УДК 004.738.52

ИНФОРМАЦИОННО-ПОИСКОВЫЕ СИСТЕМЫ

Т.Т.КАРИМБАЕВ, И.Э.ТОРГОЕВ, А.А.КУЛЬЖИГИТОВ
E.mail. ksucta@elcat.kg

Издөөчү системалар Интернеттин башкы бөлүгү болуп калды. Бүгүнкү күндө бул - өзүнө информацияны издөө функциясын гана эмес, бизнес үчүн азгыруучусфераны да камтыган атайын механизмдер.

Издөө системаларын колдонуучулардын чоң бөлүгү ошол системалардын иштөө принциби жөнүндө колдонуучулардын суроо-талаптарын иштеп чыгуу схемасы жөнүндө ал системалар эмнеден турушарын жана кантип функцияланышы жөнүндө эч убакта ойлошкон эмес. Бул макалада сиздер ушул суроолорго жооп таба аласыздар.

Поисковые системы стали главной частью Интернета. На сегодня это невероятные механизмы, включающие в себя не только функции поиска информации, но и заманчивые сферы для бизнеса.

Большая часть пользователей поисковых систем никогда не думали о принципе работы этой системы, о схеме обработки запросов пользователей, о том, из чего эти системы состоят и как функционируют. В данной статье вы найдете ответы на эти вопросы.

Search engines have become a major part of the Internet. As of today - it's incredible arrangements, which include not only the ability to search, but also attractive areas for business.

Most users of search engines have never thought about the principle of operation of this system, the schema processing user requests, the fact of what these systems are and how to operate. In this article you will find the answers to these questions.

Информационно-поисковая система – программно-аппаратный комплекс с веб-интерфейсом, предоставляющий возможность поиска информации в Интернете. Под поисковой системой обычно подразумевается сайт, на котором размещен интерфейс системы. Программной частью поисковой системы является поисковая машина – комплекс программ, обеспечивающий функциональность поисковой системы и обычно являющийся коммерческой тайной компании – разработчика поисковой системы. Наиболее крупные международные поисковые системы: «Google», «Yahoo», «MSN», «Яндекс», «Рамблер» /2/.

Рассмотрим подробнее понятие поискового запроса. Для примера возьмем поисковую систему «Google» (рис. 1). Поисковый запрос нужно сформулировать пользователем в соответствии с тем, что он хочет найти, максимально кратко и просто. Допустим, мы хотим найти информацию в «Google» о том, как выбрать ноутбук. Для этого открываем главную страницу «Google» и вводим текст поискового запроса «как выбрать ноутбук». Однако мы можем и не найти нужную нам информацию. В таких случаях нужно перефразировать свой запрос, так как в базе поисковой системы может не оказаться информации по нашему запросу (такое может быть при задании очень «узких» запросов, как, например, «как выбрать ноутбук в Таласе»).

Главная задача поисковой системы – предоставлять людям именно ту информацию, которую они ищут. А научить пользователей делать «правильные» запросы к системе, т.е. запросы, соответствующие принципам работы поисковых систем, невозможно. Поэтому разработчики создают такие алгоритмы и принципы работы поисковых систем, которые бы позволяли находить пользователям искомую ими информацию.

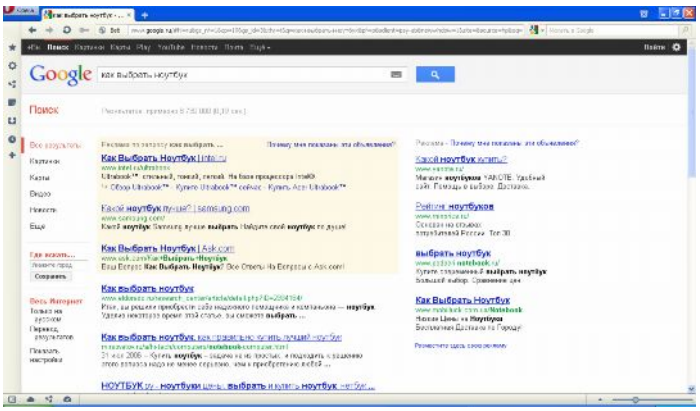

Рис. 1. Поиск информации в «Google.ru»

Улучшение поиска – это одна из приоритетных задач современного Интернета. Разработчики поисковых систем постоянно совершенствуют алгоритмы и принципы поиска, добавляют новые функции и возможности, всячески пытаются ускорить работу системы /4/.

