Google, Rambler и Yandex

Искать по-русски

Для поиска в Интернете создан не один десяток поисковых систем. Все они работают по-разному и на один и тот же запрос выдают неодинаковые результаты. Но при помощи несложных прикладных средств найти нужную информацию не составит труда.

Существует два принципиально разных подхода к целенаправленному сбору информации о содержании страниц в Интернете: каталогизация и индексация. Каталог — это список сайтов, разделенный по категориям. Составляют такой список не специальные программы-роботы, а живые люди. Они просматривают содержимое сайтов, решают, к какой теме относится тот или иной ресурс, составляют к нему описание и помещают в соответствующую рубрику. Объем ссылок, который получается обработать таким образом, составляет в лучшем случае единицы процентов от всех сайтов Сети, но плюс ручной обработки в том, что сайт будет точно находиться в рубрике, которая подходит ему по содержанию.

Индексацию же проводит робот. Он обходит существующие ссылки на документы и создает так называемый индекс. Информация о содержимом документа и о самом документе записывается в базу данных поисковой машины в специальном формате. Уже на этом этапе могут происходить некоторые дополнительные преобразования данных, например приведение слов к единой форме — нормализация. То есть глаголы приводятся к неопределенной форме, существительные — к именительному падежу и единственному числу и т. д. Это дает возможность в дальнейшем осуществлять поиск не только по точному написанию слова, но и по различным его формам. Обработанные страницы сохраняются в кеше и связываются с проиндексированным документом. Позже пользователь сможет получить документ в том виде, каким он был в момент индексации. Это полезно, если страница по каким-либо причинам станет недоступной или ее содержимое изменится. »

Объем данных, который приходится обрабатывать для построения индекса, огромен. Поэтому поисковая машина представляет собой большой программно-аппаратный комплекс. Различными этапами обработки информации занимаются разные серверы. Одна их группа отвечает за скачивание страниц, другая — за построение части индекса, третья — за объединение индексов в единую базу, четвертая — собственно модуль поиска.

Запрос

Когда пользователь отправляет свой запрос поисковой системе, на основе построенного индекса из базы выбираются документы, содержащие запрошенные слова, и выдаются в качестве результата поиска в определенном порядке. В большинстве случаев требуемые слова содержатся во множестве документов, поэтому очень важен порядок, в котором пользователь увидит результаты. Если документ, который содержит лучший ответ на вопрос пользователя, будет отображен на тридцатой странице результатов, то это равнозначно тому, что он вообще не будет найден, потому что до просмотра тридцатой страницы вряд ли кто-то доберется.

Характеристика страницы, отражающая мнение поисковой машины, насколько она удовлетворяет запросу пользователя, называется релевантностью. Для того чтобы наиболее релевантные документы были размещены на первых страницах выданного результата, применяется система присвоения веса каждому документу в зависимости от множества факторов. Каждая поисковая машина имеет свое собственное секретное ноу-хау в подсчете релевантности. Чем чаще искомые слова встречаются в документе, тем обычно больше его вес. (Например, в борьбе со спамом поисковая машина может следить за чрезмерным употреблением одного и того же слова на странице и понижать вес за резкое отклонение от среднестатистических показателей.) Также вес документа часто увеличивается, если ключевые слова присутствуют в заголовке (часть документа, выделенная тегом <title>) или являются названиями разделов (присутствуют в тегах <H1>...<H7>). Логическое усиление слов — выделение тегами <b>, <u> и — также может увеличивать весомость документа.

▲ Дополнительные возможности предлагает Google services

▲ Первым Google.ru нашел официальный сайт Chip Украина

Помимо этого, многие поисковые машины учитывают еще и взаимное расположение слов. То есть, если в найденном тексте слова расположены в том же порядке, что и в запросе, то документ получит больший вес. Кроме того, может учитываться расстояние между словами. Документ, в котором слова расположены подряд или в одном предложении, имеет больший вес, чем документ, содержащий все те же слова, но рассредоточенные по всей странице.

