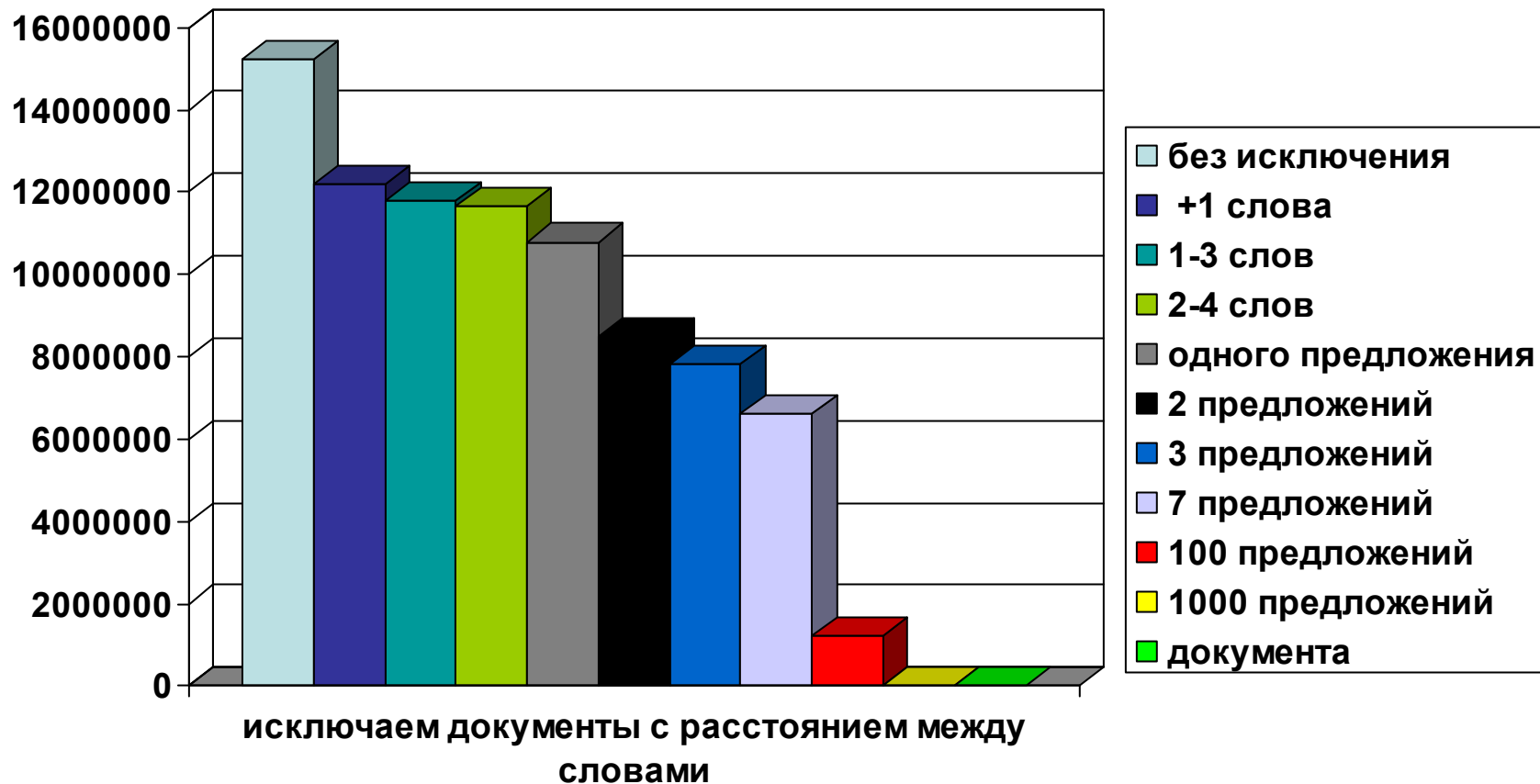
Попробуем поискать эти ограничения.

1. Берем запрос (+новый +год), оба слова должны находиться
2. Исключаем результаты поиска, в которых слова находятся «слишком близко» (от 1 слова до 10 тыс. предложений):
~~(+новый & +год)
3. Смотрим, как меняется число найденных документов... Надеемся, что оно станет нулевым тогда, когда расстояние совпадет с расстоянием в заколдованном запросе...



Число найденных результатов: «далее, чем»

(+ноутбуки +toshiba) ~~(+ноутбуки [ОПЕРАТОР] +toshiba)
исключаем страницы с близкими парами слов





Количество найденных документов

исключаем документы с расстоянием между словами

- без исключения 15227964
- +1 слова 12173770
- 1-3 слов 11806832
- 2-4 слов 11674561
- одного предложения 10788034
- 2 предложений 8482816
- 3 предложений 7813345
- 7 предложений 6604327
- 100 предложений 1223668
- 1000 предложений 7942
- документа 0

Точное количество найденных документов смотрим через Яндекс.XML



Яндекс.XML: релевантность «phrase», «strict», «all»

- `<?xml version="1.0" encoding="utf-8" ?>`
- `<yandexsearch version="1.0">`
- `<request>`
- `<query>(+ноутбуки +toshiba) ~~(+ноутбуки &&/(-3 3) +toshiba)</query>`
- `<page>0</page>`
- `<sortby order="descending" priority="no">rlv</sortby>`
- `<maxpassages>2</maxpassages>`
- `<groupings>`
- `<groupby attr="d" mode="deep" groups-on-page="10" docs-in-group="1" curcateg="" />`
- `</groupings>`
- `</request>`
- `<response date="20090302T140211">`
- `<reqid>00000000</reqid>`
- `<found priority="phrase">0</found>`
- `<found priority="strict">0</found>`
- `<found priority="all">7813345</found>`
- `<results>`
-



Из документации Яндекс.XML

<http://help.yandex.ru/xml/?id=362990>

«приоритеты» соответствия запросу:

«phrase» — число документов с буквальным соответствием запросу

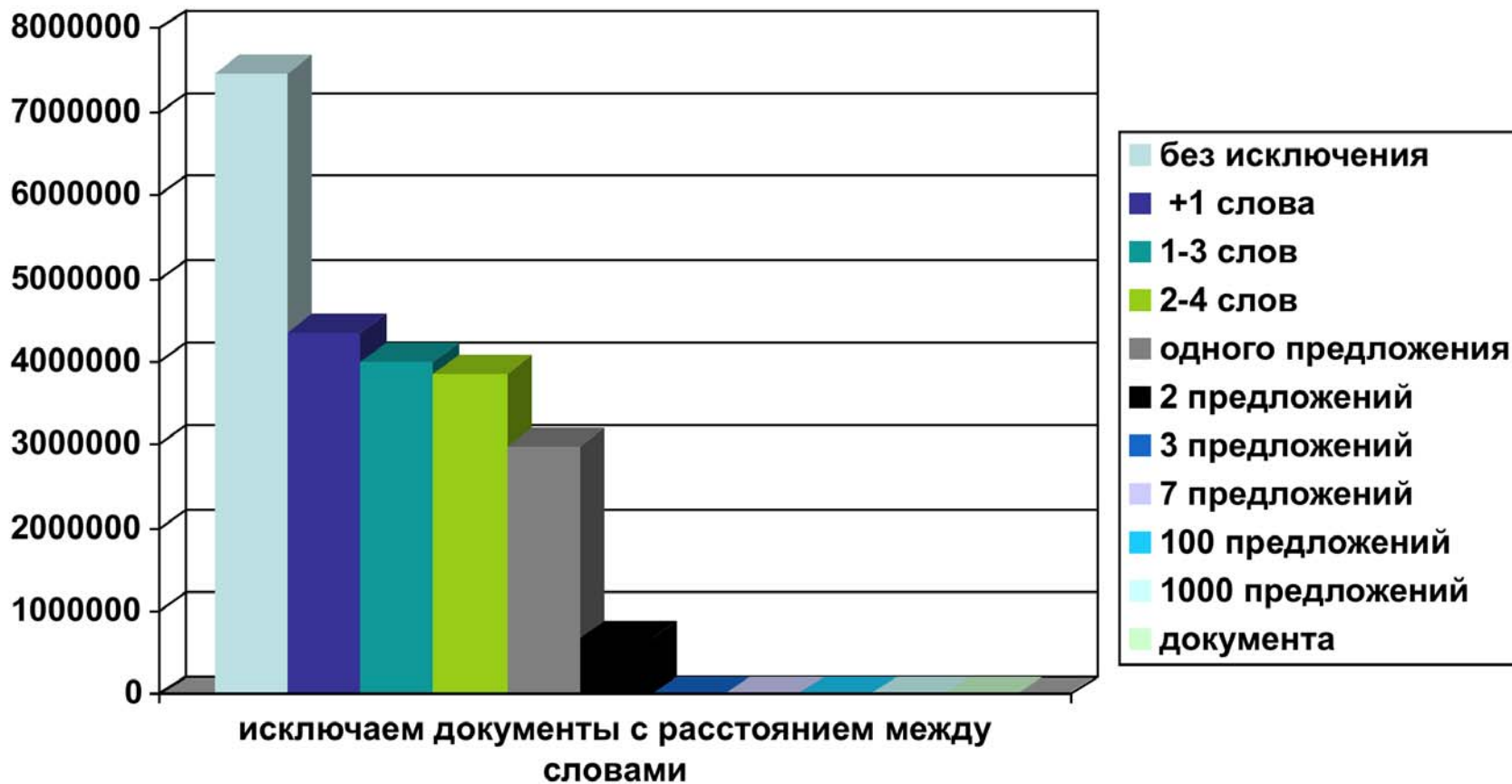
«strict» — число документов с вхождением всех слов запроса в ИСКОМЫЙ КОНТЕКСТ

«all» — общее число найденных документов, которое и было показано



«strict» число найденных результатов: «далее, чем»

(+ноутбуки +toshiba) ~~(+ноутбуки [ОПЕРАТОР] +toshiba)
исключаем страницы с близкими парами слов





«strict» расстояния

	Сейчас	Было
НОВЫЙ ГОД	НОВЫЙ /(-1 1) ГОД	/(-1 3)
шкафы купе	шкафы & купе	
ноутбуки toshiba	ноутбуки &&/(-3 3) toshiba	
toshiba satellite	toshiba &&/(-7 7) satellite	&



Смотрим выдачу: ошибок можно избежать?

