Переформулировки поисковых запросов в Яндексе

Трофименко Евгений
сЭо-эксперт
info@promosite.ru
http://tools.promosite.ru/

Я: Переформулировки-2004

Были обнаружены осенью 2004 года.
«Переколдовки» в URL сохраненной копии

&reqtext=(реальный::111 & запрос::222)//6

Использовались для:

1. Расширения запроса другими словами
2. Усиления влияния отдельных слов
3. Ограничения расстояний между словами
4. Установка контрастности слов

Примеры-2004

Расширение запроса другими словами
что такое ААА => ААА !%**это**::359 &/(-2 4) %**означает**::16316 &/(-2 4)
%**аббревиатура**::334021 &/(-2 4) %**расшифровывается**::183623

Усиление влияния отдельных слов
лоренциан => (лоренциан::2063133498 &/(0 0) !!%
лоренциан::2063133498)

Ограничение расстояний между словами
новый год => (новый::532 &/(-1 3) год::502)

Установка контрастности слов – двоеточечный вес.

В общем, все как и сейчас. Прожил reqtext до весны 2007.

Переформулировки -2008

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).

А также существенно ослаблены ограничения расстояний (поиск соседних слов в пределах документа)

Ограничения расстояний можно было подобрать перебором
(+слова +запроса) ~~ (+слова [ОПЕРАТОР] +запроса)

Однако сейчас это «вылечено».

Но зато появились подсказки в XML по опечаткам и варианты исправления исходного запроса.

Опечаточник в XML

Есть несколько типов опечаток: **Volapyuk, Undash** и др.

При комбинации нескольких вариантов опечаток «случайно» выдавался переформулированный запрос:
(fizi-olog) (поисковая оптимизация)
=>

```
<reask>
	<rule>Undash</rule>
	<source-text/>
	<text-to-show/>
	<text>(fizi::61543020-olog::1234567) ((поисковая::17483 ^
поисковик::65545) &&/(-32768 32768) (оптимизация::32653 ^
оптимизировать::95157 ^ оптимизироваться::4208069))
	</text>
</reask>
```

Евгений Трофименко info@promosite.ru

Выкачка

1. Генерим экспериментальный массив со всеми словами русского языка (было 5М)

2. Ищем другие воляпюки (20К)

=>

выкачиваем переформулировки по **1.3М запросов**

… нашли и закрыли. ☺

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

транслитерация слов

минск
минск::14882 ^ minsk::262094 ^ минский::86345

minsk
minsk::1082600 ^ !!минск::10130050

белоруссия
белоруссия::33069 ^ belorussia::8274691 ^ беларусь::10779 ^ беларусь::10779

трактор беларусь
(трактор::57459 ^ tracktor::1529019 ^ tractor::866803 ^ traktor::1341537 ^ тракторный::306947) &&/(-32768 32768) (беларусь::10779 ^ белоруссия::33069)

ограничения расстояний

10% запросов!

6300 nokia
6300::270856 **&/(-3 3)** nokia::12493

беларусь мтз
(беларусь::10779 ^ белоруссия::33069) &&/(-32768 32768) (мтз::334928 ^ mtz::9468156 ^ ((минский::86345 &/(1 1) тракторный::306947 &/(1 1) завод::9716)))

ндс беларусь
(ндс::15903 ^ ((налог::10340 &/(1 1) на::90 &/(1 1) добавленную::725 &/(1 1) стоимость::4415))) &&/(-32768 32768) (беларусь::10779 ^ белоруссия::33069)

работа с фрагментами слов

разбиение и склейка

автозапчасти минск
(автозапчасти::26701 ^ ((авто::2979-запчасти::7418)))
&&/(-32768 32768) (минск::14882 ^ minsk::262094 ^
минский::86345)

туроператор по белоруссии
(туроператор::80911 ^ ((тур::3736-оператор::9437))) &&/(-32768
32768) по::194 &&/(-32768 32768) (белоруссии::33069 ^
беларусь::10779 ^ беларусь::10779)

dsl 200
(dsl::91438 &/(-1 1) 200::4936) | dsl200::709103565

работа с фрагментами слов

Сколько бывает вариантов…

w200i

w200i::4958766
^ (!(w::1737 &/(1 1) 200::5303 &/(1 1) i::199))
^ ((w200::633693 &/(1 1) !i::199))
^ ((!w::1737 &/(1 1) 200i::23636785))

