

ИСПОЛЬЗОВАНИЕ ОСОБЕННОСТЕЙ ЯЗЫКА ЗАПРОСОВ ПОИСКА ЯНДЕКСА ДЛЯ ИССЛЕДОВАНИЙ.

Трофименко Е.А.
начальник отдела исследований и аналитики
Корпорации РБС

Введение

Яндекс пока что наиболее популярный поисковик в рунете и более других открыт к исследованиям его алгоритмов. Эксперименты в SEO важно не только провести, но и правильно интерпретировать, а в реальности в формировании поисковых результатов вмешивается слишком много факторов. В данной работе рассмотрены особенности работы поиска по текстам ссылок, возможности для изучения трактовки Яндексом многозначных запросов и их расширения, а также элементы учета текстовой релевантности — основываясь только на результатах выдачи, без создания экспериментальных страниц.

Особенности поиска по ссылкам

Как известно, поиск производится по текстам ссылок и текстам страниц.

Если в содержимом документа содержится достаточное количество слов запроса, документ отображается как найденный по текстам, с включением фрагментов документа в сниппет — описание результата выдачи.

Если же документ не содержит необходимых слов запроса или считается не проиндексированным (временно отсутствующим на сервере, выдающем ошибку, выпавшим из базы Яндекса, запрещенным в файле robots.txt, новым для робота Яндекса), но слова запроса на этот документ содержатся в тексте ссылок, то он может находиться с надписью «найден по ссылке» (т.н. НПС) (кроме случая, когда сайт находится в Яндекс-каталоге и НПС-слова есть в описании сайта — сниппет базируется на описании сайта из ЯК).

Некоторые операторы поиска Яндекса по-разному применяются к текстам ссылок и текстам страниц.

Оператор «-» (минус) — поиск НПС-результатов

Для целей исследования часто возникает необходимость найти НПС-результаты с нужными словами в ссылках, например, для изучения особенностей ранжирования по редким и частотным словам. Как бы то ни было, а найти в собственных базах такое не всегда возможно.

Оператор «-» (минус) является недокументированным [1]. Документированный оператор исключения «~~» (двойная тильда) работает в пределах документа, а «минус» всегда можно было использовать для исключения слов в пределах контекстных ограничений. Сейчас в описании языка есть оператор «~» (одинарная тильда) — он заявлен как оператор исключения в пределах предложения, и, похоже, не работает.

Так или иначе, а оператор «-» (минус) работает только в применении к контенту документа, т.е., к проиндексированным текстам. К текстам ссылок он не применяется.

Поэтому можно задать запрос вида **слово-слово** и найти все результаты, которые найдены по ссылкам или по текстам, содержащим **слово**, и все найденные по текстам результаты будут убраны из выдачи — тогда останутся только НПС результаты по **слову** (исключение — сайты из Яндекс-каталога, но и их можно убрать добавлением **-cat=9000000**).

Таким образом, довольно просто можно отыскивать НПС-результаты с интересующими нас словами в тексте ссылок.

Примеры:

Запрос	Число найденных страниц
сайт -сайт	780 тыс.
порно -порно	29 тыс.
интернет -интернет	470 тыс.
продвижение -продвижение	12 тыс.
реферат -реферат	76 тыс.
москва -москва -московский -cat=9000000	230 тыс
секс -секс	51 тыс.
ашманов -ашманов	137 страниц
трофименко -трофименко	114 страниц

Операторы «~» и «~~» (тильда) — исключаем НПС

При использовании операторов исключения «~» и «~~», напротив, можно избавиться от присутствия НПС-результатов в выдаче. Даже если мы пытаемся исключить из документа

заведомо отсутствующую в нем абракадабру, найденные по ссылке результаты исчезают из выдачи. Например:

Сайт /{(1 1) производителя /{(1 1) Canon /{(1 1) Inc — на 1 месте www.canon.ru НПС, всего 19 результатов.

Сайт /{(1 1) производителя /{(1 1) Canon /{(1 1) Inc ~~jcaenhcqawv — найдено 18 страниц, НПС результат пропадает.

По-видимому, это происходит из-за того, что алгоритм поиска (и выдачи НПС) по тексту ссылок принципиально не принимает во внимание контент страницы. Т.е., алгоритм, выдавая НПС, как бы не знает ничего о содержимом документа, хотя документ практически всегда проиндексирован, и Яндекс знает, что исключенной абракадабры там нет. Но в реальности он делает вид, что не знает этого, и при исключении абракадабры убирает НПС-результаты «на всякий случай». Это относится именно к тому случаю, когда в тексте документа нет фразы запроса.