В начальный период развития Интернета число его пользователей было невелико, а объем доступной информации сравнительно небольшим. В большинстве своем доступ к сети Интернет имели лишь сотрудники научно-исследовательской сферы. В это время задача поиска информации в Интернете не была столь актуальной, как в настоящее время.

Практически все крупные поисковые системы имеют свою собственную структуру, отличную от других. Однако можно выделить общие для всех поисковых машин основные компоненты. Различия в структуре могут быть лишь в виде реализации механизмов взаимодействия этих компонентов.

Архитектура современных ИПС для Интернета

Рассмотрим типовую схему информационно-поисковых систем Web (рис. 2).

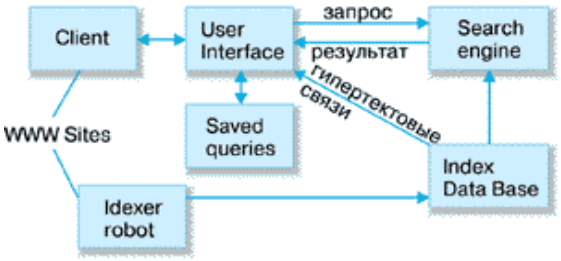

Рис. 2. Типовая схема информационно-поисковой системы

Client (клиент) на этой схеме – это программа просмотра конкретного информационного ресурса (браузеры). В свою очередь, все эти информационные ресурсы являются объектом поиска информационно-поисковой системы.

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 (ИПЯ) в формальный запрос системы, поиска ссылок на информационные ресурсы сети и выдачи результатов этого 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 проходил максимально

быстро и при этом можно было бы оценить 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 нужные ответы.

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

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

Рассмотрим назначение и принципы построения каждого из этих компонентов более подробно и определим, в чем отличие данной системы от традиционной ИПС локального типа.

Индекс поисковой системы

Индекс поисковой системы – это хранящаяся на поисковом сервере база данных, по которой осуществляется поиск запрошенной пользователем информации. Как правило, содержит ссылки на проиндексированные ресурсы и сжатые копии веб-страниц.

Копия страницы в индексе представляет собой инвертированный файл, где для каждого слова, имеющегося в исходном документе, перечислены позиции, в которых оно встречается. Индекс пополняется поисковым роботом во время периодических обходов **Интернета**.

Цель использования индекса – в повышении скорости поиска релевантных документов по поисковому запросу. Без индекса поисковая машина должна была бы сканировать каждый документ в корпусе, что потребовало бы большого количества времени и вычислительной мощности. Например, в то время, как индекс 10 000 документов может быть опрошен в пределах миллисекунд, последовательный просмотр каждого слова в 10 000 больших документов мог бы занять часы. Дополнительное хранилище, требуемое для хранения индекса, а также значительное увеличение времени, требуемого для его обновления, являются компромиссом за экономию времени при поиске информации. Эффективность поиска в каждой конкретной ИПС определяется исключительно архитектурой индекса.

Интерфейс системы

Важным фактором является вид представления информации в программе-интерфейсе. При этом различают два типа интерфейсных страниц: страницы запросов и страницы результатов поиска.

При составлении запроса к системе используют либо меню-ориентированный подход, либо командную строку. Меню-ориентированный подход позволяет ввести список терминов, обычно через пробел, и выбрать тип логической связи между ними. Логическая связь распространяется на все термины. В большинстве систем это просто фраза на ИПЯ, которую можно расширить за счет добавления новых терминов и логических операторов. Но это только один тип использования сохраненных запросов. В традиционных системах это называется расширением или уточнением запроса, в зависимости от того, что получаем в результате преобразования запроса: увеличение размера выборки или ее сокращение. При этом традиционная система хранит не запрос как таковой, а результат поиска, т.е. список идентификаторов документов, который объединяется, пересекается со списком, полученным при поиске документов по новым терминам. К сожалению, сохранение списка идентификаторов найденных документов в Интернете не практикуется. Вызвано это особенностью протоколов взаимодействия программы-клиента и сервера системы, которые не поддерживают сеансовый режим работы.

Информационно-поисковый язык

Информационно-поисковый язык (ИПЯ) – искусственный язык, предназначенный для выражения семантических аспектов информационных источников и запросов в форме, пригодной для осуществления поиска информации. По своим знаковым системам и правилам синтаксиса ИПЯ различаются.

