М.А. Шумская (Минск, МГЛУ)

ПРИНЦИПЫ ПОСТРОЕНИЯ И ФУНКЦИОНИРОВАНИЯ ПОИСКОВОЙ СИСТЕМЫ В СЕТИ ИНТЕРНЕТ (ЛИНГВИСТИЧЕСКИЙ АСПЕКТ)

Информационно-поисковые системы (ИПС) сети Интернет предназначены для нахождения и выдачи пользователям информации по заданным критериям. ИПС представляют собой совокупность информационно-поискового языка, программных средств и правил перевода текстов на этот язык (индексирования), а также обеспечения поиска необходимых документов и/или данных. В ответ на запросы пользователя информационно-поисковая система выдает список ссылок на информационные источники, которые, по мнению системы, наилучшим образом отвечают потребностям пользователя. При этом списки содержат количество документов, найденных на каждое слово запроса, и условные веса для каждого из документов списка.

Информационно-поисковые системы, при всем их внешнем разнообразии, делятся на три основных класса: систематические, предметные и словарные (алфавитные) [1]. В систематической ИПС используется иерархическая (древовидная) организация информации, которая называется классификатором, а его разделы — рубриками. Классификатор разрабатывается и совершенствуется коллективом авторов. Затем его использует другой коллектив специалистов — систематизаторов. Они читают документы и приписывают им классификационные индексы, указывающие, каким разделам классификатора эти документы соответствуют. Предметная

ИПС с точки зрения пользователя устроена очень просто. Для организации поиска необходимо знать название предмета, поскольку именно с ним связаны списки соответствующих ресурсов сети Интернет. Этот вид систем был бы наиболее удобным, если бы полный перечень предметов был невелик. Основная идея словарной (алфавитной) информационно-поисковой системы заключается в создании словаря из слов, встречающихся в документах сети Интернет. В таком словаре при каждом слове будет храниться список документов, из которых взято данное слово. Схема поиска информации в словарной ИПС проста. Пользователь набирает ключевое слово/словосочетание и активизирует поиск, после чего получает подборку документов по сформулированному ранее запросу. Этот список документов ранжируется по определенным критериям так, чтобы вверху списка оказались те документы, которые наиболее релевантны запросу пользователя.

Принципиальная схема работы классической поисковой системы включает в себя следующие этапы [2]; [3]; [4].

1. Поиск информации поисковыми агентами и составление списка слов сайта. Для этих целей используется программный продукт под названием «спайдер» («паук», «робот»). Спайдер — это программа, которая посещает веб-страницы и считывает (индексирует) их содержимое. Порождаемый спайдером процесс, который углубляет поиск, перемещаясь далее по ссылкам, найденным в документе, называется «краулером». Его основные задачи — отслеживание изменений на веб-страницах и повторное сканирование Интернет-ресурсов через определенные промежутки времени, а также определение того, куда должен идти спайдер, основываясь на найденных ссылках.

2. Индексация всех слов документа и построение базы данных. Все, что находит и считывает спайдер, поступает в индексы поисковой системы, которые представляют собой огромное хранилище информации, состоящее из копий текстовой составляющей всех посещенных и проиндексированных спайдером веб-страниц. На этом этапе работает специальная программа–индексатор, которая анализирует и разбирает такие элементы страницы, как заголовки, ссылки, тексты и т.п. Здесь же происходит построение базы данных, которая представляет собой совокупность структурированных и обработанных индексатором данных.

3. Перенос уже обработанной информации в базу данных. По окончании работы индексатора вся собранная информация помещается в базу данных и сортируется по ключевым словам. Далее база данных сохраняется в удобном для поиска виде.

4. Обработка поискового запроса и сопоставление слов из запроса со словами в базе данных. На данном этапе поисковая система принимает запросы от пользователей, проводит их анализ и извлекает соответствующие запросу результаты из базы данных. Пользователь, желающий найти информацию в Сети, заходит на страницу поисковой системы и вводит в

форму запрос об интересующей его информации. Чаще всего запрос представляет собой набор ключевых слов.

5. Составление списка сайтов, на которых встречается искомая ключевая единица, и их последующая сортировка по определенным правилам с выдачей результатов. Поисковая система находит предмет запроса, основанный на информации, указанной в заполненной пользователем форме, и выводит список соответствующих документов. Для определения порядка следования и сортировки веб-страниц используется алгоритм ранжирования, целью которого является помещение наиболее релевантных запросу пользователя документов первыми в результатах поиска. Каждая информационно-поисковая система имеет свои критерии ранжирования, однако, к наиболее часто используемым можно отнести следующие [5]:

1) наличие слов из запроса пользователем в сетевом документе, их близость друг к другу и количество;

2) наличие слов из запроса пользователя в заголовках и подзаголовках сетевого документа;

3) количество ссылок на данный документ с других документов;

4) авторитетность ссылающихся документов.