С помощью этого способа можно оценивать долю НПС-результатов в выдаче. Например:

Запрос	Найдено страниц по точному запросу	Найдено страниц по запросу с добавлением ~~sfjhdssnv	Доля НПС-страниц в выдаче по исходному запросу
авто	280 млн	279 млн	~0.35%
авто москва	124 млн	118 млн	~4.8%
купить авто москва	67 млн	58 млн	~13.4%
купить поддержанное авто москва	4 млн	3 млн	~25%

Общая тенденция — для длинных поисковых запросов доля НПС-результатов в выдаче выше. Это нормальная ситуация, и ее можно использовать для оценки степени коммерциализированности низкочастотных запросов. Так как при оптимизации на страницах появляются целевые слова и фразы, они уже не будут находиться как НПС. Логично считать, что с НПС-результатами не работали оптимизаторы, следовательно, и конкуренция по этим запросам должна быть ниже.

Для детальных исследований более точные данные (количество найденных сайтов и страниц) можно получить с использованием Яндекс.XML.

Особенности расширения пользовательских запросов

Летом 2008 года Яндекс сообщил [4] об изменениях в алгоритме переформулировки запросов, которые предполагают сильные изменения, включающие добавление новых слов в запрос:

Одну и ту же поисковую потребность пользователь может выразить разными запросами. Например, запрос «гамбургские гостиницы» кажется эквивалентным по смыслу запросу «гостиницы Гамбурга».

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).

Как именно происходит эта переколдовка запроса? По смыслу, некоторые слова могут быть заменены (в случае замены транслитерации или аббревиатуры), в некоторых случаях запрос может быть расширен с переходом между частями речи (недвижимость в Москве -> московская недвижимость).

Поскольку эти изменения затрагивают большую часть коммерческих запросов, важно знать точно, как это происходит. Например, может оказаться, что дополненный/измененный запрос — менее конкурентный и легкий для продвижения.

Для того чтобы исследовать особенности переформулировок, используем операторы исключения. Например, из результатов поиска по запросу **мазда** исключаем все результаты, содержащие слово **мазда**. Получаем все результаты с переформулировкой — новыми словами и частями речи. Последовательно исключая различные переформулировки, можно найти их все.

Исключение оператором «~~» уже не работает

Еще в сентябре 2008 особенность оператора исключения состояла в том, что исключалась форма слова во всевозможных падежах и числах, но без переформулировок. В результате,

задавая запрос **мазда ~~мазда** можно было получить результаты, содержащие **mazda** — переформулировку.

Аналогично, запрос **недвижимость в москве** расширялся с использованием слов **московская, московский** [sic!], **московск** вместо **москвы**. Не все из этих слов подсвечивались в выдаче, что говорит об использовании разных алгоритмов в основном поиске и в подсветке выдачи. **Ауди** переводилось как **audi, audis, аудь**. Все это приводится по памяти.

Однако на момент написания этого доклада оператор «~~» уже исключает все формы слова, включая переформулировки. В результате выдача по **мазда ~~мазда** пустая.

Исключение оператором «-» работает, но неудобно

Оператор «минус» исправно исключает выбранную форму слова, оставляя переформулировки (**мазда -мазда** показывает результаты с **mazda**). Однако неудобство состоит в том, что он не работает по текстам ссылок, и очень часто вместо тех текстовых результатов выдачи, которые содержат интересующие нас новые переформулировки, мы получаем выдачу, состоящую из НПС-результатов (видимо, достаточно релевантных).

Исключение точной формы (!) оператором «~~» неудобно, но можно

Выход — использование оператора «!» (восклицательный знак) для обозначения точной формы слова, которая исключается оператором «~~». В этом случае мы вынуждены записывать полный список всех возможных форм слова для исключения.

Например, **мазда ~~!(мазда|мазду|мазде|мазд)** дает искомые результаты с **mazda**. Кстати, операторы исключения не применяются к поиску по URL, поэтому в любом случае выдача может содержать результаты с переформулировкой в URL найденной страницы.

Однако иногда этот оператор не позволяет посмотреть все дополнительные слова в расширениях запроса (исключает доп. слова), лучше для страховки проверять его оператором «минус».

Переформулировки запросов

Из особенностей переформулирования запросов можно отметить, что при исследовании транслитерации в основном находятся односторонние переформулировки: русскоязычным «мазда, ауди» соответствуют англоязычные варианты, но обратные переформулировки таким способом не находятся.