Яндекс

Нашлась
1 страница

Поиск [Почта](#) [Новости](#) [Маркет](#) [Карты](#) [Словари](#) [Блоги](#) [Картинки](#) [ещё](#) ▼

(новый год) << url="cd.sportmaster.ru"

в найденном в Москве

[рас](#)

1. [Каталог товаров СПОРТМАСТЕР](#)

найден по ссылке: Компания "Спортмастер" представляет Вашему вниманию лучшие спортивные товары из **новых** коллекции сезона весна-лето 2006 **года** от ведущих мировых... **новые** коллекции...

cd.sportmaster.ru · 1 КБ

[Сохраненная копия](#) · Рубрика: [Спортивные товары](#)

НПС-результат, в котором:

1. Очень большое расстояние между словами
2. Отдельная ссылка с одним из двух слов

... при переколдовке && такое было раньше...



...это «all» - результат:

- `<?xml version="1.0" encoding="utf-8" ?>`
- `<yandexsearch version="1.0">`
- `<request>`
- `<query>(НОВЫЙ ГОД) << url="cd.sportmaster.ru"</query>`
- `<page>0</page>`
- `<sortby order="descending" priority="no">rlv</sortby>`
- `<maxpassages>2</maxpassages>`
- `<groupings>`
- `<groupby attr="d" mode="deep" groups-on-page="10" docs-in-group="1" curcateg=""`
`/>`
- `</groupings>`
- `</request>`
- `<response date="20090302T145647">`
- `<reqid>00000000</reqid>`
- `<found priority="phrase">0</found>`
- `<found priority="strict">0</found>`
- `<found priority="all">1</found>`
- `<results>`



2. Статистика использования операторов в поиске Яндекса

Попробуем регулярно пробивать “прямой эфир”:

<http://stat.yandex.ru/queries/last20.xml>

И искать «неправильные» символы...

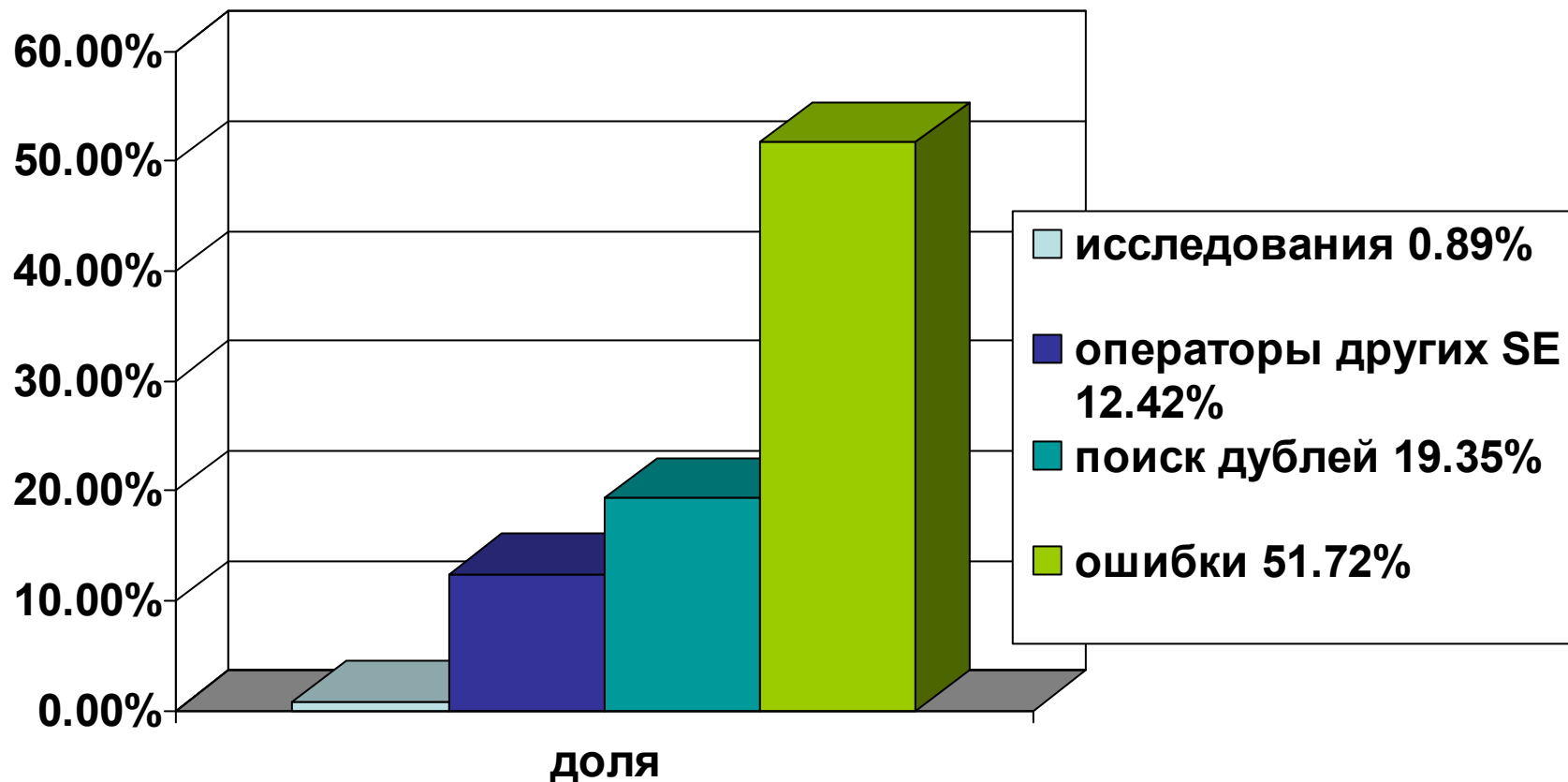
По базе ~300 тыс. запросов:

Только 28% - использование операторов,

Из которых:



Как используют операторы поиска в реальной жизни?





«ошибки»?

<code>\"recomendatcia.ru\"</code>	30.86%
<code>"*.top-famous-recipes.com"</code>	19.35%
<code>#*="www.uchaly.net*"</code>	1.50%



Поиск дублей?

+ " для лечения горла болезней щитовидной "	15.31%
!(командная строка windows)	3.22%
аудиокнига && боярская & сотня && прозоров && скачать && letitbit	0.52%
"!+фильтры !+для !+воды"	0.18%
+(правила проведения сертификации) +(лекарственных средств) +(для животных)	0.13%



Чужие операторы +парсинг для линкспама

url:"maksim-fanclub.ru"	9.51%
site:FLIKKENGAME.COM	1.35%
link:BOMINSOLAR.COM	1.11%
will inurl:/addurl.php	0.26%
intitle:'добавить новую ссылку'	0.21%



Исследования и сборы баз?

((добавить новость)<<(*" .divan-tut.ru"))	0.49%
(Работа в москве) &&/+1000 (Работа в москве)	0.20%
двухконтурные газовые котлы ~~!!(двухконтурные)	0.15%
http -http << (domain="с*" /+2 domain="root")	0.04%
кондиционер::10000 кондиционер::10	0.02%
тут -тут date="20081202"	0.01%