В итоге поисковая система выводит ранжированный список результатов поиска сетевых документов. Различные поисковые системы выбирают свои способы представления такого списка: некоторые выводят только ссылки на документы, другие показывают ссылки на документы с первыми предложениями или с его заголовком, третьи выводят ссылки на сетевые документы с их краткой аннотацией.

Как отмечалось выше, используя поисковые системы, пользователь взаимодействует с информационным наполнением Сети посредством запросов. В строке поиска он набирает ключевое слово, словосочетание или фразу и активизирует поиск. Для формирования различных запросов используются специальные (« », |, ~) и математические (*, +, ?) символы. Освоение критериев уточнения запроса и приемов расширенного поиска позволяет увеличить эффективность поиска и достаточно быстро найти необходимую информацию. Сделать поиск более эффективным можно, прежде всего, за счет использования в запросах логических операторов (операций) *or*, *and*, *near*, *not*. С помощью операторов и/или специальных символов пользователь связывает ключевые слова в нужной последовательности, чтобы получить наиболее адекватный запросу результат.

Большинство поисковых машин поддерживают два вида запросов — простой и расширенный. Простой запрос дает большое количество ссылок на документы, которые содержат одно из введенных при запросе слов или простое словосочетание. Оператор *and* позволяет указать на то, что в содержание документа должны быть включены все ключевые слова. Тем не менее, количество документов может быть все еще велико, и их просмотр займет достаточно времени. Поэтому в ряде случаев гораздо удобнее применить контекстный оператор *near*, указывающий, что слова должны располагаться в документе в достаточной близости. Использование *near* зна-

чительно уменьшает количество найденных документов. Наличие символа * в строке запроса означает, что будет осуществляться поиск слова по его маске. Например, если в строке запроса записано *gov**, то пользователь получает список документов, содержащих слова, начинающиеся с буквосочетания *gov — government*, *governor* и т.д. Рассмотрим оба вида оформления запросов на примере англоязычной поисковой системы. Формы запросов приведены в таблице.

Таблица

Виды запросов к поисковой системе

Простой запрос	**Расширенный запрос**	**Расширенный запрос с использованием математических символов**
account	internet merchant account **and** online payments	+internet +merchant +account
merchant account	internet merchant **near**	internet ~merchant ~gov*
internet merchant account	internet merchant **near** education	internet ~merchant ~governor
«merchant account»	—	internet ~merchant ~(governor | account)
«internet merchant account»	—	—

Список ссылок на документы, который поисковая машина выдает в результате обработки запроса, ранжируется по определенным критериям. Причем каждый поисковый инструмент использует свои правила ранжирования ссылок, как при анализе результатов поиска, так и при наполнении индексной базы данных веб-страниц. Таким образом, если для каждой поисковой системы указать в строке поиска запрос одинаковой конструкции, то можно получить различные результаты. Однако в верхней части списка всегда находятся ссылки на те документы, которые в большей степени соответствуют запросу пользователя. Чем выше процент релевантных документов среди всех найденных, тем выше точность поиска. Чем больше процент найденных документов по отношению к общему числу ресурсов, хранящихся в базе данных поисковой системы, тем больше полнота поиска.

ЛИТЕРАТУРА

1. Тактаев, С. Поиск информации в компьютерных сетях: новые подходы. [Электронный ресурс]. — Режим доступа: http://www. taktaev.com. — Дата доступа: 22.11.2014.

2. Факторы ранжирования Google [Электронный ресурс]. — Режим доступа: http://www.seonews.ru/article/.publication/284/. — Дата доступа: 25.11.2014.

3. Tabke, B. Search Engine Theme Pyramids / B. Tabke [Electronic resource]. — Mode of access: http://www.searchengineworld.com/links /theme_engines.htm. — Date of access: 15.10.2014.

4. Миклуха, В. Технология поиска информации в Интернет. Виды поисковых инструментов [Электронный ресурс]. — Режим доступа: http://www.seonews.ru/ master class/16/98/. — Дата доступа: 18.12.2014.

5. Search Engine Copywriting [Electronic resource]. — Mode of access: http://www.redalkemi.com/articles/seo-copywriting-article.php. — Date of access: 30.11.2014.