

Тема 1.7 Поиск информации в
Интернет. Поисковые машины



Сеть Интернет растет очень быстрыми темпами, и найти нужную информацию среди сотен миллионов Web-страниц и файлов становится все сложнее.

Для поиска информации используются специальные поисковые серверы, которые содержат более или менее полную, и постоянно обновляемую информацию о Web-страницах, файлах и других документах, хранящихся на десятках миллионов серверов Интернета.

Различные поисковые серверы могут использовать различные механизмы поиска, хранения и предоставления пользователю информации.

Существует два основных вида поиска информации в сети Интернет:

1. Поиск по адресам URL

URL-адрес

URL (*Uniform Resource Locator*) – универсальный адрес документа в Интернете.

<http://city.samara.ru/admin/document/15.txt>

Протокол

Адрес сайта

Путь доступа
к документу (каталог)

Имя файла

Это самый быстрый и надежный вид поиска информации в Интернете. Многие из URL-адресов приводятся в печатных изданиях, специальных справочниках, звучат в эфире популярных радиостанций и с экранов телевизора. Это самый быстрый способ поиска, но его можно использовать только в том случае, если точно известен адрес документа или сайта, где расположен документ.

2. Поисковые системы



В Интернете сосредоточено огромное количество документов. Чтобы облегчить поиск нужной информации, создаются специальные **поисковые машины**.

Поисковые машины - это автоматические системы, опрашивающие серверы, подключенные к глобальной сети, и сохраняющие в своей базе информацию об имеющихся на серверах данных.

По специальному образом сформулированному запросу поисковые машины предоставляют информацию о том, где можно получить необходимые данные.



Как правило, поисковые машины состоят из трех частей: [робота](#), [индекса](#) и [программы обработки запроса](#).

- **Робот (Spider, Robot или Bot)** - это программа, которая посещает веб-страницы и считывает (полностью или частично) их содержимое.

Роботы поисковых систем различаются индивидуальной схемой анализа содержимого веб-страницы.

Индекс - это хранилище данных, в котором сосредоточены копии всех посещенных роботами страниц.

Индексы в каждой поисковой системе различаются по объему и способу организации хранимой информации. Базы данных ведущих поисковых машин хранят сведения о десятках миллионов документов, а объемы их индекса составляют сотни гигабайт. Индексы периодически обновляются и дополняются, поэтому результаты работы одной поисковой машины с одним и тем же запросом могут различаться, если поиск производился в разное время.

- **Программа обработки запроса** - это программа, которая в соответствии с запросом пользователя «просматривает» индекс на предмет наличия нужной информации и возвращает ссылки на найденные документы.

Множество ссылок на выходе системы распределяется программой в порядке убывания релевантности, то есть от наибольшей степени соответствия ссылки запросу к наименьшей.

Контрольные вопросы:

Дайте характеристику следующим поисковым системам:

- *Google;*
- *Yahoo! Search;*
- *Bing;*
- *Ask.com;*
- *Baidu;*
- *AltaVista;*
- *Нигма;*
- *Яндекс;*
- *Рамблер;*
- *Апорт.*