…все варианты разбиений буква-цифра

основное: расширения слов

белорусский металлургический завод

((белорусский::31308 ^ белорусско::996648) &&/(-32768 32768)
(металлургический::78749 ^ металлургия::78617)
&&/(-32768 32768)
(завод::9716 ^ заводик::2584483)) ^ !бмз::4579945

белорусский квн

(белорусский::31308 ^ белорусско::996648) &&/(-32768 32768)
(квн::77388 ^ kvn::935126 ^ ((клуб::2716 &/(1 1) веселых::15270
&/(1 1) и::54 &/(1 1) находчивых::1102674)))

курьезы переформулировок

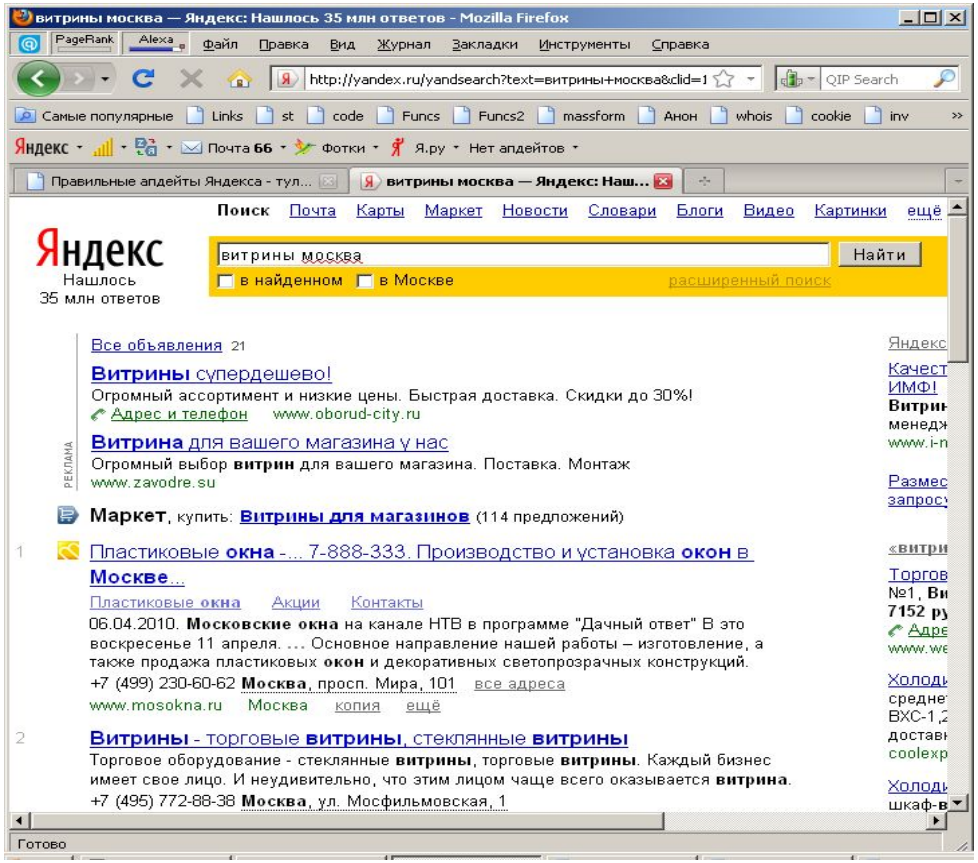
вряд ли только машины работают:

партия единая россия

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

витрина – это ведь окно?

Да… взгляд с той стороны витрины 😊



Евгений Трофименко info@promosite.ru

ФИО – новые зоны и термы

!!! Экстракция сущностей в большом поиске !!!