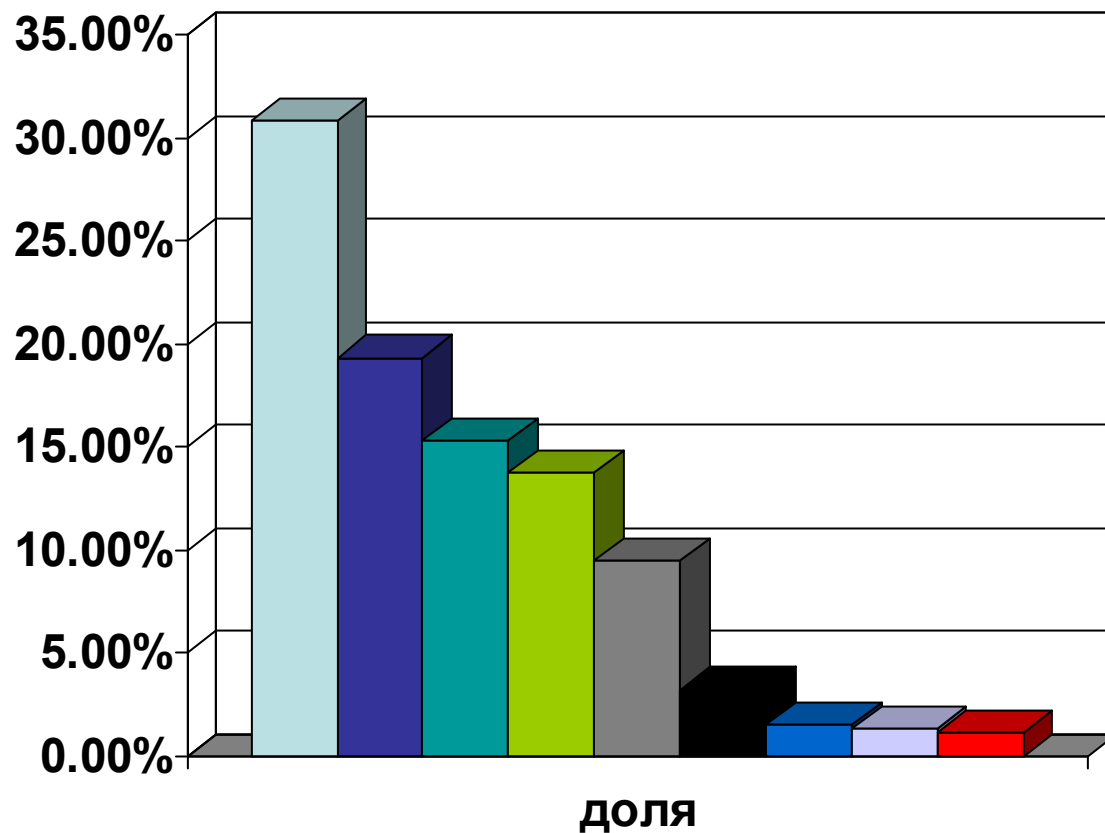


Пробивки и сборы баз

"zhilex-fito.ru/" "www.zhilex-fito.ru/"	13.78%	поиск главного домена + "тыпоследний"?
растекаться \$title/phpBB)	0.55%	сбор форумов
передвижение ~~!!(передвижение)	0.15%	сбор баз переформулировок
(Дмитрий & Лесневский) (!Рен && Лесневский) (REN && Лесневский)	0.12%	мониторинг
(!903/+1!527/+1!36/+1!97/+1)//1	0.07%	мониторинг телефонов, ограничение по расстоянию



Популярные поиски



- "recomendatcia.ru"
- *.top-famous-recipes.com
- +" для лечения горла болезней щитовидной"
- zhilex-fito.ru/|"www.zhilex-fito.ru/"
- url:"maksim-fanclub.ru"
- !(командная строка windows)
- #*="www.uchaly.net"
- site:FLIKKENGAME.COM
- link:BOMINSOLAR.COM



Самое интересное:

Отсутствие в «прямом эфире» результатов
пробивки проиндексированности

`url=" domain/path"`

При этом операторы `domain`, `rhost`
присутствуют...

Значит ли эта фильтрация что-то плохое?

<http://tools.promosite.ru/last20.php>



Анализ и мониторинг алгоритма

Проблемы и вопросы:

1. Влияет ли ТИЦ на выдачу? 😊
2. Почему в Яндексе засилье википедии?
3. Мало пассажиров в результате - к чему?
4. Релевантность: страницы или сайта?
5. Т.д.



Попробуем анализировать:

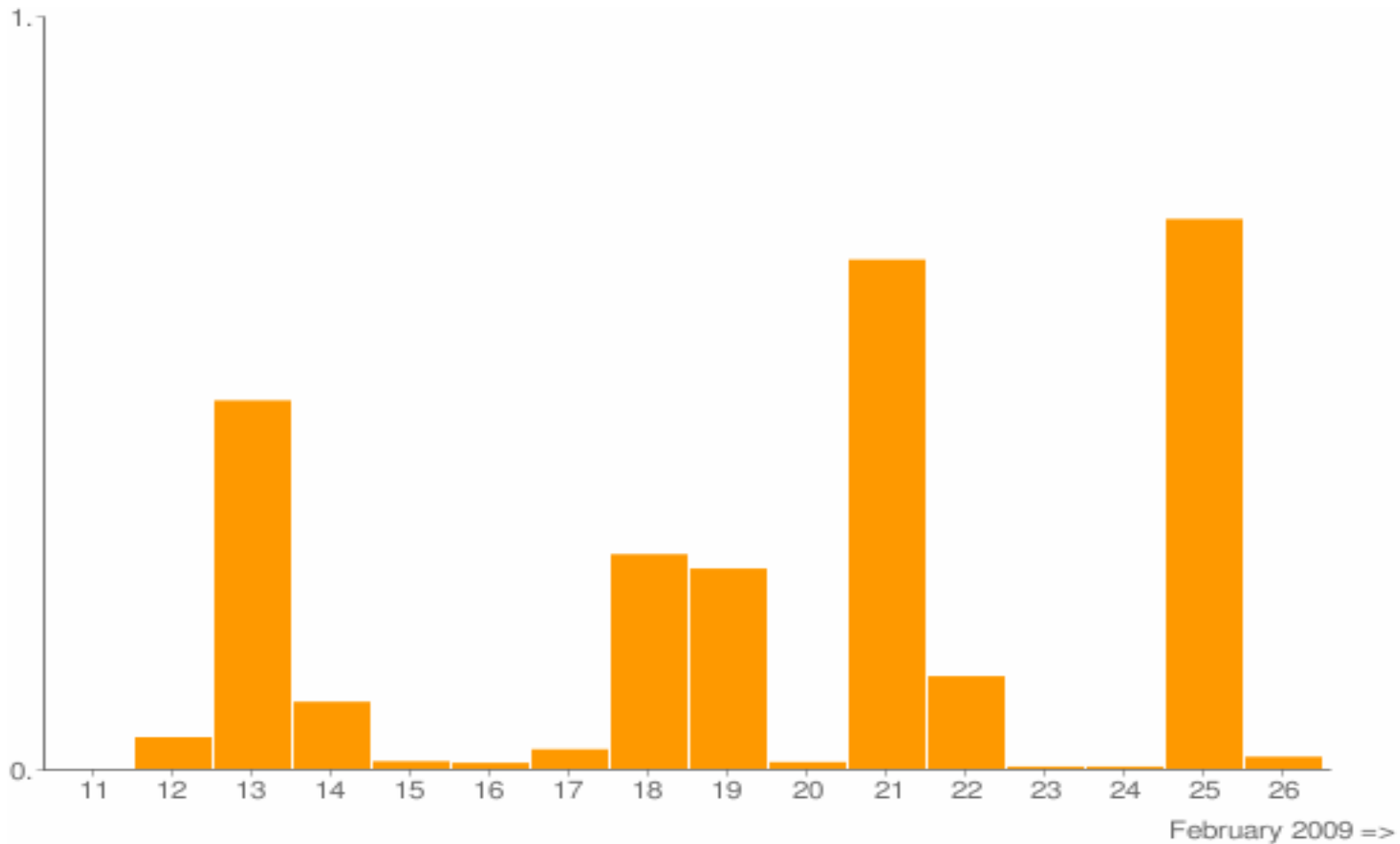
Составим группы запросов, разбитые по

- числу слов в запросе
- конкурентности
- поиску в зонах и НПС
- Т.д.

И посмотрим, как именно изменяются «средние по больнице» параметры во времени?



Апдейты? Релизы? Во времени.





Как взвешивать параметры?

По принципу видимости:

1. Больше место - ниже вес
2. Сумма видимостей по топ50 равна 1
3. Сумма весов 1-10 вдвое выше 11-20

