Автор: Administrator
16.01.2010 19:22 -

Поисковые системы Навык, просто жизненно необходимый для любого обитателя Сети, однако, как это ни парадоксально, более половины посетителей грамотно искать информацию в Интернете вообще не умеют! Обучиться этому нужно еще до того, как Вы начнете создавать свой первый серьезный интернет-продукт. Кто владеет информацией, то владеет миром! А тот, кто умеет в совершенстве пользоваться поисковыми машинами, всегда будет владеть оперативной и актуальной информацией! За редкими (очень редкими!!!) исключениями сегодня в Интернете можно найти практически ВСЁ!! Многое можно найти в Сети за пять минут, даже не обладая изощренной фантазией в составлении поисковых запросов.

Для начала – несколько слов о сути работы поисковых систем и общие, так сказать, принципы. В отличие от каталогов (специальных списков сайтов, разбитых по категориям и снабженных кратким описанием), практически все основные поисковые системы работают по принципу индексации информации, содержащейся на тех или иных интернет-страницах.
Что это значит? Это значит, что если каталогизацию производят живые люди (увы, ограниченные в своих возможностях, а потому объем ссылок в каталогах составляет ничтожно малый процент от общего объема сайтов Сети), индексацию в поисковиках производит поисковый робот. Поисковый робот без устали, двадцать четыре часа в сутки бороздит Сеть в поисках появления новых ссылок на документы и обновления информации о ссылках старых, уже проиндексированных ранее.
Поисковая машина, это огромный программно-аппаратный комплекс, в котором различными этапами обработки индексируемой информации занимаются различные службы. Одни поисковые сервера заняты скачиванием интернет-страниц, другие эти страницы индексируют, третьи группируют индексы в единую базу, осуществляют нормализацию (приведение слов к единой форме)…
При подаче пользователем поискового запроса системе, из ее базы выбираются проиндексированные документы, содержащие слова, которые были введены пользователем в строке запроса. Далее эти документы ранжируются по определенному, довольно сложному алгоритму, чтобы первыми в выданном по запросу пользователя списке, оказались ссылки на те странички, которые содержат наиболее точный ответ на запрос пользователя. Это называется релевантностью.
Если говорить просто, то релевантность, - это соотношение между желаемой и действительно получаемой информацией. Это то, насколько реально полученный документ соответствует тому, что следовало бы получить из поисковой системы.
Несмотря на то, что все поисковые системы построены на общих принципах (чем чаще искомые слова встречаются в документе, тем выше его вес, как правило), алгоритмы у них, все же, разные.
Каждая поисковая машина использует свой собственный алгоритм вычисления релевантности, не похожий на алгоритмы других поисковиков (например, для

большинства поисковых систем высокорелевантным текстом считается тот, где вхождение запроса в текст равно приблизительно 4-7%. Если больше, то система может принять текст за поисковый спам и наложить на страницу понижающий фильтр или вообще убрать ее из результатов выдачи по искомому запросу).
Так же, многие поисковики учитывают взаимное расположение слов в документе – если в найденном тексте слова расположены в том же порядке, что и в поисковом запросе, документ будет проранжирован выше. Может учитываться расстояние между словами – если искомые слова содержатся в одном предложении, документ будет иметь больший вес, чем, если бы, искомые слова содержались в пределах абзаца или даже страницы. Еще вес искомого документа может увеличиваться поисковой машиной, если на данный документ имеется большее количество ссылок с других сайтов, чем на аналогичный документ, но с меньшим количеством ссылок. Значимость могут добавить ссылки с наиболее весомых страниц (PageRank).
* * *
Для точного поиска Вам потребуется знание синтаксиса языка запросов. Это специальные символы, которые пишутся в поисковой строке вместе с ключевыми словами и уточняют критерии Вашего поиска. Синтаксис языка запросов в разных поисковых системах может отличаться (обычно в справочных данных на поисковом сервере приводится подробная информация о синтаксисе запросов данной конкретной системы), однако основные поисковики, такие как Yandex, Google и Rambler, имеют некоторое сходство в использовании ряда специальных символов.
Поисковая фраза, “заключенная в кавычки”, будет найдена в точном соответствии поисковому запросу. То есть, слова в документе будут находиться в той же форме и расположены точно в таком же порядке, что и в закавыченной фразе поискового запроса.
Символ «+» (плюс) перед словом поискового запроса задает параметр, согласно которому данное слово ОБЯЗАТЕЛЬНО должно присутствовать в искомом документе.
Символ «-» (минус) или «~» (тильда) перед словом поискового запроса задает противоположный параметр, согласно которому данное слово НЕ ДОЛЖНО присутствовать в найденных документах. В пределах предложения – «~» или в пределах всего документа – «~~» (В Rambler’е вместо «-» используется восклицательный знак «!»…)
В Google «~» (тильда) обозначает поиск синонимов. То есть, если в Google перед искомым словом поставить тильду «~», будут найдены документы, содержащие не только само слово, но и его синонимы. К сожалению, словарь синонимов представлен только на английском языке.
Так же очень широко в поисковых запросах используется логическая связка «или». В поисковых машинах Yandex и Rambler она имеет вид «|», а в Google вид «OR».
Несомненное достоинство Yandex и Rambler заключается также в том, что в этих поисковиках можно строить сложные поисковые запросы с использованием скобок и оператора логического сложения «&» (в Yandex оператор «&» означает, что искомые слова должны находиться в одном предложении, в Rambler – что они присутствуют в одном документе. Для того, чтобы и Yandex искал по всему документу, используйте «&&»).
Так, если Вам нужно найти картинку доллара или евро, Ваш поисковый запрос может выглядеть следующим образом: (фото | изображение | картинка | рисунок) & (доллар |