Для запросов, содержащих имена в виде 2+ слов

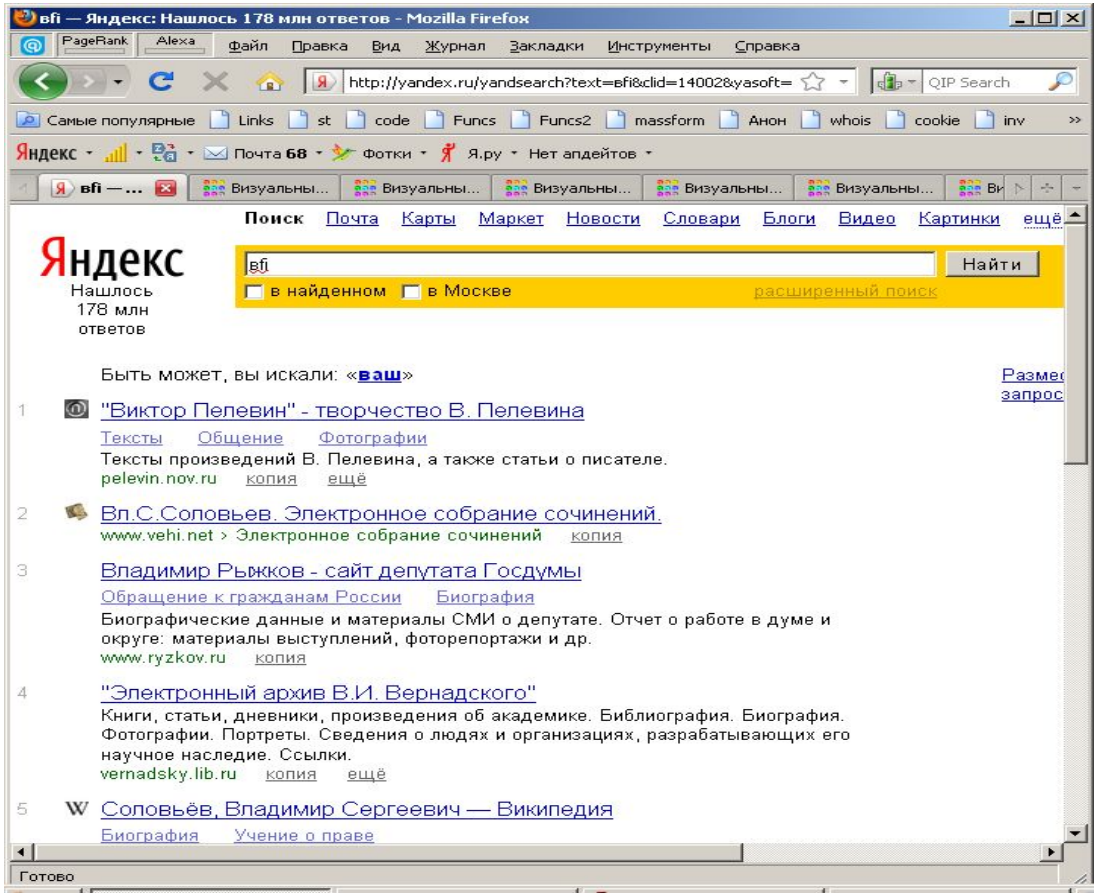
вася пупкин

Переформулируется с фрагментом
*** (
fioname[((**васяfi**::332552 &&/(-32768 32768) !!пупкин::901729))] |
fiinname[((**вfi**::1574 &&/(-32768 32768) !!пупкин::901729))] |
fiinoinname[((**вfi**::1574 &&/(-32768 32768) !!пупкин::901729))] |
finame[((**васяfi**::332552 &&/(-32768 32768) !!пупкин::901729))]
)

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

bfi – все имена на букву B

экстракция объектов из текста…



Евгений Трофименко info@promosite.ru

поиск по зоне ФИО?

Не очень нужные операторы

fioname[ ]
fiinname[ ]
fiinoinname[ ]
finame[ ]

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

И ведь без микроформатов и разметки…

Евгений Трофименко info@promosite.ru

Оператор [^]

Похож на %

Доп.слово не обязательное

Для запроса вида **слово1 ^ слово2**

Слово1 обязательно находится, ему приоритет
Слово2 не обязано находиться.