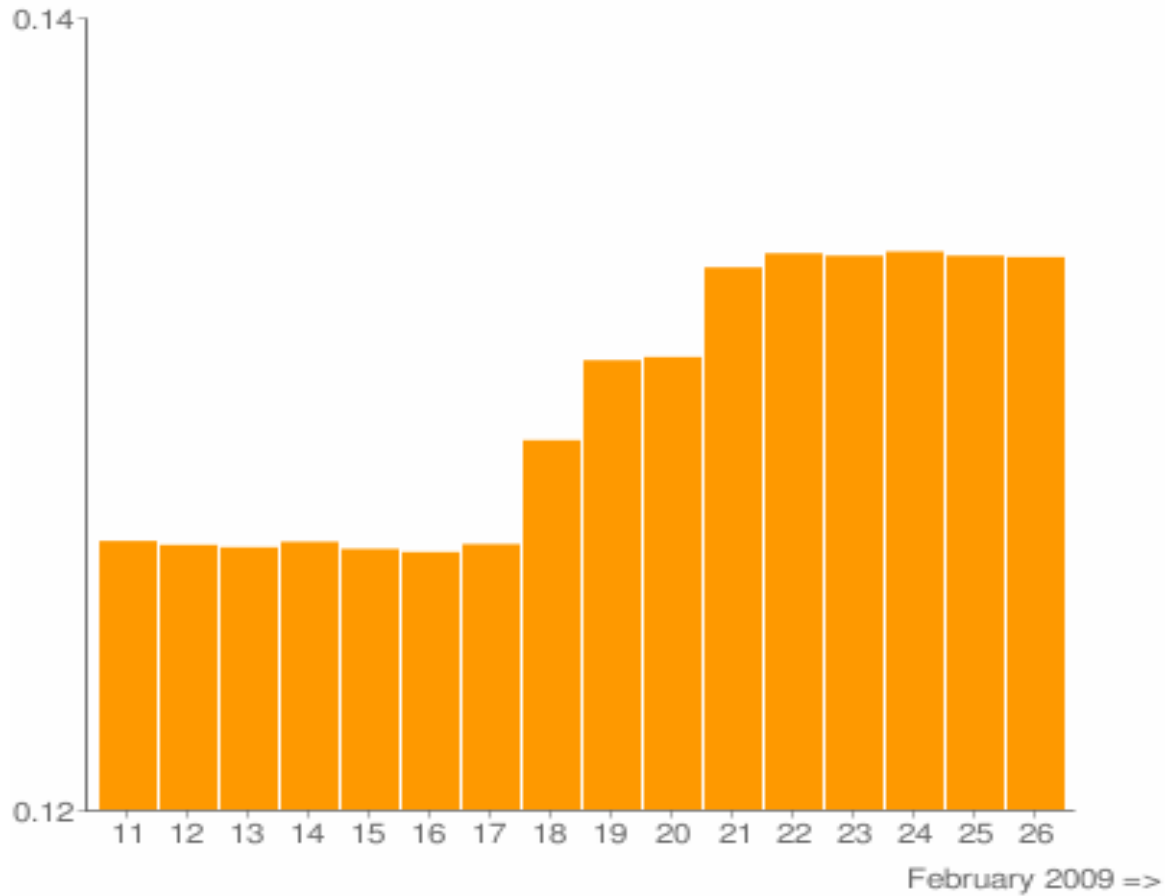
Использую весовую функцию

$$Wt(pos) = 0.074 * 2^{(-pos/10)}$$



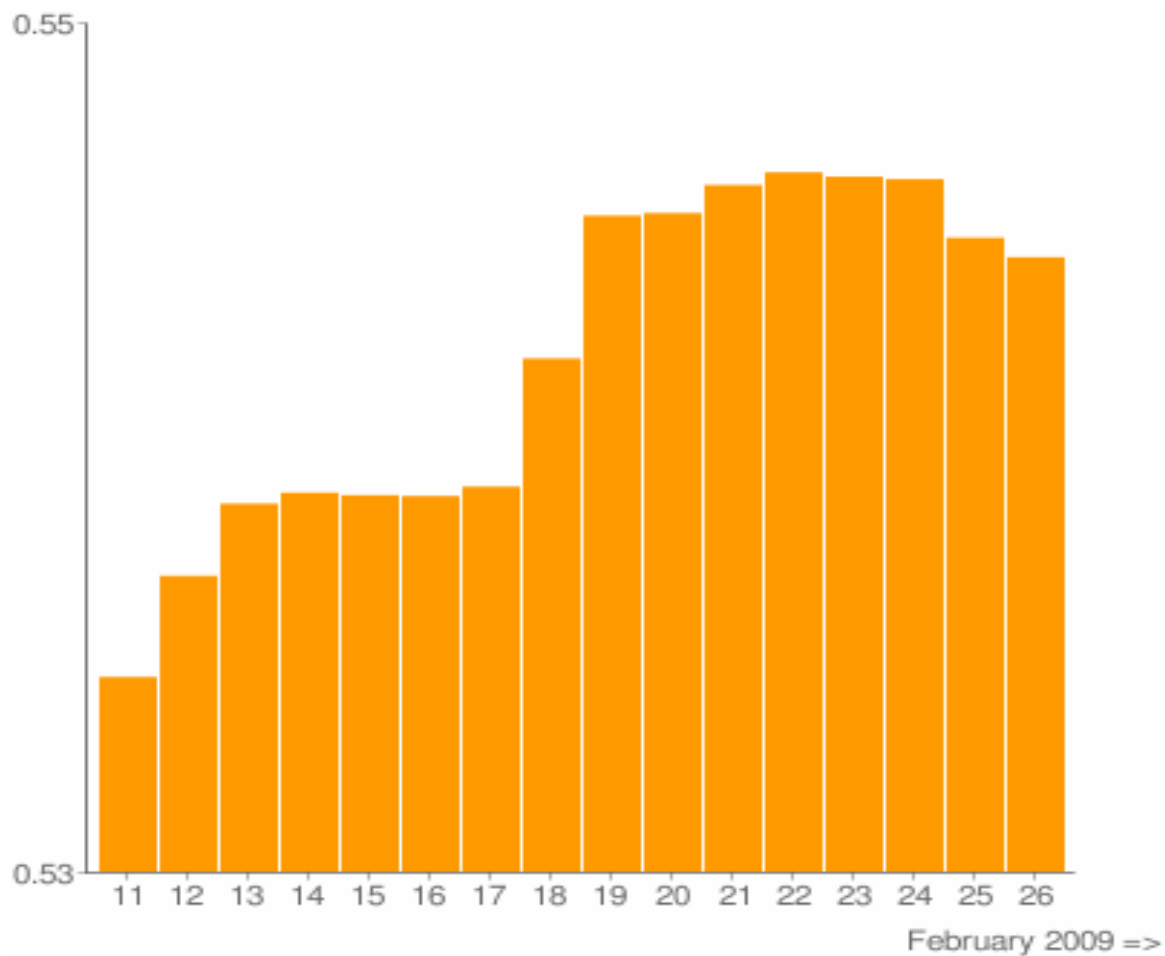
Изменение доли НПС

взв. доля НПС (min=94.4% of max (0.134072))