Вес каждого документа может зависеть не только непосредственно от его содержания, но и от так называемой авторитетности. То есть документы, на которые ведет большее количество ссылок, имеют больший вес. Также значимости добавляют ссылки с наиболее весомых страниц (так работает технология PageRank от Google).

Истории поисковых систем

Yandex

Разработка системы Yandex как алгоритма поиска в текстовых документах началась в начале девяностых годов прошлого века.

В 1993 году родилось слово «Яndex», еще никак не связанное с поиском в Интернете. Придумал его Илья Сегалович, один из главных разработчиков поискового механизма, сейчас — технический директор компании Yandex. Изначально «Яndex» означало «Языковой index», или, по программистской традиции, «yandex» — «Yet Another indexer», как, говорят, «yahoo» — это в том числе «Yet Another Hierarchicall Organized Oracle». Позже была разработана технология, позволяющая осуществлять поиск с учетом морфологии русского языка. До 1996 года на основе существующей технологии создавались прикладные программы для поиска в различных справочниках и текстовых массивах (например, Библии).

В 1996 году добавилась возможность строить гипотезы о морфологии слова. То есть, даже если слово не содержится в словаре, система в состоянии предположить, как выглядят различные формы этого слова.

21 ноября 1996 года впервые была установлена система Яndex.Site — система полнотекстового поиска на веб-сервере.

В апреле 1997 года на сайте yandex.ru заработала система поиска по русскому Интернету. Основные разработчики — Сергей Ильинский, Михаил Маслов, Илья Сегалович, Дмитрий Тейблюм — до сих пор работают в компании Yandex.

В марте 2001 года Yandex стал лауреатом Национальной Интел интернет-премии сразу в нескольких номинациях. В 2003 году Yandex научился искать документы в форматах RTF, PDF и DOC. На сегодняшний день он хранит информацию о более чем 150 миллионах документов, что составляет больше 4000 Гбайт.

» Разные поисковые системы используют различные алгоритмы и формулы для вычисления веса и различные способы сопоставления всех этих факторов. Поэтому релевантность документов оценивается по-разному. То есть один и тот же запрос к разным поисковым системам даст разные результаты.

Синтаксис языка запросов

Хотя расширенный запрос и предназначен для уточнения критериев поиска, полностью настраиваемый поиск можно обеспечить с помощью применения языка запросов. Язык запросов — это специальные символы и операторы, которые пишутся в ту же строку для поиска, что и ключевые слова, и обрабатываются поисковой машиной. Google, Yandex и Rambler имеют сходство в применении некоторых специальных символов.

Строка, заключенная в кавычки, будет найдена именно в том виде, что и в запросе — слова расположены в том же порядке и находятся в той же форме. Символ «+» перед словом говорит о том, что слово должно обязательно присутствовать в найденных документах. На самом деле по умолчанию между всеми словами и так подразумевается логический оператор «И», то есть будут найдены документы, которые содержат одновременно все слова из запроса. Поэтому символ «+» имеет смысл для так называемых «стоп-слов». Это такие слова, которые часто встречаются в текстах и вряд ли могут являться критерием для поиска. Например, предлоги, союзы, местоимения, артикли и т. п. Противоположное значение имеет символ «-». Слово, которому предшествует этот знак, не должно попадаться в документе. В Rambler вместо «-» используется знак «!». Исключение слов — очень простой, но полезный прием, позволяющий сразу отсеять множество документов, которые точно не подходят.

▲ Расширенный поиск в Yandex: язык, дата, формат

Иногда можно использовать логическое «ИЛИ». В Google оно выглядит как «OR». В Yandex и Rambler — как символ «|». Также в Yandex и Rambler можно строить запросы с применением скобок и оператора логического сложения «&». К примеру, запрос «(фотография | фото | фотоснимок) & (тигр | носорог)» выдаст страницы с фото какого-либо из двух животных. Yandex оператор «&» указывает на то, что слова должны находиться в одном предложении. Rambler же достаточно, чтобы они просто присутствовали в документе. Для того чтобы Yandex искал слова по всему документу, нужно использовать оператор «&&».

