Мир ПК 2009 № 05

Торные дороги WEB

08.04.2009 Автор: Валерий Коржов

Тем не менее специалисты, занимающиеся мониторингом содержимого Интернета или контролем за утечками конфиденциальных сведений, выработали определенные приемы эффективного поиска информации, позволяющие таким образом составить запрос к поисковой системе, чтобы нужная ссылка оказалась на первой странице выдачи поисковика.

В частности, приемам подобного поиска обучает на своих семинарах Андрей **Масалович**, руководитель направления конкурентной разведки в компании "ДиалогНаука". Суть их заключается в том, чтобы правильно подобрать поисковый запрос. Если не удалось обнаружить требующейся информации на первой странице выдачи, то нужно не искать ее на других страницах, а модифицировать запрос до получения ожидаемого результата. Чтобы правильно выбрать слова запроса, рекомендуется обратиться к расширенному поиску, механизмы которого включены практически во все известные поисковые машины.

Свои знают больше

Целесообразно обращаться с запросами к поисковику, имеющему более полную базу индекса. О русскоязычном Web больше всего известно "Яндексу", который обнаружил уже более 3 млрд. страниц на русском языке и обрабатывает половину русскоязычных поисковых запросов. К сожалению, поисковый робот "Яндекса" довольно медлителен -- он обходит известные ему адреса примерно 2 раза в месяц. В то же время роботы Google (которые, кстати, на 11 лет моложе «Яндекса») за день анализируют до 8 млрд. страниц, правда, во всем Интернете и на всех языках. Поисковая машина «Рамблер» сейчас также достаточно компетентна в русскоязычном Интернете: у нее хороший стартовый каталог и быстрые роботы, обходящие известные ей сайты дважды в день. В общем, трех этих поисковиков достаточно, чтобы найти необходимые данные. Однако каждая поисковая система отличается своими особенностями при обработке запроса, и их следует учитывать при поиске.

Спроси у «Яндекса»

Например, "Яндекс" уделяет повышенное внимание первым двум словам запроса, а остальные слова, начиная с третьего, могут вообще не участвовать в поиске. Кроме того, в данной системе предусмотрен механизм учета морфологии русского языка с порождением гипотез для неизвестных слов. В языке запросов "Яндекса" для отключения морфологии используется символ "!". Чтобы сделать все слова

запроса одинаково значимыми, их нужно разделить знаком "неранжирующее И" -- ">>"; чтобы слово обязательно присутствовало в найденном документе, следует поставить перед ним знак "+"; а чтобы исключить документы, содержащие определенное слово, необходимо пометить его символом "-". Причем в запросе не рекомендуется писать слова со знаком дефиса, поскольку "Яндекс" может воспринять его как минус, что порой приводит к самым неожиданным результатам. Если известна точная цитата из искомого документа, то ее требуется выделять кавычками. Нужно отметить такую полезную возможность "Яндекса", как поиск по типу документов: PDF (Adobe Acrobat Reader), RTF, DOC (Microsoft Word), PPT (PowerPoint) и SWF (Macromedia Flash). Важно, что эта система публикует статистику запросов, помогающую оценить, какие слова чаще всего интересуют тех, кто ищет информацию в Интернете.

Google it!

У российских пользователей вторым по популярности поисковиком является Google, несколько лет назад открывший свое представительство в России. Он активно наращивает не только поисковые возможности, но и различные дополнительные веб-сервисы. Так, система показывает распространенность используемых в запросе слов сразу же после их набора в поисковой строке. При работе с Google целесообразно приводить точные цитаты в кавычках и указывать ограничения по типам документов, выполняющиеся командой "filetype:". Здесь, в отличие от "Яндекса", допустимы любые расширения. Кроме того, Google предоставляет возможность вести поиск по определенным HTML-тегам. Так, команда "allinurl:" позволяет искать слова лишь в ссылках на документы, "allintitle:" -- в их заголовках. Кроме того, можно ограничить зону поиска, указав конкретный сайт с помощью директивы "site:". Впрочем, и у "Яндекса" есть аналогичная директива -- "host=". Именно ограничения по сайту и типам документов и употребляются для поиска конфиденциальной информации, зачастую содержащейся в виде таблиц Excel, презентаций и PDF-документов.

Старик «Рамблер»

Третий по распространенности поисковый механизм российского Интернета -- «Рамблер». Это старейший поисковый сервис. Поисковый индекс «Рамблера» не очень велик, но, как уже было отмечено, роботы у него достаточно быстрые, и потому индекс быстро наполняется. К тому же в качестве стартового каталога служит Rambler Top100, т. е. список наиболее популярных сайтов Рунета, что позволяет поисковику оперативно находить наиболее востребованные ресурсы.

В этом поисковике принят более простой и интуитивно понятный язык запросов. Так, чтобы увеличить важность слова в результатах поиска, можно перед ним поставить знак "+" и даже повторить его несколько раз. А вот знак "-" снижает «вес» слова при подсчете релевантности документа. В «Рамблере» также допустимо проводить поиск по частям документа посредством директив $All, $URL, $Title, $Header, $Essence, $Address. Группировать слова и директивы помогают логические операторы ИЛИ (символ "|"), И (символ "&") или NOT (символ "!"), принятые в языках программирования. Благодаря всему этому удается составлять более точные поисковые конструкции.

Поисковый консилиум

Объединить возможности поисковиков помогают метапоисковые системы, позволяющие обратиться к ним с одним запросом одновременно. Полученные из разных источников результаты представляются в наиболее удобном для изучения виде. К примеру, такая система, как Nigma, обрабатывает результаты и категоризует их, а Quintura позволяет пользователю управлять смыслом и контекстом запроса. Кроме того, метапоисковые системы помогают уточнять и наращивать запросы. Поэтому, прежде чем заниматься поиском информации по не очень известной теме, стоит изучить предметную область с помощью метапоисковой машины, чтобы быстрее найти дорогу к цели. Однако в таких машинах невозможно применять особенности языков поисковых систем.

Иногда, чтобы улучшить результаты поиска, рекомендуется включать в запрос специальные "магические" слова для фильтрации ссылок определенного вида. Например, когда требуется найти контактную информацию о какой-нибудь фирме, то стоит ввести слово "судебный" -- и оно выделит в результатах документы судебных разбирательств, где обычно указаны юридические адреса фирм. Аналогично, чтобы получить полные тексты книг, стоит употребить слово "загрузить" (или download). Такие же слова помогут подбирать и метапоисковые системы. В общем, для эффективного поиска достаточно запроса максимум из четырех слов, причем из них одно-два зададут тему, максимально расширяя количество нужных документов, и еще одно-два "магических" выделят из большой выборки те документы, что содержат наиболее ценную информацию.

Основные приемы поиска информации в глобальной сети