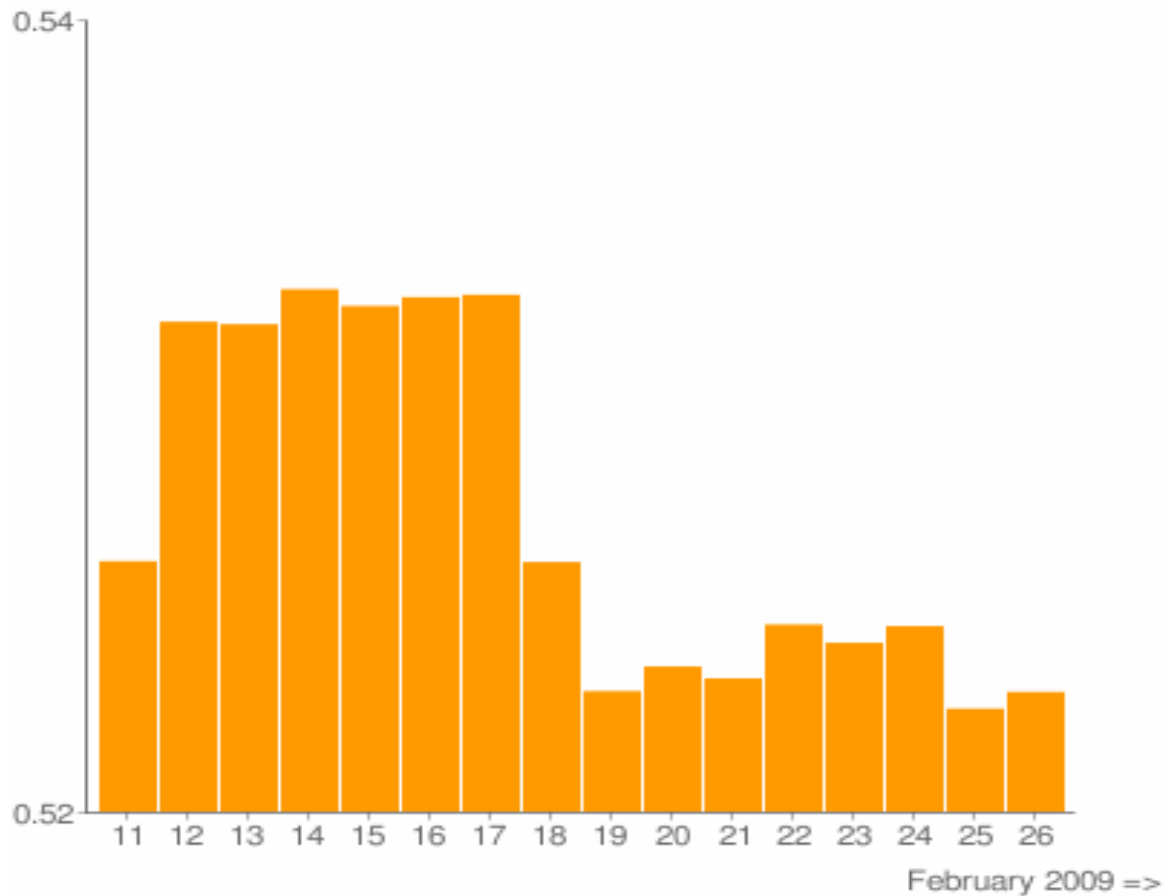
Изменение доли главных страниц в выдаче





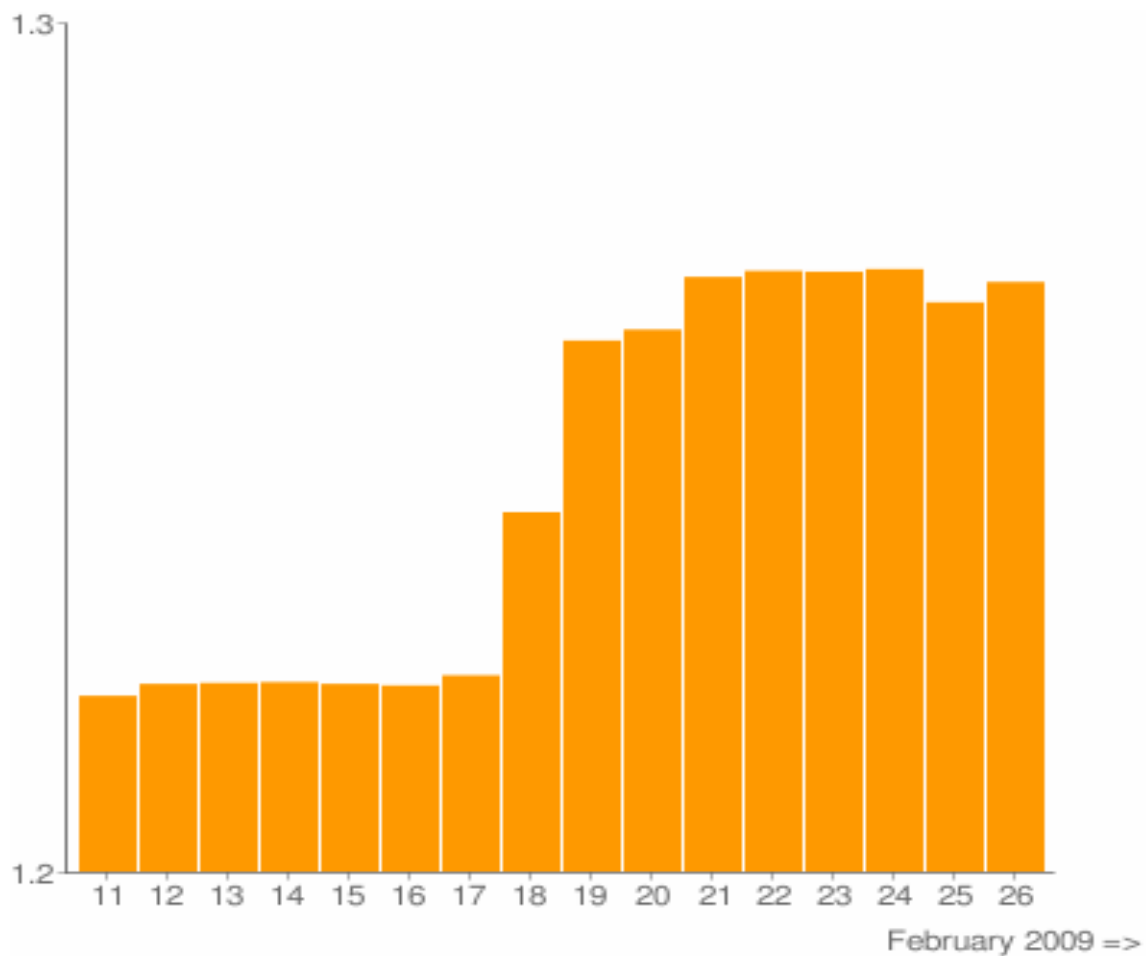
Изменение доли Яндекс Каталога

взв. доля ЯНДЕКС-КАТАЛОГА (min=98.0% of max (0.53318))