Также Yandex и Rambler позволяют указать расстояние между искомыми словами в предложении. В Rambler для этого используется конструкция '(число, запрос)', где число — это расстояние между словами, представленными в запросе, измеряемое в словах. В Yandex используется конструкция вида «/(n m)», где n и m — расстояние назад и вперед в словах между ключевыми выражениями. Кроме того, можно применять упрощенную конструкцию — «/n» — или указывать расстояние не в словах, а в предложениях — «&&/(n m)».

Yandex отличается чувствительностью к регистру букв. Если в запросе присутствует слово, написанное со строчной буквы, то будут найдены документы, где это слово написано как со строчной, так и с прописной. Если же в запросе содержится слово, написанное с прописной буквы, то будут найдены только слова, начинающиеся с прописной (если это слово не первое в предложении).

Для исключения слов в пределах предложения служит оператор «~», в пределах документа — «~~» (то есть «~~» эквивалентно «-»). Для поиска точной формы слова (без учета морфологии) нужно поставить перед ним «!». При помощи операторов «$» и «#» можно, как и в расширенном поиске, задать зону поиска (заголовок документа или текст ссылки) или элемент документа (описание картинки, ключевое слово и т. д.). Кроме того, у Yandex существует возможность влиять на ранжирование результатов. »

Истории поисковых систем

Rambler

Английское слово «rambler» имеет множество значений. Самим работникам компании больше по душе перевод «бродяга», под которым подразумевается бродяга по Интернету.

Разрабатывать поисковый механизм начала в 1991 году группа единомышленников из подмосковного научного города Пущино. Через пять лет, в 1996 году, программист Дмитрий Крюков создал первую уникальную российскую поисковую программу, которую сразу и запустили эксплуатацию. Первая в России поисковая система с самого своего начала расположилась по адресу rambler.ru. Постепенно небольшая группа единомышленников выросла в крупный интернет-холдинг. В феврале 1997 года заработала рейтинговая система Rambler's Top100 (top100.rambler.ru). Спустя 3 года, 7 марта 2000 года, был зафиксирован миллиардный посетитель страниц, зарегистрированных в рейтинге. То есть все ресурсы, которые стоят в рейтинге Rambler (и более к Rambler никакого отношения не имеют), за три года получили миллиард посетителей.

» Через двоеточие после ключевого слова или выражения можно указать число, которое будет влиять на вес этого слова или выражения. Также можно использовать оператор «<-» для задания уточняющего слова или выражения — это увеличит релевантность документов, содержащих уточняющее выражение.

Кто ищет лучше?

У каждого человека есть свое мнение о том, как должен выглядеть естественный запрос к поисковой системе. Поэтому оценить, насколько результат поиска соответствует запросу пользователя, довольно сложно. Google и Yandex обладают самыми большими базами по русскому Интернету. Но Rambler, так как это первая поисковая машина, начавшая индексировать российский Интернет, лучше ведет поиск по старым документам, которые в силу каких-либо причин не стали популярны. Кроме того, ресурсы с установленным счетчиком Rambler Top 100 (а это одни из самых популярных рейтингов) имеют на Rambler больший вес и индексируются еще чаще.