USD | евро | EUR) По такому поисковому запросу вам будут выданы ссылки на изображения евро и доллара, а если Вы перейдете на вкладку «картинки» поисковика, то и сами эскизы изображений искомых картинок.
Весьма немаловажно, что Yandex чувствителен к регистру букв. Если в поисковом запросе присутствует слово, написанное с заглавной буквы, то Yandex выдаст Вам документы, в котором искомые слова написаны именно с заглавной буквы (если это слово не первое в предложении). Если же в поисковом запросе слово написано строчными буквами, Yandex выдаст документы, где данное слово встречается как написанное со строчной буквы, так и с прописной.
В Yandex, независимо от того, в какой форме вы употребили слово в запросе, поиск учитывает все его формы. Например, если задан запрос «идти», то в результате поиска будут найдены ссылки на документы, содержащие слова «идти», «идет», «шел», «шла» и т.д. На запрос «окно» будет выдана информация, содержащая и слово «окон», а на запрос «отзывали» – документы, содержащие так же и слово «отозвали»…
* * *
Поисковых машин в Сети существует достаточно много. Приведем лишь десять самых известных и наиболее распространенных из них:

Yandex - http://www.yandex.ru ,
Google - http://www.google.com ,
Rambler - http://www.rambler.ru ,
Aport - http://aport.ru ,
Yahoo - http://www.yahoo.com ,
Mail.ru - http://mail.ru ,
AltaVista - http://www.altavista.com ,
Webalta - http://www.webalta.ru ,
MSN - http://www.msn.com ,
All The Web - http://www.alltheweb.com .

Какому же поисковику отдать предпочтение? Какой по праву может считаться лучшим? Боюсь, что однозначного ответа на эти вопросы просто не существует. Сколько людей, столько и мнений, столько предпочтений и пристрастий. Попробуйте поработать с разными поисковыми машинами и выберите для себя ту, что Вам лично понравится больше остальных.

Nota Bene: Как выбрать поисковую машину

1. Охват и глубина

Под охватом имеется в виду объем базы поисковой машины: который измеряется тремя показателями - общим объемом проиндексированной информации, количеством уникальных серверов и количеством уникальных документов. Под глубиной понимается – существует ли ограничение на количество страниц или на глубину вложенности директорий на одном сервере.
Некоторые машины пишут на своем сайте статистику робота. Но можно проверить и самому – надо задать несколько поисковых запросов, состоящих из одного слова (чтобы исключить влияние языка запросов, в том числе – различного трактования пробела), и при этом смотреть на статистику результатов, выдаваемую машиной – обычно в начале списка указано, сколько всего было найдено документов. Помимо того, что слова должны быть из разных областей, хорошо еще взять слова разных "весов" – редкие, "средние" и "тяжелые" (частотные), и сравнить количество найденного. Тяжелые слова, в частности, тестируют полнотекстовость (индексацию всех слов документа) поисковой машины.
Глубину хождения робота проверить сложнее – для этого надо взять какие-то сайты, например, с разветвленной структурой архивов, и проверить, проиндексированы ли документы, на которые можно попасть только, например, за 6 переходов по ссылкам.

2. Скорость обхода и актуальность ссылок

Скорость обхода Сети показывает, насколько быстро происходит индексация вновь добавленного ресурса и насколько быстро обновляется информация в базе. Важным показателем качества поисковой машины (ее робота) является не только "захват" новых территорий: но и отслеживание состояния уже охваченных. Сервера исчезают и появляются, страницы на них обновляются. Ссылки, которые выдает поисковая машина в списке найденного, должны, во-первых, существовать, и, во-вторых, их содержание должно соответствовать запросу.
Объективную информацию можно получить, проанализировав логи серверов – робот поисковой машины представляется обычно именем своей машины (или похожим образом), так что можно увидеть, как часто он бывает на сервере, сколько страниц просматривает и т.д. К сожалению, обычно для изучения бывает доступен лог только своего сайта, поэтому остается экспериментальный способ.
Для определения скорости обхода надо создать где-нибудь страничку текста, добавить ее в поисковики и посмотреть, как быстро она начнет находиться. Или изменить уже имеющуюся страничку. Для определения актуальности ссылок – проверить документы

хотя бы на первой странице списка найденного по нескольким запросам. Сообщение "Not Found" свидетельствует о том, что документ более не существует.