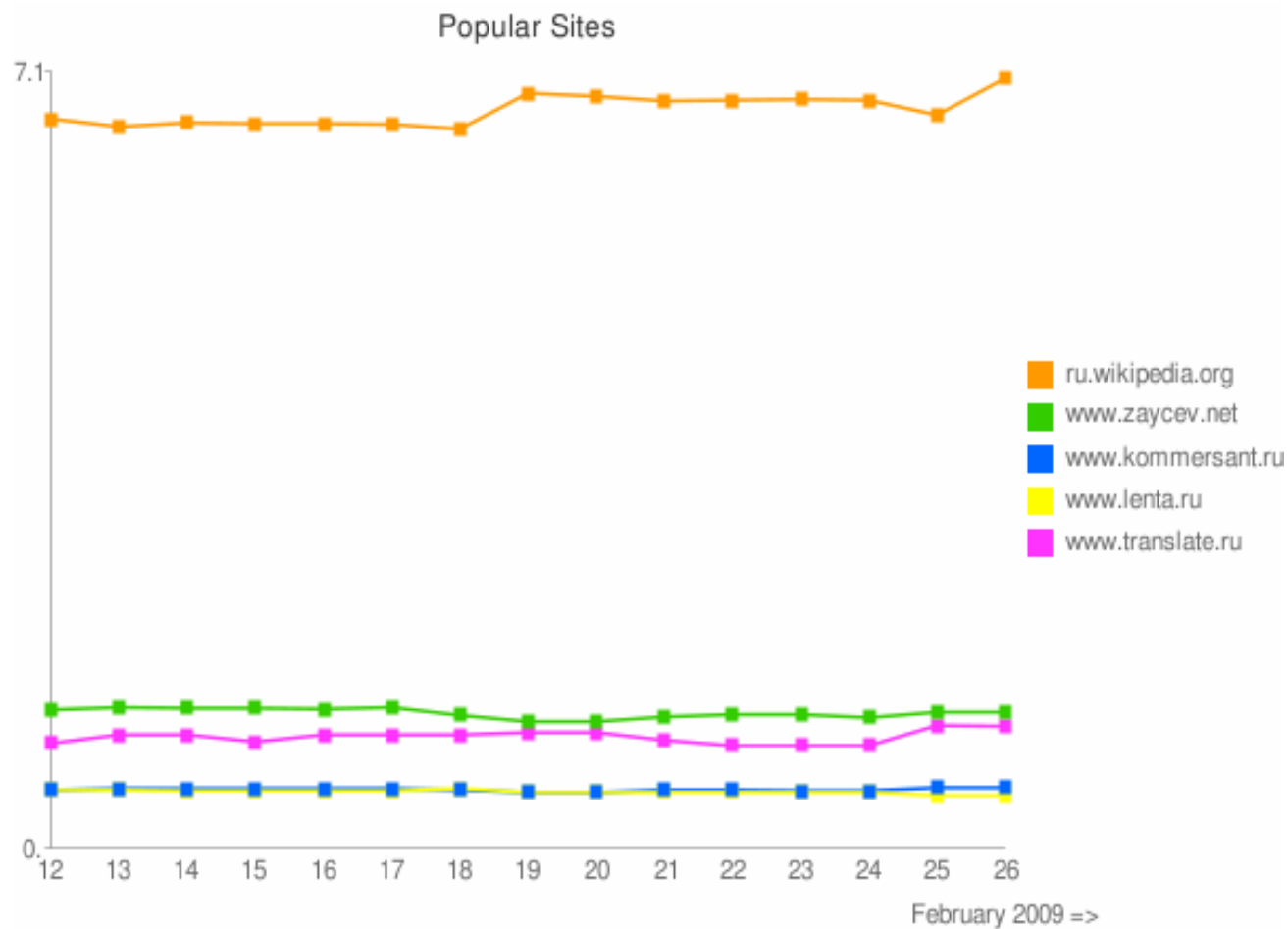




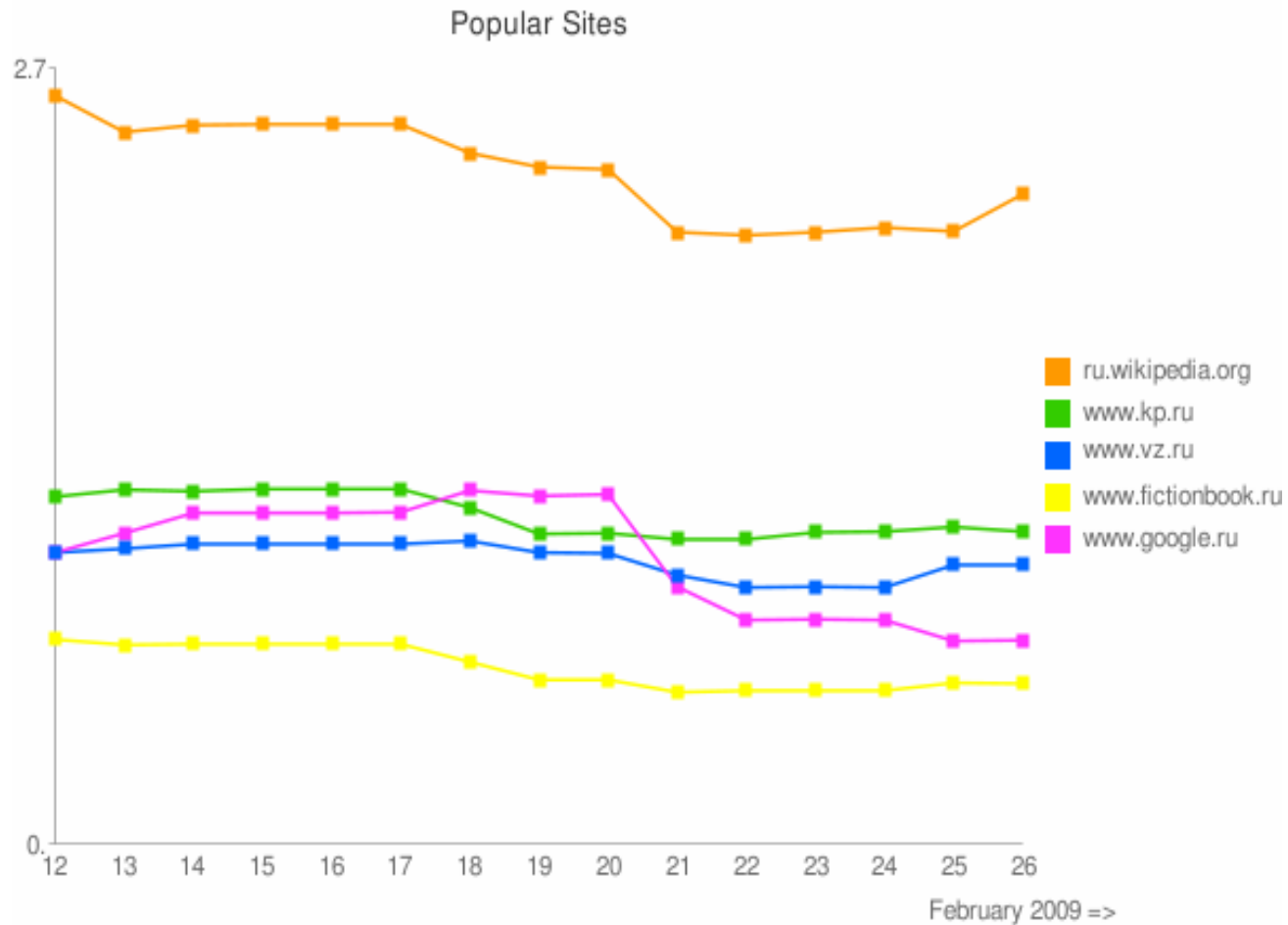
Число найденных пассажиров



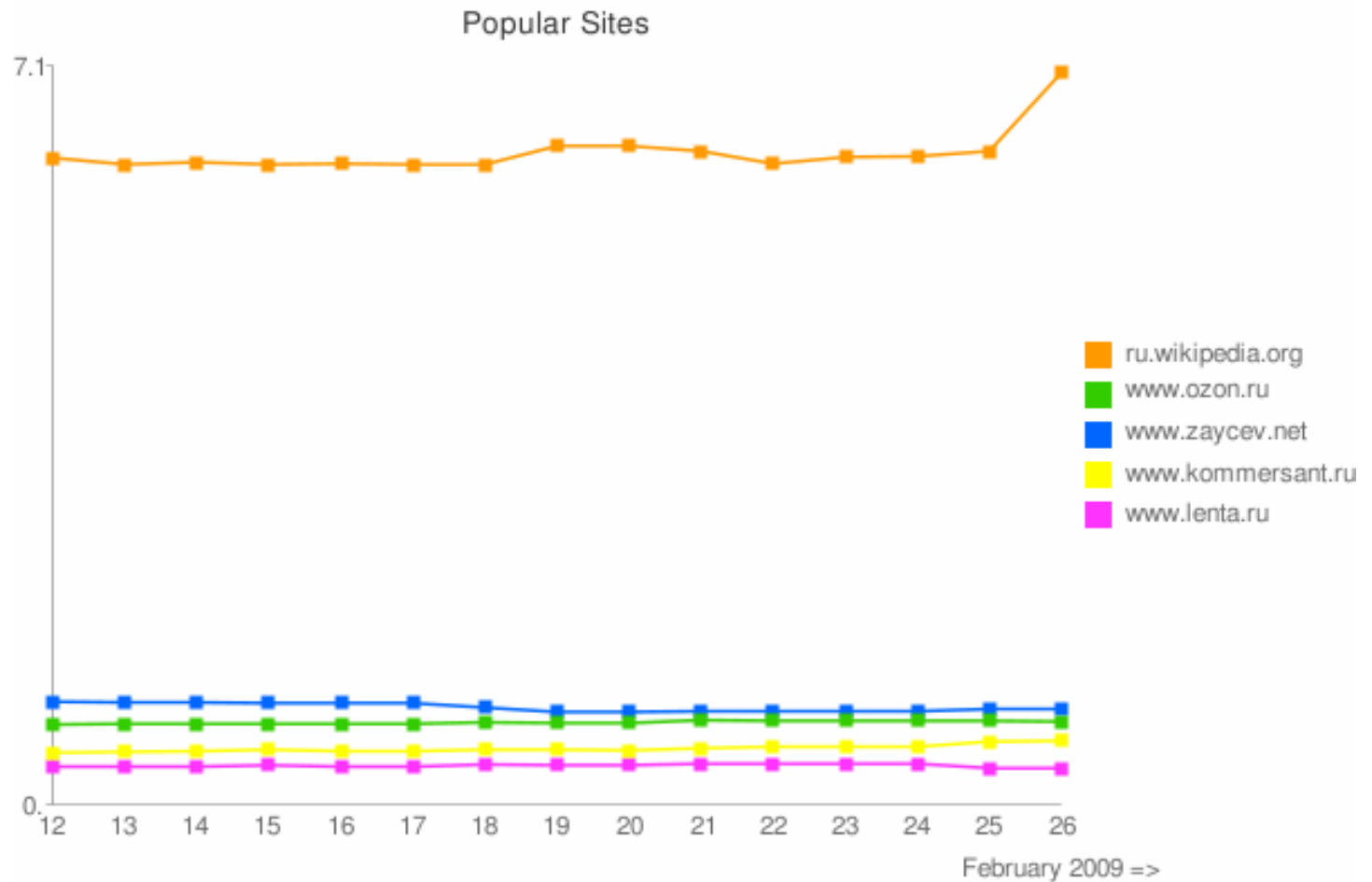
Топ-сайты: обычные запросы



По НПС-запросам слово - слово