окна ^ мебель – окна первые
мебель ^ окна – мебель первая
domain:root ^ мебель ^ окна –окна выше!

А для [%]

Не совсем похоже:

Для запроса вида **слово1 % слово2**

Слово1 обязательно находится
Слово1 и Слово2, похоже, равноправны в смысле ранжирования

окна %мебель – (окна+мебель) первые
мебель %окна – (окна+мебель) первые
domain:root %мебель %окна - (окна+мебель)

И выдачи похожие.

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

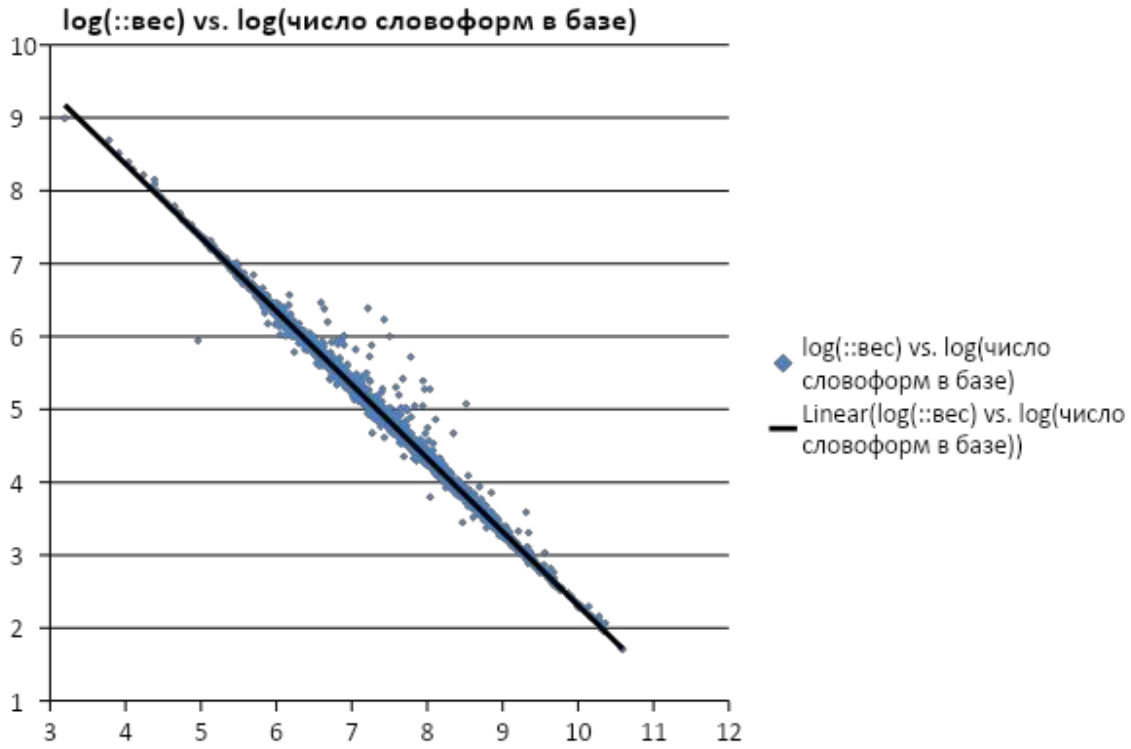
Догадываемся - **::вес=D/Di**
Это отношение числа документов.
Чтобы получить IDF, берем логарифм:

=> IDF=log(::вес)

Евгений Трофименко info@promosite.ru

::веса -не документные?

А от словоформ? Не IDF, а ICF?



::веса по разным коллекциям

веса разные по RU, EN, UK коллекциям

Слово **fizi** присутствовало во всех запросах.
Оно имело разные веса в 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.

Итого 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

это частично внедрено в сервис http://tools.promosite.ru/

Вопросы?

Переформулировки поисковых запросов в Яндексе

Трофименко Евгений

сЭо-эксперт

info@promosite.ru

http://tools.promosite.ru/