3. Качество поиска (субъективный показатель)

Каждая поисковая машина имеет свои алгоритм сортировки результатов поиска. Чем ближе к началу списка оказывается нужный вам документ, тем лучше работает релевантность.
Только путем эксперимента. Рекомендуется для сравнения делать запросы разной длины. Можно также использовать язык запросов, при этом те, кому неохота читать описание, могут воспользоваться развернутой страницей запроса ("расширенный поиск" в Апорте и Яндексе, "детальный запрос" в Рамблере – варианты перевода на русский язык "advanced search").

4. Скорость поиска

Если поисковая машина отвечает медленно, работать с ней неэффективно. Стоит добавить, что видимая пользователю скорость зависит не только от самой поисковой машины, но и от Интернет-каналов.
Путем эксперимента – надо поискать запросы разной длины, разной <тяжести> слов и в разное время суток (загрузка серверов существенно неравномерна по суткам, пик – около трех-четырех часов дня).

5. Поисковые возможности (работа с языком документа, язык запросов)

Еще один пункт сравнения – что именно и как поисковая машина вносит в индекс. Полнотекстовая поисковая машина индексирует все слова видимого пользователю текста. Наличие морфологии дает возможность находить искомые слова во всех склонениях или спряжениях. Кроме этого, в языке HTML существуют тэги, которые также могут обрабатываться поисковой машиной (заголовки, ссылки, подписи к картинкам и т.д.).
Язык запросов в виде стандартных логических операторов (И, ИЛИ, НЕ) есть практически у всех машин. Некоторые умеют искать словосочетания или слова на заданном расстоянии – это часто важно для получения разумного результата.

Дополнительной возможностью является поиск в зонах документа – заголовках, ссылках, ключевых словах (META KEYWORDS) и т.д. Дополнительная возможность языка запросов – естественно-языковый запрос, который не требует знания операторов.

Обычно эта информация публикуется на сервере поисковой машины (в Help'e). Тем не менее, рекомендуется проверить на реальных запросах, поскольку иногда желаемое выдается за действительное.

6. Дополнительные удобства

Это дополнительные возможности, которые предоставляет пользователям поисковая машина. Сюда входит всевозможные варианты поиска (специализированные страницы, поиск похожих документов, ограничение области поиска), и список найденных серверов, и поиск по датам и серверам, и удобный интерфейс поисковой машины, и возможность его персонализации.
Информация может быть частично опубликована на сервере поисковой машины, но лучше всего попробовать самому поработать с этими возможностями.

На мой взгляд, новичку удобнее всего начинать освоение искусства сетевого поиска с Yandex, так как эта система не только обладает одной из самых больших баз в русскоязычном Интернете, но и выгодно отличается от ближайших конкурентов своим развитым языком запросов и широким диапазоном русскоязычной морфологии.
Так же, Yandex обеспечивает высокую точность поиска при естественно-языковом запросе, когда Вы просто вводите в поисковую строку вопрос, ответ на который хотите получить.
Хотя, конечно же, разумнее всего использовать сразу несколько поисковиков одновременно. Это существенно расширит Ваши возможности. Не нашли то, что искали, при помощи одной машины, переходите к следующей.
Осуществляя поиск, избегайте общих слов. Чем уникальнее ключевое слово, по которому осуществляется поиск, тем больше шансов найти именно то, что Вам нужно. Ищите больше, чем по одному слову. Сократить объем ссылок можно, определив несколько ключевых слов. Используйте синонимы.
Используйте различные инструменты для поиска информации разного профиля.
Пользуйтесь расширенным запросом. Если один из найденных документов ближе к искомой теме, чем остальные, нажмите на ссылку «Найти похожие документы».
Пользуйтесь языком запросов. С помощью языка запросов можно сделать запрос более точным. Во многих поисковых системах есть форма расширенного запроса, в которой можно использовать основные механизмы сужения поиска.
Вывод: Использование поисковых систем является одним из самых важных и ценных интернет-навыков, которые Вы когда-либо приобретали и приобретете в будущем, ибо

умение отыскивать в Сети нужную информацию – основа основ, как электронной коммерции, так и любой другой интернет-деятельности вообще!