**Лекция №7. Поисковые машины. Механизм работы поисковой машины. Построение индекса**

**Цель лекции:** ознакомиться с механизмом работы поисковой машины. **Содержание лекции:** Поиск по индексу. Механизм работы поисковой машины. Поиск по индексу.

Поисковые машины.

B ответ на запрос вы обычно получаете длинный список документов, многие из которых не имеют никакого отношения к тому, о чем вы спрашивали. Такие документы называются нерелевантными (от англ, relevant - подходящий, относящийся к делу). Таким образом, релевантный документ - это документ, содержащий искомую информацию. Очевидно, что от умения грамотно составлять запрос зависит процент релевантных документов в списке найденных ссылок. Доля релевантных документов в списке всех найденных поисковой машиной документов называется точностью поиска. Нерелевантные документы называют шумовыми. Если все найденные документы релевантны (шумовых нет), то точность поиска составляет 100%. Если найдены все релевантные документы, то полнота поиска - 100%. Таким образом, качество поиска определяется двумя взаимозависимыми параметрами: точностью и полнотой поиска. Увеличение полноты снижает точность и наоборот. Механизм работы поисковой машины. Поисковые системы можно сравнить со справочной службой, агенты которой обходят предприятия, собирая информацию в базу данных. При обращении клиента в службу информация выдается из этой базы. Данные в базе устаревают, поэтому агенты их периодически обновляют. Некоторые предприятия сами присылают данные о себе, и к ним приезжать агентам не приходится. Иными словами, справочная служба имеет две функции: создание и постоянное обновление данных в базе и поиск информации в базе по запросу клиента. Аналогично, поисковая машина состоит из двух частей: так называемого робота, который обходит все предписанные серверы, формируя базу данных, и поискового механизма. У термина робот множество синонимов, помимо робота его также называют сетевым агентом, или, подчеркивая тот факт, что он ползает по сети, - пауком или червяком. База робота в основном формируется им самим (робот сам находит ссылки на новые ресурсы) и в существенно меньшей степени - владельцами ресурсов, которые регистрируют свои сайты в поисковой машине. Помимо робота, формирующего базу данных, существует программа, определяющая рейтинг найденных ссылок. 30 Принцип работы поисковой машины сводится к тому, что она опрашивает свой внутренний каталог (базу данных) по ключевым словам, которые пользователь указывает в поле запроса, и выдает список ссылок, ранжированный по релевантности. Следует отметить, что поисковая система оперирует именно внутренними. Несмотря на то, что база данных поисковой машины постоянно обновляется за счет опроса узловых адресов в Сети, внутренние ресурсы поисковой машины и ресурсы Сети несопоставимы, и поэтому всегда есть вероятность, что машина даст устаревший адрес или не найдет нужный ресурс. Проблема состоит не только в ограниченности внутренних ресурсов, но и в том, что скорость робота ограничена. Увеличение внутренних ресурсов поисковой машины не решает проблемы в силу того, что скорость обхода конечна. При этом нельзя сказать, что поисковая машина внутри имеет копию определенной части исходных ресурсов Интернета, разложенных по каталогу. Полная информация (исходные документы) хранится отнюдь не всегда, чаще хранится лишь ее часть - так называемый индексированный список или индекс, который гораздо компактнее текста документов. Для построения индекса исходные данные преобразуются таким образом, чтобы объем базы был минимальным, а поиск осуществлялся очень быстро и давал максимум полезной информации. Объясняя, что такое индексированный список, можно провести параллель с его бумажным аналогом - так называемым конкордансом, т. е. словарем, в котором в алфавитном порядке перечислены слова, употребляемые определенным писателем, а также указаны ссылки на них и частота их употребления в произведениях писателя. Очевидно, что конкорданс (словарь) гораздо компактнее исходных текстов произведений, и найти в нем нужное слово намного проще, нежели перелистывать книгу в надежде наткнуться на ключевое слово. Построение индекса. Сетевые агенты, или роботы-пауки, «ползают» по Сети, анализируют содержимое WEB-страниц и собирают информацию о том, что и на какой странице было обнаружено. При нахождении очередной HTML-страницы большинство поисковых систем фиксирует слова, картинки, ссылки, скрипты и другие элементы (в разных поисковых системах по-разному), содержащиеся на ней. Причем при отслеживании слов на странице фиксируется не только их наличие, но и местоположение, т. е. где эти слова находятся: в заголовке (title), подзаголовках (subtitles), в метатэгах (meta tags) или в других местах. Обычно фиксируются значимые слова, а союзы и междометия типа «а», «но» и «или» игнорируются. Метатэги позволяют владельцам страниц самим определять ключевые слова и тематику, по которым индексируется страница. Это особенно актуально в случае, когда ключевые слова имеют несколько значений. Метатэги могут сориентировать поисковую систему при выборе из нескольких значений слова на единственно правильное. Однако метатэги 31 работают надежно только в том случае, когда заполняются честно. Недобросовестные владельцы WEB-сайтов помещают в свои метатэги наиболее популярные в Сети слова, не имеющие ничего общего с темой их сайта только для того, чтобы заманить к себе новых посетителей и повысить рейтинг посещаемости своего ресурса. Исключение из поиска подобных сайтов - это еще одна задача, которую должна решать хорошая поисковая система. Каждый робот поддерживает свой собственный список ресурсов, наказанных за недобросовестную рекламу. После сбора информации на предписанных WEB-страницах идет индексация полученных данных. Роботы-пауки просматривают информационное наполнение WEB-страниц и создают индексированную базу поиска по ключевым словам, а затем по запросу пользователя система выдает ранжированный по релевантности список сайтов. Очевидно, что если вы ищете сайты по ключевому слову «цветы», то поисковый механизм должен найти не просто все страницы, где находится слово «цветы», а те, где это слово имеет отношение к теме сайта. Для того чтобы определить, насколько то или иное слово имеет отношение к профилю некоторой WEB-страницы, необходимо оценить, насколько часто оно встречается на странице, есть ли по данному слову ссылки на другие страницы или нет. Короче говоря, необходимо ранжировать найденные на странице слова по степени важности. Словам присваиваются весовые коэффициенты в зависимости от того, сколько раз и где они встречаются (в заголовке страницы, в начале или в конце страницы, в ссылке, в метатэге и т. п.). Каждый поисковый механизм имеет свой алгоритм присваивания весовых коэффициентов - это одна из причин, по которой разные поисковые машины по одному и тому же ключевому слову выдадут различные списки ресурсов. Поскольку страницы постоянно обновляются, то и процесс индексирования должен выполняться постоянно. Роботы-пауки путешествуют по ссылкам и формируют файл, содержащий индекс, который может быть довольно большим. Для уменьшения его размеров прибегают к минимизации объема информации и сжатию файла. После обработки данные сохраняются в базе, которая постоянно обновляется. Имея несколько роботов, поисковая система может обрабатывать сотни страниц в секунду. Сегодня мощные поисковые машины хранят сотни миллионов страниц и получают десятки миллионов запросов ежедневно. При построении индекса решается также задача снижения количества дубликатов - задача нетривиальная, учитывая, что для корректного сравнения нужно сначала определить кодировку документа. Еще более сложной задачей является отделение очень похожих документов (их называют «почти дубликаты»), например таких, в которых отличается лишь заголовок, а текст дублируется. Подобных документов в Сети очень много - например, кто-то списал реферат и опубликовал его на своем сайте под другим заголовком. Современные поисковые системы позволяют решать все эти проблемы. 32 Поиск по индексу. Поиск по индексу заключается в том, что пользователь формирует запрос и передает его поисковой машине. При использовании нескольких ключевых слов весьма полезно применение языка запросов, основу которого составляют булевы операторы. Наиболее часто используемые булевы операторы:

- AND - все термины, соединенные «AND», должны присутствовать в предлагаемом документе. Некоторые поисковые системы используют значок «+» вместо «AND»;

- OR - как минимум одно из ключевых слов, соединенных «OR», должно присутствовать в искомом документе;

- NOT- ключевое слово (слова), следующее за «NOT», не должно появляться в искомом документе. Некоторые поисковые системы используют значок «-» вместо «NOT»;

- FOLLOWED BY - одно из ключевых слов должно следовать непосредственно за другим;

- NEAR - одно из слов должно отстоять на определенное количество слов от другого; - кавычки - слова внутри кавычек - это фрагмент текста, который должен быть найден в пределах документа или файла.

Отметим, что семантика языка запросов может несколько отличаться в конкретных поисковых машинах, но обычно описана в справке поисковой машины. Текст, в пределах которого проверяется логическая комбинация, называется единицей поиска. Это может быть предложение, абзац или весь документ. В разных поисковых системах могут использоваться различные единицы поиска. Поиск в пределах предложения возможен для тех систем, которые имеют в индексе подробный адрес. После того как пользователь передал запрос поисковой системе, она обрабатывает синтаксис запроса, сравнивает ключевые слова со словами в индексе. Затем составляется список сайтов, отвечающих запросу, они ранжируются по релевантности и формируется результат поиска, который выдается пользователю.