Исходный запрос	Новые слова в запросе
ауди	audi, audy
mazda	mazdae
ноутбуки в СПб	Петербург
недвижимость москвы	Московская
квартиры в москве	Московская, москві
МГУ	Московский, mgu, МГОУ, mgou
гамбургский счет	Гамбург

Однако обратные переформулировки есть: по запросам **mazda, audi** в выдаче подсвечиваются русскоязычные слова **мазда, ауди**. Вероятно, это связано с выбором главного слова из набора «синонимов» и с особенностями работы операторов исключения, которые постоянно меняются.

В любом случае, самым надежным методом для исследования переформулировок было бы создание специальных страниц, содержащих, в том числе, и разные части речи, с изучением подсветки разных терминов при поиске интересующих запросов в пределах сайта.

Особенности переформулировки запросов («колдунщика»)

Колдунщик – механизм переформулировки запросов, известный с 2004 года и описанный авторами алгоритма [3]. Перед исполнением запрос пользователя изменялся специальным образом, индивидуальным для каждого запроса. Если исключить редкие случаи, когда в запрос добавляются «новые» слова или усиливается влияние отдельных слов, то работу колдунщика можно было свести к расстановке между словами «расстояний» — операторов контекстных ограничений и отмене влияния стоп-слов.

Расстояния между словами [3] важны для выбора последовательностей слов (пассажей), которые находятся в документе не слишком далеко друг от друга и будут учтены в дальнейшем. При этом «не слишком далеко» для каждой пары слов в запросе было индивидуальным — расстояния изменялись от «в пределах документа» до «строго по порядку».

Дальнейшая оценка релевантности документа запросу производилась с учетом выбранных пассажей и «весов» слов.

Стандартные переколдовки запросов

Почти все используемые в колдунщике операторы были описаны в [1,2] и представляли собой обычные операторы ограничения расстояния между словами.

Хотя эти расстояния теоретически могли быть любыми, реально использовался набор из «стандартных» 7 расстояний. Примерная относительная «популярность» этих операторов приведена ниже – по массиву запросов клиентских сайтов Корпорации РБС [6]

Оператор	Относительная частотность	Смысл оператора: поиск в пределах
&	100%	в пределах одного предложения
&/(-2 4)	9%	в пределах -2 +4 соседних слов
&/(-1 3)	10%	в пределах -1 +3 соседних слов
&/(1 1)	2%	строго по порядку
&/(0 0)	0.035%	усиление влияния одного слова
&&/(-7 7)	15%	в пределах 7 предложений
&&/(-3 3)	15%	в пределах 3 предложений
&&	7%	в пределах документа

В середине октября 2007 года Яндекс отменил показ переколдовки в ссылке на сохраненную копию в результатах поиска [7], однако она все еще работала, и отработка операторов отключена не была. Реальные расстояния между словами можно было подобрать – используя 7 «любимых» Яндексом вариантов с помощью перебора можно было найти тот вариант, для которого выдача по «заколдованному вручную» запросу совпадала с выдачей Яндекса.

Так продолжалось недолго.

Новые аспекты переколдовки

Весной 2008 в релизе «Магадан» разработчики сообщили [5], что

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.

Наиболее заметных улучшений мы смогли достичь в обработке многословных запросов.

И действительно, отработка запросов с подставленными «любимыми» операторами перестала давать совпадения с реальной выдачей. Реальная отработка запросов Яндексом изменилась.

Возьмем запрос **(+скачать +реферат)** и исключим оператором «~» из результатов поиска все страницы, где два слова запроса находятся на некотором расстоянии друг к другу (от одного слова до всего документа):