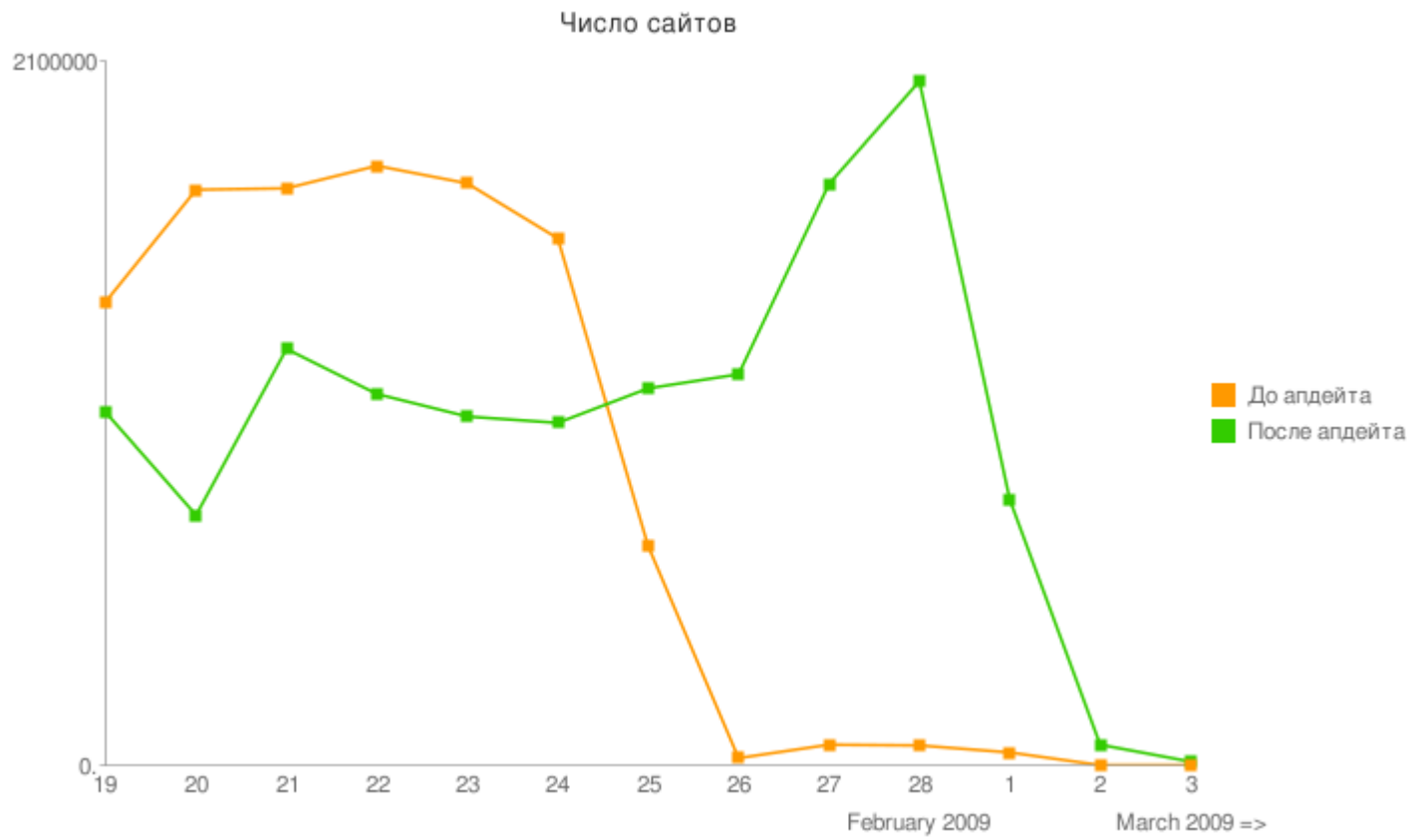
Поиск точной формы !слово





Апдейт 3 марта 2009

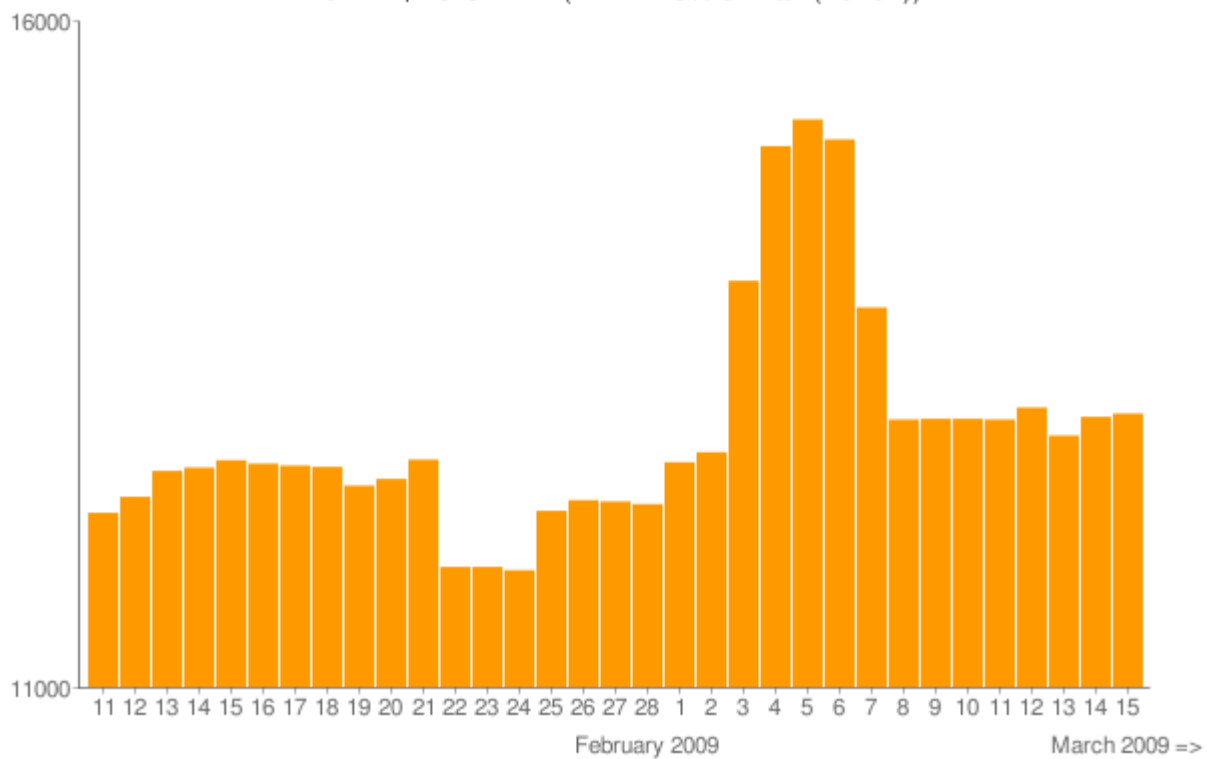
«Чистка индекса»





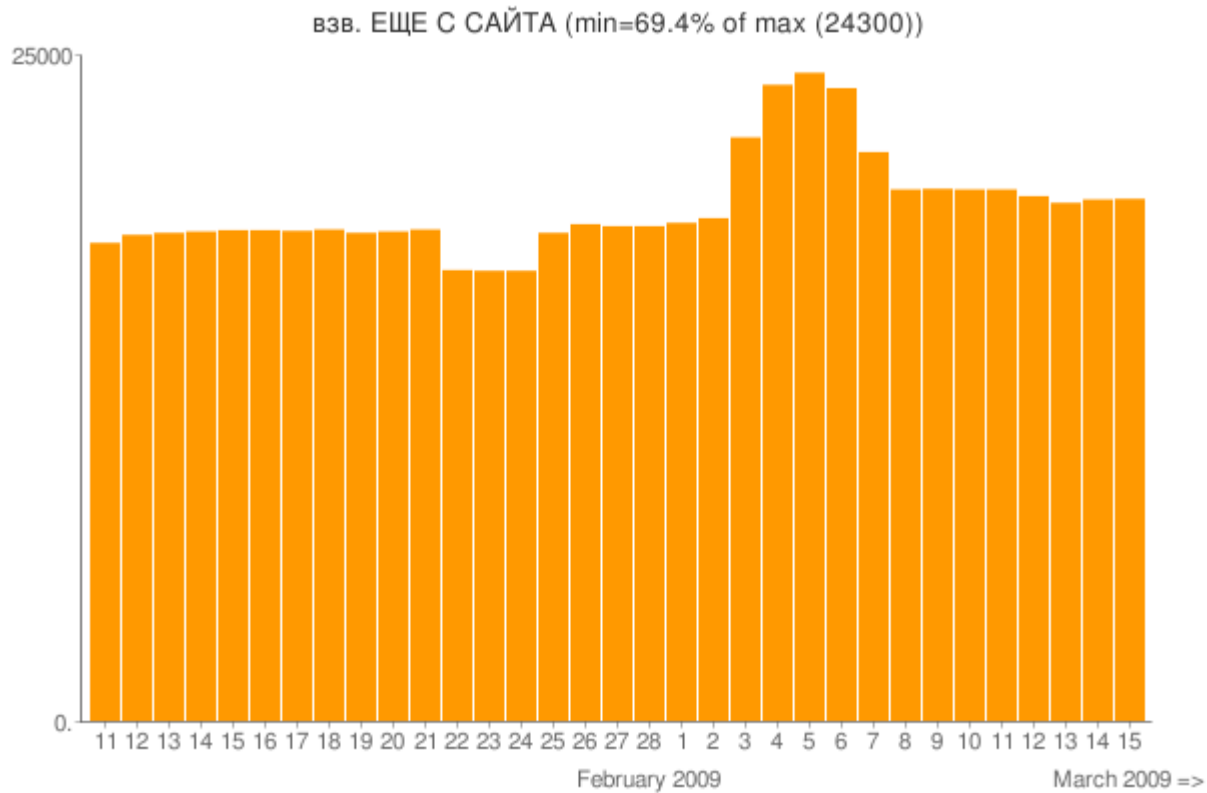
«еще с сайта»

взв. ЕЩЕ С САЙТА (min=77.8% of max (15254))