Процесс поиска информации предусматривает взаимодействие в режиме «запрос – ответ» пользователя и информационно-поисковой системы через посредство заранее согласованного ИПЯ. Таким образом, предпосылками для проведения информационного поиска являются:

а) предварительное индексирование информационного массива, т.е. создания поискового образа каждого информационного источника в массиве;

б) перевод информационного запроса пользователя определенного ИПЯ.

Информационно-поисковые языки делятся на два основных типа:

1. ИПЯ классификационного типа.

К языкам этого типа относятся иерархические, алфавитно-предметные и фасетные классификации.

2. ИПЯ дескрипторного типа

Словарь такого языка состоит из фиксированного набора слов и словосочетаний одной или нескольких естественных языков. Таким образом, индексирование информационного источника предполагает создание его поискового образа как определенного набора слов и словосочетаний, которые характеризуют его ключевые содержательные признаки. Методы полнотекстового поиска информации, в основном, предусматривают использование ИПЯ дескрипторного типа.

Популярные поисковые системы

По данным компании Net Applications, использование поисковых систем распределялось следующим образом:

Google – 83,87 %;

Yahoo! – 6,20 %;

Baidu – 4,22 %;

Bing – 3,69 %;

Yandex – 1,7 % (рис. 3);

Ask – 0,57 %;

AOL – 0,36 % /5/.

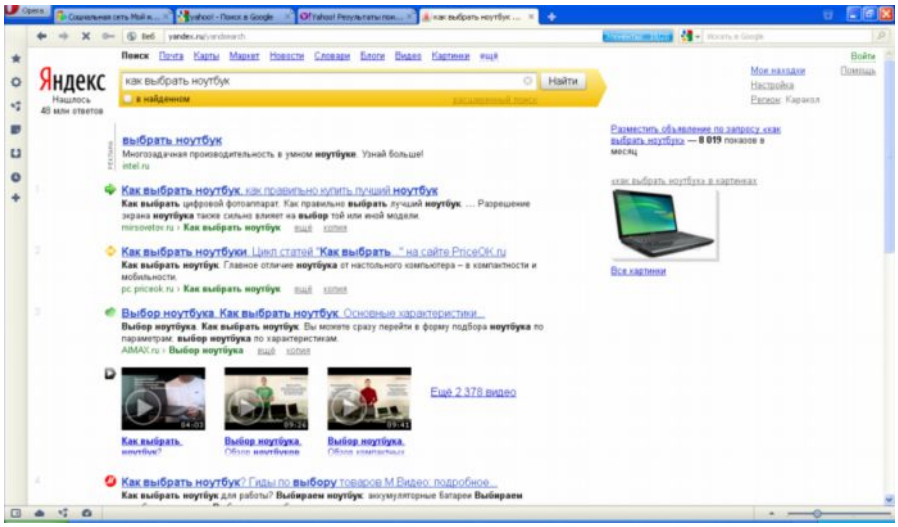

Рис. 3. Поиск информации в «Yandex.ru»

Заключение

Считается, что идеальная поисковая машина должна отвечать следующим требованиям:

1. Простота в использовании.
2. Четко организованный и обновляемый индекс.
3. Быстрый поиск в базе данных и быстрое реагирование.
4. Надежность и точность результатов поиска.

Масштабы информационных ресурсов и их количество постоянно расширяются. Становится ясно, что база данных не является совершенной. Интеллектуальные агенты – новое направление, лежащее в основе нового поколения поисковых машин, которые могут фильтровать информацию и получать более точный результат. Internet продолжает развиваться с неослабевающей интенсивностью, по сути дела стирая ограничение на распространение и получение информации в мире. Однако в этом информационном океане бывает не очень легко найти необходимый документ. Следует также иметь в виду, что в сети наряду с давно действующими серверами возникают новые.

Информационные системы, в которых представлены хранение и обработка информации, осуществляемые с помощью вычислительной техники, называют автоматизированными, они направлены на различные виды деятельности и в наиболее бурно развивающиеся отрасли индустрии информационных технологий.

Список литературы

1. Автоматизированные информационно-поисковые системы http://otherreferats.allbest.ru/programming/00103809_0.html
2. Курсовая работа: «Автоматизированные информационно-поисковые системы». http://bibliofond.ru/view.aspx?id=6973
3. Храмцов П. Информационно-поисковые системы Интернета. http://www.osp.ru/os/1996/03/178885/
4. Информационные поисковые системы http://referats.allbest.ru/programming/9000060837.html
5. Поисковая система. http://ru.wikipedia.org/wiki/Поисковая_система