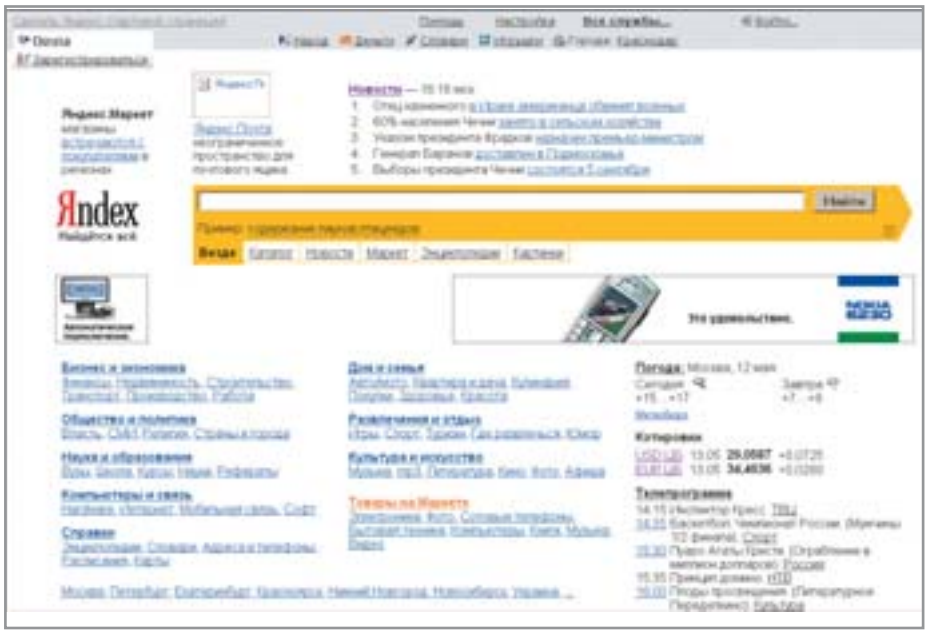
Особенность Google состоит в том, что благодаря применяемой там системе присвоения веса PageRank хорошо ищутся авторитетные сайты. В этом отношении Google был первым, но сейчас подобные ссылочные алгоритмы используют почти все поисковики.

Yandex отличается своим развитым языком запросов (которым пользуются менее 1% пользователей) и большими познаниями в морфологии русского языка, но разработчики системы всегда видели своей задачей обеспечение точности поиска при так называемом естественно-языковом запросе, то есть когда неподготовленный человек просто пришел и просто спросил.

Надо заметить, бытуют мнения, что Yandex наиболее релевантен, а через Google лучше искать файлы. Но это каждый сможет опробовать лишь на личном опыте и собственных ощущениях.

Дополнительные возможности

Кроме главной своей функции — полнотекстового поиска по документам Интернета — поисковые системы часто предоставляют ряд дополнительных услуг. Например, у всех трех рассматриваемых поисковых систем есть возможность поиска в каталоге.



▲ Yandex помимо поиска предлагает и другие сервисы

Для поиска графических изображений на Yandex отведен отдельный раздел. Обычно изображение находится в каком-то документе и связано с некоторым текстом. По этому тексту и можно попытаться его найти. Тут можно использовать текст подписи к картинке (параметр «alt» тега <img>, задающий поясняющую надпись) или же текст ссылки на нее. Также информацию об изображении можно почерпнуть из текста, который расположен в документе рядом с картинкой, и из названия графического файла. При этом ключевые слова подвергаются и транслитерации, и переводу на английский язык. Таким образом, если вы ищете изображение по ключевому вы- »

Истории поисковых систем

Google

Название компании происходит от английского слова «googol», обозначающего число — единицу и сто нулей. Сергей Брин и Лари Пейдж встретились случайно в Стэнфорде, где и завязалось их знакомство. В январе 1996 года они начали работу над поисковым механизмом BackRub. Первую половину 1998 года молодые люди занимались совершенствованием своей системы, установив оборудование прямо в спальне Лари. 7 сентября 1998 новоиспеченная Google, Inc. уже въехала в своей первый настоящий офис. Поисковая система обслуживала 10 000 запросов в день. И, все еще находясь в стадии разработки, попала в Top 100 веб-сайтов журнала PC Magazine. В 1999 году основатели Google дважды сменили офис, значительно расширили штат, получили множество наград, а поисковая машина обрабатывала уже 500 000 запросов ежедневно.

В 2000 году Google стал крупнейшей мировой поисковой системой. Последующие годы поисковая машина совершенствовалась, добавлялись новые службы и возможности. Сегодня его индекс содержит сведения о более чем 4 млрд различных URL, а сама система обрабатывает 200 миллионов запросов ежедневно. Россия является редчайшим исключением из правил — страной, где позиция Google не первая, и даже не вторая. По статистике Rax.ru и Spylog, через Yandex на сайты попадают около половины всех ищущих, через Rambler — около четверти, а через Google — около 15%. При этом заметно, что пользователи Google, попадающие на русские сайты, как правило, находятся не в России — доказательством является относительный рост трафикогенерации Google в отечественные праздники, не совпадающие с мировыми.

