УДК 681.3.06

А.А. Шайдуров, П.М. Зацепин

Информационно-поисковые системы для intranet-сетей

В последнее время помимо глобальной сети Internet широко развиваются intranet-сети. Все увеличивающаяся популярность intranet-сетей приводит к их стремительному росту [1]. Неизбежно встает проблема быстрого поиска необходимой информации. И тем логичнее вырисовывается необходимость создания информационно-поисковых систем (ИПС) для intranet-сетей. Несмотря на то, что существует множество различных ИПС, работающих в Internet, в большинстве случаев применение их в intranet-сети весьма затруднительно.

Глобальная сеть Internet и локальная сеть intranet очень сильно отличаются друг от друга. Например, форматами данных. В Internet распространены такие форматы, как HTML-документы, FTP-архивы, электронная почта и т.д. В intranet наряду с HTML-документами присутствуют документы, с которыми мы работаем в повседневной жизни. Это различные текстовые документы, базы данных, электронные таблицы и т.д. Чаще всего ИПС Internet-типа не могут полноценно производить поиск в intranet-сети. Каждая поисковая система обладает рядом особенностей. Проанализируем некоторые из них [4].

1. Тип поисковой машины

«Поисковые машины индексируют каждое слово в документе, исключая лишь некоторые стоп-слова. Любое слово, встречающееся на Web-странице, подвергается анализу при определении его релевантности к запросам пользователей. Это может исходить от алгоритма экстрагирования, например, по частоте употребления на странице одних и тех же слов.

При поиске в intranet-сети необходимо индексировать слова, встречающиеся не только в Web-документе, но и в других форматах. Следовательно, если идет поиск в текстовом документе, то принцип определения релевантности слова остается прежним. Однако если индексируются электронные таблицы или базы данных, то релевантность слова будет определяться иными методами.

2. Размер

Размер поисковой машины определяется количеством проиндексированных страниц. Например, в поисковой машине с большим размером могут быть проиндексированы почти все страницы. При среднем объеме сервер может быть частично проиндексирован, а при малом объеме некоторые страницы могут вообще не попасть в каталоги поисковой машины. *В свою очередь ИПС intarnet-типа должна индексировать все документы, находящиеся в сети.*

Имеются и другие отличия, которые естественным образом вытекают из описанных выше.

Существует еще один момент, о котором стоит упомянуть: ИПС должна выдавать полученный результат в Web-формате, следовательно, необходим конвертер, который бы преобразовывал данные из разных форматов в формат HTML. В этой ситуации выбирается несколько направлений, которые должен выбрать программист при написании своей ИПС :

1. Исходный документ конвертируется в HTML вид, а затем в нем производится поиск;

2. Вначале производится поиск в документе в соответствии с его форматом, а затем найденная информация выдается в виде HTML файла.

Каждый из этих путей имеет свои достоинства и недостатки. Если исходный документ имеет небольшие размеры, то вначале его надо конвертировать в HTML-формат, а затем производить поиск по разработанным алгоритмам. Однако если размер файла достаточно большой (например, электронная таблица в несколько тысяч записей и ссылками на другие таблицы), то тогда предпочтительнее будет второй вариант. Сначала ИПС находит в этом файле запрошенную информацию, а затем найденные строки конвертирует в HTML-формат и выводит полученный результат. Исходя из этих соображений можно предложить также третий вариант поиска, при котором ИПС сама будет «решать», что для нее лучше – конвертировать и найти или найти и конвертировать. Исходя из вышесказанного рассмотрим в общем случае схему ИПС. В различных публикациях, посвященных конкретным системам [1, 2], приводятся схемы, которые отличаются друг от друга только способом применения конкретных программных решений, а не принципом организации различных компонентов системы [3]. Поэтому рассмотрим ИПС на примере, взятом из работы [2] (рис. 1).

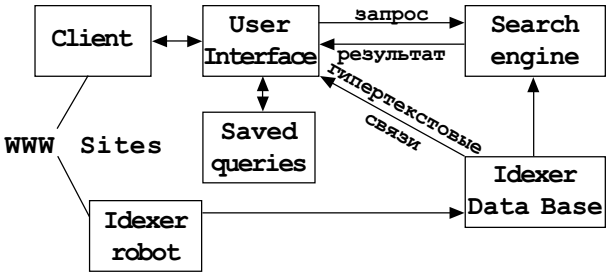

Рис. 1. Типовая схема информационно-поисковой системы

Client (клиент) на этой схеме – это программа просмотра конкретного информационного ресурса. В настоящее время наиболее популярны мультипротокольные программы типа Netscape Navigator. Такая программа обеспечивает просмотр документов WWW, Gopher, Wais, FTP-архивов, почтовых списков рассылки и групп новостей Usenet. В свою очередь, все эти информационные ресурсы являются объектом поиска информационно-поисковой системы. *Для ИПС intranet-типа необходим уже другой* ***Client. Client*** *для нашей ИПС должен уметь просматривать более широкий спектр документов, в том числе базы данных, электронные таблицы, различные текстовые форматы и т.д. В свою очередь в локальной сети, как правило, не присутствуют FTP-архивы, почтовые ссылки и т.д.*

User interface (пользовательский интерфейс) – это не просто программа просмотра. В случае информационно-поисковой системы под этим словосочетанием понимают также способ общения пользователя с поисковым аппаратом: системой формирования запросов и просмотров результатов поиска.

Search engine (поисковая машина) служит для трансляции запроса на информационно-поисковом языке, в формальный запрос системы – поиска ссылок на информационные ресурсы Сети и выдачи результатов поиска пользователю.

Index database (индекс базы данных) – индекс, который является основным массивом данных ИПС и служит для поиска адреса информационного ресурса. Архитектура индекса устроена таким образом, чтобы поиск происходил максимально быстро и при этом можно было бы определить ценность каждого из найденных информационных ресурсов сети.

Queries (запросы пользователя) – сохраняются в его (пользователя) личной базе данных. На отладку каждого запроса уходит достаточно много времени, и поэтому чрезвычайно важно запоминать запросы, на которые система дает хорошие ответы.

Index robot (робот-индексировщик) – служит для сканирования Intranet и поддержания базы данных индексов в актуальном состоянии. Эта программа является основным источником информации о состоянии информационных ресурсов сети.

WWW sites – это весь intranet или точнее – информационные ресурсы, просмотр которых обеспечивается программами просмотра.

Таким образом, ИПС intranet-типа, по сравнению с ИПС Internet, должна обладать рядом отличительных свойств:

- работать с различными форматами данных;
- уметь конвертировать различные форматы данных друг в друга или в HTML;
- самостоятельно определять тип поиска.

Литература

1. Brain Pinkerton. Finding What People Want: Experiences with the WebCrawler. http://www.citforum.ru.

2. Bodi Yuwono, SavioL.Lam, Jerry H. Dik L.Lee. A World Wide Web Resource Discovery System. http://www.citforum.ru.

3. Храмцов П. Открытые Системы. 1996. № 3.

4. Талантов М. КомпьютерПресс. 1999. № 7.