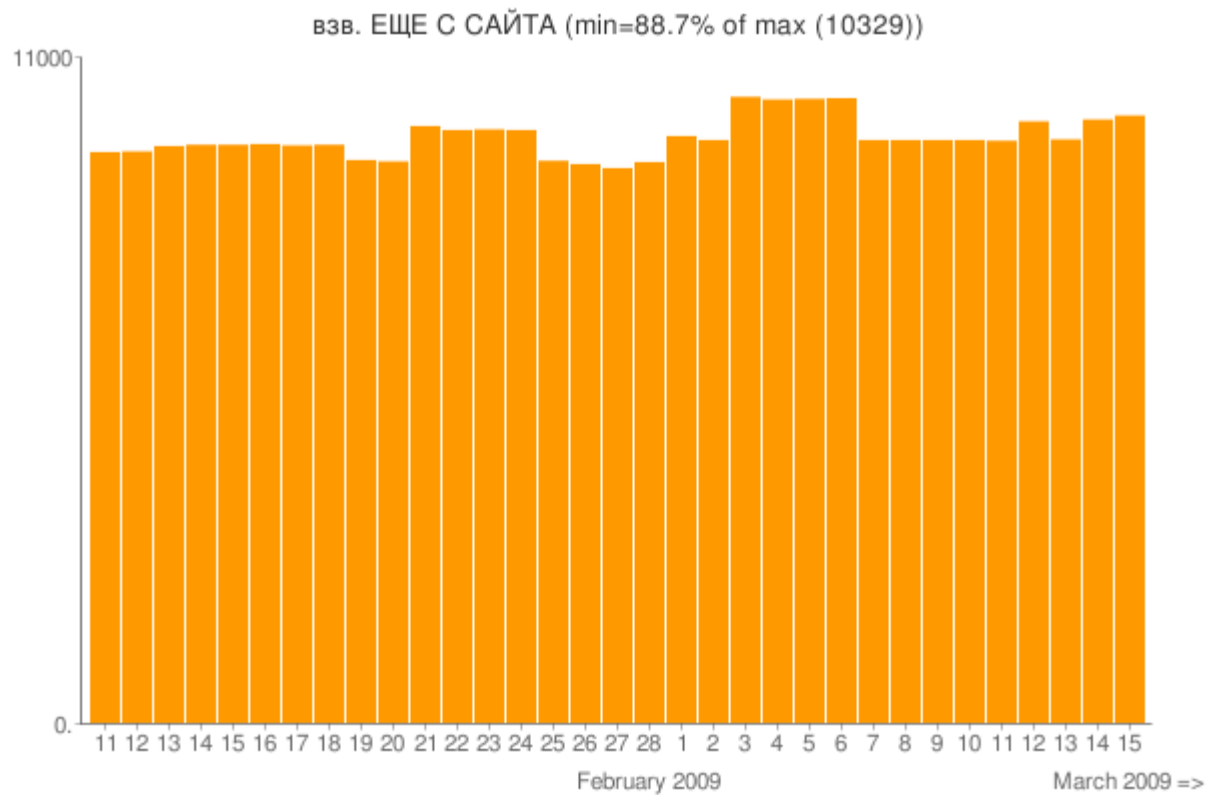


«еще с сайта» для однословных запросов



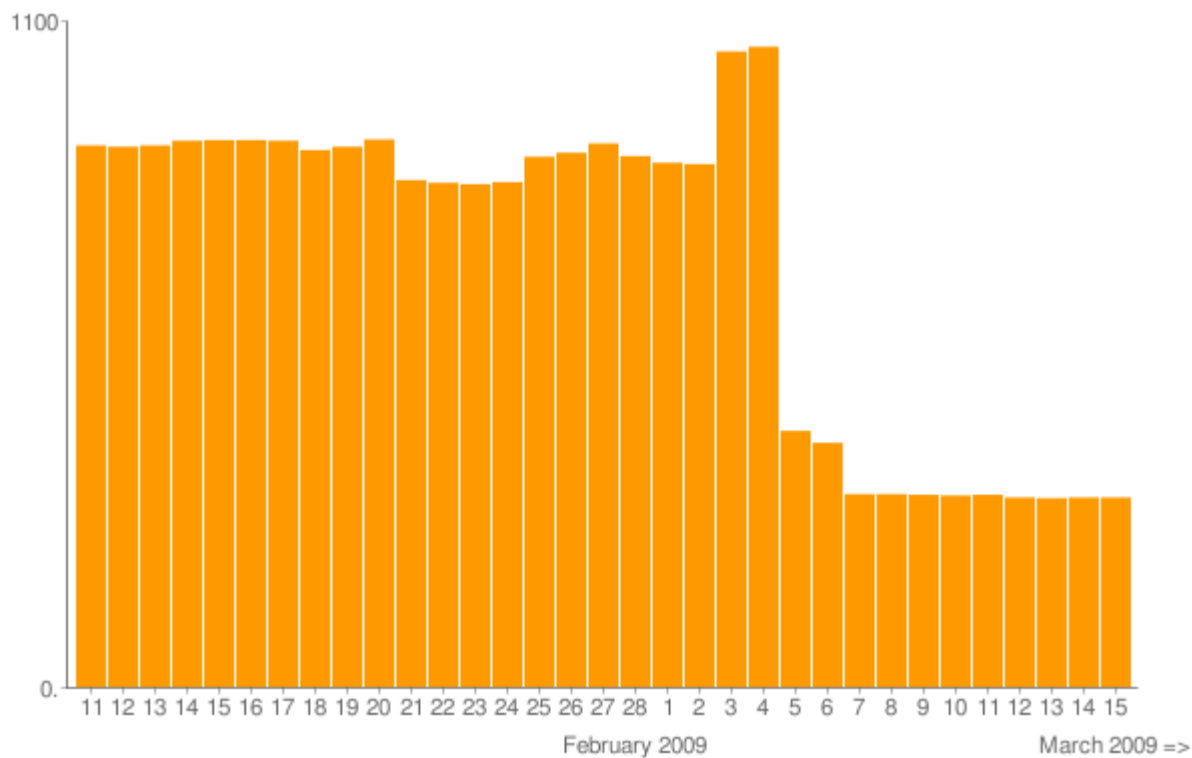


«еще с сайта» для длинных запросов



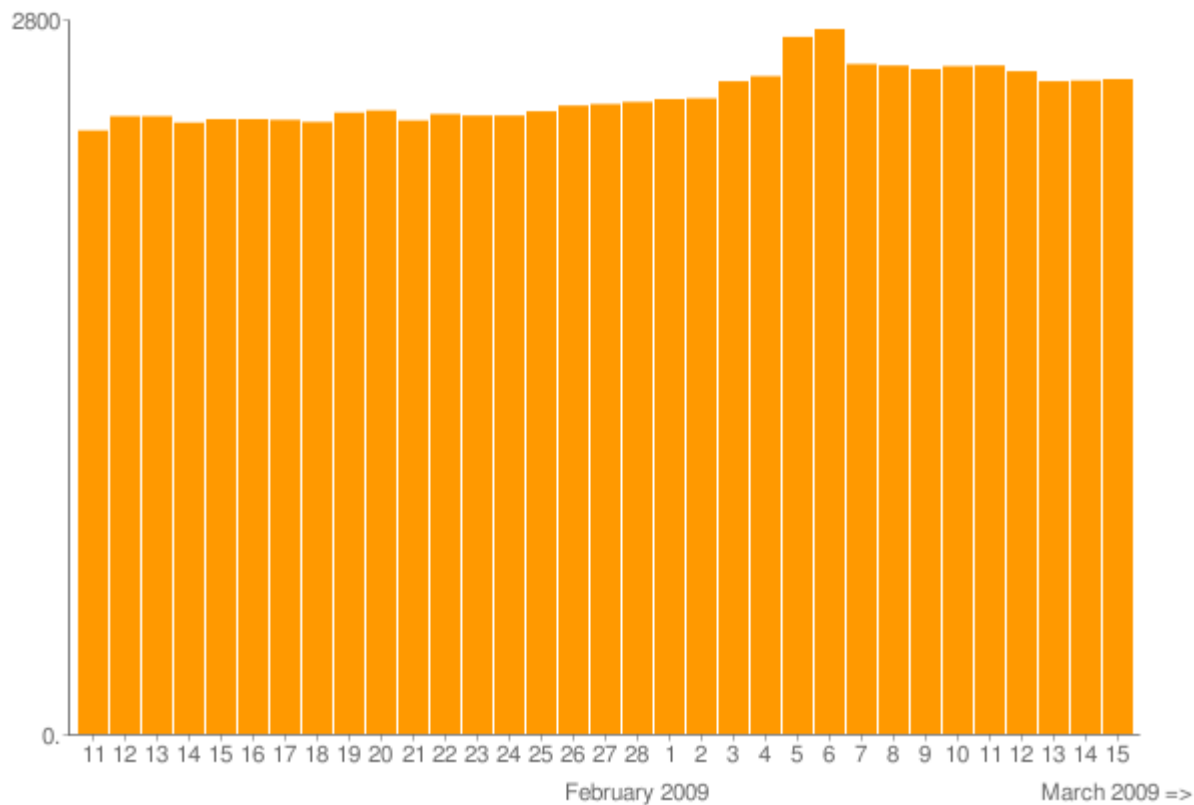


Взвешенный ТИЦ по сайтам, отсутствующим в ЯК -cat=(9000000)



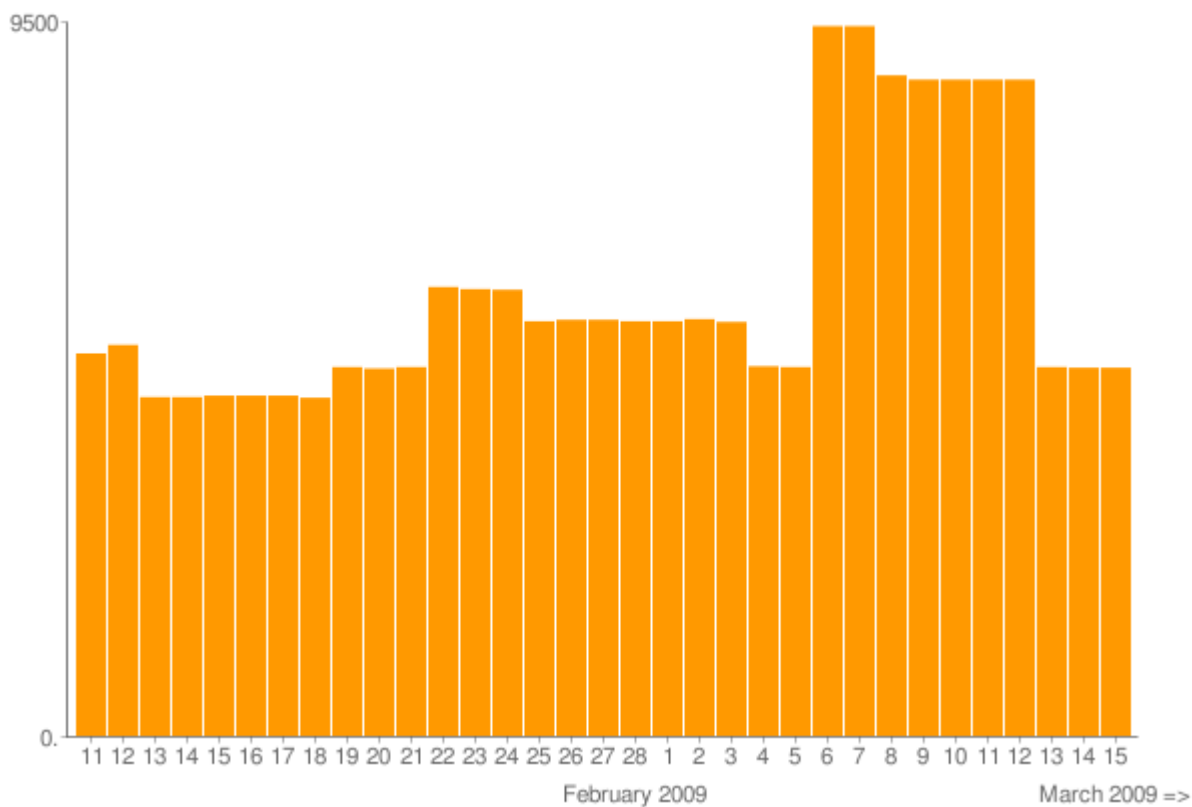


Взвешенный ТИЦ по сайтам из ЯК << cat=(9000000)





Взвешенный ТИЦ по «широким» запросам url, domain, rhost, www, lang, стопслова





Пока этот анализатор - игрушечный. В дальнейшем:

- Выбрать «правильные» параметры мониторинга
- Выбрать «правильные» группы запросов
- Увеличить число запросов в группах для хорошей статистики (от ~10 до ~300)

<http://tools.promosite.ru/monitoring/>



Использование особенностей языка запросов поиска Яндекса для исследований

Трофименко Е.А. Корпорация РБС, начальник отдела
исследований и аналитики

trofimenko.evgeny@rbscorp.ru, <http://www.bdbd.ru>