Поиск информации в сети

Проблема поиска и средства его организации

Гигантские и непрерывно увеличивающиеся объемы доступной в Интернет информации, в том числе оперативной, делает проблему поиска необходимых сведений весьма актуальной и сложной. Скорость поиска нужной информации определяет в значительной степени профессионализм пользователя Интернет. Для автоматизации этой задачи разработаны различные, как зарубежные, так и отечественные системы поиска, представляющие собой Web-страницы специального вида.

Однако, несмотря на наличие многочисленных средств автоматизации поиска, эта задача остается достаточно трудоемкой, требующей от пользователя определенного опыта, интуиции, знания терминологии, используемой в его предметной области.

По оценке, опубликованной в журнале Nature от 8 июля 2009 г., число публично индексируемых Web-страниц составляло 800 млн. Спустя год автор исследования (Стив Лоуренс из института NEC Research Institute) полагал, что их число увеличилось почти вдвое – до 1,5 млрд. Даже лучшие поисковые механизмы индексируют не более чем одну страницу из шести. Для того чтобы извлечь полезную информацию из сети Интернет, нужно знать, где и как вести поиск.

Имеющийся в Internet Explorer инструмент Поиск упрощает обращение к средствам поиска, избавляя от знания адресов поисковых машин. Однако лучше непосредственно обращаться к поисковым системам, загружая соответствующую страницу.

Виды средств поиска

- каталоги и специализированные базы данных;
- поисковые системы;
- метапоисковые системы.

Каталоги и базы данных

Каталоги в WWW аналогичны систематическим библиотечным каталогам. Поиск по каталогам состоит в последовательном движении по иерархическому списку ссылок, называемых рубриками или категориями. На первой странице каталога содержится ссылки на крупные темы, например, Культура и искусство; Медицина и здоровье; Общество и политика; Бизнес и экономика; Развлечения и др. Щелчок мыши на соответствующей ссылке (категории) открывает страницу, содержащую ссылки, детализирующие выбранную тему (рубрику). Двигаясь вниз по детализирующим категориям, можно найти страницу с нужной информацией. На каждой странице, открываемой при движении по каталогу тем или иным способом, указывается последовательность просмотренных вложенных рубрик.

Поисковые системы

Существуют десятки крупных и тысячи малых и специализированных Web-узлов, предназначенных для поиска в Интернете. Средства поиска этой группы позволят пользователю по определенным правилам сформулировать требования к необходимой ему информации (с помощью языка запросов создать запрос). После этого машина поиска автоматически просматривает документы на контролируемых (индексируемых) ею сайтах и отбирает те из них, которые, «по мнению» поискового сервера, соответствуют сформулированным пользователем требованиям (релевантны запросу). В поисковых узлах используются собственные индексы Интернета, постоянно обновляемые особыми программами, называемыми пауками (spiders). Программа-паук обследует Web, проверяя каждую ссылку на данной странице, затем на страницах, адресуемых ссылками, и т. д., и сообщает своему владельцу сведения обо всех страницах для последующей индексации.

Достоинство автоматизированного поиска состоит в том, что он обеспечивает просмотр очень больших объемов информации, имеющейся в Интернет в данный момент. Однако сложность точного описания запроса, адекватно отражающего ваши информационные потребности, а также еще большая сложность задачи автоматического определения степени соответствия вашему запросу просматриваемых страниц, приводит к тому, что количество страниц, отобранных «с первого захода», как правило, или очень мало, или чрезмерно велико. В целом поиск с использованием поисковой машины представляет собой итерационный (многоходовой) процесс, в результате которого постепенно уточняется форма запроса.

Метапоисковые системы

Поисковая система просматривает определенный набор серверов и отбирает документы в соответствии с присущими ей критериями. В итоге поиск разными системами по одним и тем же ключевым словам дает различные результаты. Это привело к идее создания так называемых метапоисковых (или мультипоисковых) систем, которые сами ничего не ищут, но обращаются за помощью сразу к нескольким поисковым системам. Каждая из метапоисковых систем имеет свой язык запросов. Система переводит сформулированный на ее языке запрос на языки запросов, используемые каждой машиной поиска. Далее, результаты поиска всеми системами объединяются и представляются в соответствующей форме. Естественно, что поиск с помощью метапоисковых систем занимает большее время по сравнению с обычными системами поиска.

Наиболее популярные поисковые системы

- 1. Google (www.google.com) Самая быстрая и самая большая поисковая система.
- 2. Яndex (www.yandex.ru) Лучшая из поисковых систем отечественного производства.
- 3. AltaVista (www.altavista.com) Предоставляет большое расширение критериев поиска.
- 4. Yahoo! (www.yahoo.com) Один из первых поисковых серверов в Интернет.
- 6. Рамблер (www.rambler.ru) До недавнего времени самая известная русская поисковая система.

Характеристики поиска

При поиске в Интернет важны две составляющие - полнота (ничего не потеряно) и точность (не найдено ничего лишнего). Обычно это все называют одним словом - релевантность, то есть соответствие ответа вопросу.

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.

Скорость обхода и актуальность ссылок

Скорость обхода Сети показывает, насколько быстро происходит индексация свежедобавленного 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.

Скорость и качество поиска

- Каждая поисковая машина имеет свои алгоритм сортировки результатов поиска. Чем ближе к началу списка оказывается нужный вам документ, тем лучше работает релевантность.
- 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.

Поисковые возможности

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.).

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.

Спасибо за внимание!