▲ http://images.google.com: поиск всех изображений по слову Chip

» ражению, к примеру, «розовый слон», то найдутся в том числе и файлы, содержащие в своем названии сочетания «slon», «elephant», «pink» и т. д.

На Rambler есть специальная форма для поиска файлов. Файлы можно искать любые или определенного типа: картинки, аудио, видео. В отличие от Yandex поиск происходит только по именам файлов или каталогов, без анализа каких-либо элементов, связанных с файлом. Имя файла можно задавать точным значением или используя шаблоны (символы «*» и «?») и регулярные выражения (более сложные формы шаблонов). Есть возможность задать каталоги, которые следует исключить из поиска или же, наоборот, искать только в них. Эти же ограничения можно наложить и на доменные зоны, в которых должен располагаться сервер с нужным файлом.

Заглавная страница каждой поисковой машины — это не просто форма для ввода запроса, но еще и внушительный портал. На сайтах Rambler и Yandex можно найти ссылки на популярные ресурсы, программу телепередач, прогноз погоды, гороскоп, курсы валют, последние новости, почтовый сервис, онлайн-словари, энциклопедии и множество других разделов. Но в тот момент, когда вам ничего этого не нужно и вы хотите воспользоваться именно поиском, к вашим услугам облегченные варианты страниц. У Rambler — http://r0.ru, у Yandex — http://ya.ru.

■ ■ ■ **Дмитрий Солошенко**

Малозаметные возможности

Секреты Google

Помимо поиска Google предлагает доступ к целому набору других возможностей. Для того чтобы осветить их все, не хватит и целой статьи, но некоторые, безусловно, стоит упомянуть.

Поиск синонимов

Если поставить перед искомым словом оператор «~», будут найдены документы, содержащие не только само слово, но и его синонимы. Словарь синонимов представлен только на английском языке. Кроме того, поисковая машина понимает числовые диапазоны — через знак «..» можно задать нижнюю и верхнюю границу некоторого числового значения, которое должно присутствовать в документе.

Панель инструментов

Google Toolbar — надстройка для браузера Internet Explorer версии 5.0 и выше, которая позволяет вести поиск независимо от того, какой сайт открыт у вас в окошке браузера. Кроме того, Google Toolbar блокирует всплывающие окна (только в Internet Explorer версии 5.5 и выше), помогает заполнять одним нажатием мыши формы, состоящие из нескольких полей, а также подсвечивает на странице искомое слово. Надстройка не работает с такими браузерами на основе Internet Explorer, как MyIE2. http://toolbar.google.com

Онлайн-перевод

Если нужный вам текст оказался на французском, немецком, итальянском, испанском или португальском языке, Google с помощью собственного механизма переведет содержимое страницы на английский язык (см. ссылку «Translate this page» справа от найденной страницы).

Хотя перевода страниц на русский язык у разработчиков нет даже в проекте, онлайн-перевод на английский может здорово помочь, если нужный вам текст оказался на немецком языке, а по-немецки вы знаете только «хенде хох» и «гитлер капут».

Калькулятор

Сюрприз! Кто бы мог подумать, что с помощью поисковой системы можно искать не только слова и целые фразы на страницах, но и результаты математических вычислений! Работает это все так же, как и обычный поиск. Попробуйте ввести в поисковую строку что-нибудь типа «15+78*4,5», нажмите «Найти» и посмотрите, что у вас получится. Подробное описание синтаксиса калькулятора (например, как вам взять натуральный логарифм числа пять) — на http://google.ru/help/calculator.html.

Кроме того, Google может работать и как переводчик единиц измерения: введите в поисковую строку «1 mile in kilometers» и узнайте, сколько километров в одной миле.

Мини-Google

Вы считаете, что целых 14 кбайт текста и графики — слишком много для заглавной страницы такого поисковика, как Google? Сделайте себе на жестком диске или собственном сайте заглавную страницу для Google размером в… 202 байта!

```
<html><meta http-equiv="content-type"
content="text/html; charset=UTF-
8"><body><form
action=http://google.ru/search
method=get name=f><input type=hidden
value="UTF-8"><Input
name=q></form></body></html>
```