Запрос	Чистим в пределах	Найдено страниц	Доля, %
(+скачать +реферат)	—	16065797	100.0
(+скачать +реферат) ~~(+скачать /(-1 1) +реферат)	1 слова	12004783	74.7
(+скачать +реферат) ~~(+скачать /(-2 2) +реферат)	2 слов	10743380	66.9
(+скачать +реферат) ~~(+скачать /(-3 3) +реферат)	3 слов	10588526	65.9
(+скачать +реферат) ~~(+скачать /(-10 10) +реферат)	10 слов	10077579	62.7
(+скачать +реферат) ~~(+скачать /(-20 20) +реферат)	20 слов	9951594	61.9
(+скачать +реферат) ~~(+скачать & +реферат)	одного предложения	9920305	61.7
(+скачать +реферат) ~~(+скачать &&/(-1 1) +реферат)	соседних предложений	8108090	50.5
(+скачать +реферат) ~~(+скачать &&/(-2 2) +реферат)	2 предложений	6032021	37.5
(+скачать +реферат) ~~(+скачать &&/(-3 3) +реферат)	3 предложений	5401261	33.6
(+скачать +реферат) ~~(+скачать &&/(-5 5) +реферат)	5 предложений	4273921	26.6
(+скачать +реферат) ~~(+скачать &&/(-10 10) +реферат)	10 предложений	3089645	19.2
(+скачать +реферат) ~~(+скачать &&/(-20 20) +реферат)	20 предложений	2326525	14.5
(+скачать +реферат) ~~(+скачать &&/(-30 30) +реферат)	30 предложений	1754794	10.9
(+скачать +реферат) ~~(+скачать &&/(-100 100) +реферат)	100 предложений	635005	4.0
(+скачать +реферат) ~~(+скачать &&/(-1000 1000) +реферат)	1000 предложений	9515	0.1
(+скачать +реферат) ~~(+скачать &&/(-10000 10000) +реферат)	10000 предложений	210	0.0
(+скачать +реферат) ~~(+скачать && +реферат)	документа	0	0.0

Данные получены с использованием Яндекс.XML

Мы видим, что при увеличении расстояния между словами, которые мы вычищаем из выдачи, число найденных страниц уменьшается постепенно от 16 миллионов до нуля. Это означает, что запросу **+скачать +реферат** соответствуют документы с очень большим расстоянием между словами в том числе, и именно эти документы остаются в выдаче после исключения «близких» пар слов. Если бы запрос скрытно от внешнего мира «переколдовывался» стандартным образом, то установленные контекстные ограничения ограничивали бы расстояние «сверху», и не позволяли бы находиться парам, в которых слова находятся далеко друг от друга.

В этом случае в какой-то момент при увеличении расстояния в приведенных запросах число найденных документов должно было в какой-то момент резко обнулиться. Именно это расстояние и соответствовало бы «новой переколдовке».

Однако по этому запросу находятся даже документы, в которых слова «скачать» и «реферат» находятся на очень больших расстояниях друг от друга. Больше того, если мы не будем использовать «плюсик» в исходном запросе, например:

(скачать реферат) ~~(+скачать && +реферат) , мы обнаружим, что по такому запросу находится **283204** документов, содержащих только одно из слов запроса.

Все это полностью согласуется с сообщением разработчиков об изменениях, позволяющих находить документы, «которые содержат слова запроса далеко друг от друга». И косвенно свидетельствует о том, что переколдовка с расстановкой операторов между словами уже не используется. Вместо этого актуален алгоритм, в котором слова ищутся в пределах всего документа, но в зависимости от расстояния между словами вносят разный вклад в конечную общую релевантность.

Заключение

Углубление знаний о наиболее популярном поисковике рунета (в данный момент — Яндексe☺) невозможно без постановки экспериментов — это ясно даже и ежу. В отличие от других поисковиков, Яндекс позволяет проводить исследования и без создания специальных экспериментальных страниц. Надеюсь, что приведенные примеры были полезны для отбора объектов эксперимента (НПС-результатов), дали возможность более детального исследования переформулировок отдельных клиентских запросов и уровня конкуренции по ним, позволили отказаться от более неактуальных представлений об алгоритме Яндекса.

Литература

1. Синтаксис языка запросов Яндекса <http://help.yandex.ru/search/?id=481939>
2. FAQ поиска Яндекса <http://help.yandex.ru/search/?id=481938#context>
3. Илья Сегалович, Михаил Маслов: Яндекс на РОМИР-2004. Некоторые аспекты полнотекстового поиска и ранжирования в Яндекс http://romip.ru/romip2004/07_yandex.pdf
4. Расширение запросов - переходы из одной части речи в другую, транслитерация, аббревиатуры http://webmaster.ya.ru/replies.xml?item_no=1030
5. О смягчении фильтрации найденных документов. Подлетая к "Магадану" http://webmaster.ya.ru/replies.xml?item_no=645
6. Статистика по операторам колдунщика Яндекса <http://blog.promosite.ru/comments.php?533>
7. Отмена показа переколдовки <http://forum.searchengines.ru/showthread.php?t=173853>

Корпорация РБС

115191, Россия, Москва, ул. Б. Тульская, д. 10, стр. 9

Телефон: (495) 232–05–91, ICQ-консультант: 377–169–437

<http://www.bdbd.ru> | <http://www.rbsgroup.ru> | <http://www.webprofy.ru> | <http://www.mediaguru.ru>

