

СИСТЕМА ПОИСКА INTERNET ДЛЯ ПОЛЬЗОВАТЕЛЕЙ ДОНДТУ

Адамов В.Г., Адамова Е.В., Кулеш С.В, Калиберда И.В.

Донецкий государственный технический университет,

Кафедра АСУ

E-mail: asu@kita.dgtu.donetsk.ua

Abstract

Adamov V., Adamova E., Kulesh S., Kaliberda I. The Global system for searching information on students and teachers in DonSTU is being developed. System consists of three parts: client's part, server part and data collection system. Collected information will be stored within university's server and anyone can access its materials. The searching will be provided due to keyword, entered by user. Also is possible to browse database information, containing our server's links to the necessary information.

Разрабатываемая поисковая система специализируется на информации по научным статьям и рефератам. Она рассчитана в основном на пользование студентами, аспирантами и преподавателями ДонДТУ. Вследствие того, что скорость работы Internet, как правило, оставляет желать лучшего, система существенно поможет сократить временные затраты при поиске нужной информации, так как все данные будут храниться на одном из серверов университета.

В настоящее время при поиске нужной информации в сети Internet возникает ряд проблем. Во-первых, каждый день появляются сотни новых Web-страниц, так что появление новой и интересной информации можно пропустить. Во-вторых, поисковые системы обычно фиксируют только ту информацию, которая содержится в заголовках Web-страниц. Заголовки некоторых страниц не всегда полностью описывают находящуюся на странице информацию. Поэтому результаты поиска не всегда отображают истинную суть информации, находящейся на той или иной Web-странице.

Система Web была разработана для того, чтобы облегчить связь между различными частями информации с помощью гипертекстовых ссылок. Однако количество имеющихся на сегодняшний день ссылок настолько велико, что довольно сложно найти что-нибудь новое и интересное, просто щелкая мышью на ссылках и перемещаясь по сети.

Существует много систем поиска информации в Web. Наравне с поисковыми системами используются сервера, так называемые каталоги, которые хранят различную информацию в виде заархивированных текстовых файлов. Например, популярный среди студенческих масс каталог коллекции рефератов (<http://www.referats.ru/>) представляет собой разбитый по темам сборник рефератов. Каждая тема представляет собой отдельный каталог, содержащий заголовки рефератов и ссылки на них.

Разрабатываемая система поиска будет производить поиск информации (научных статей и рефератов) по ключевому слову. Все работы будут храниться на институтском сервере, и пользователи института смогут получать их за короткий промежуток времени. При этом URL сервера, откуда в свою очередь были взяты эти статьи или рефераты (если они были взяты из Internet) можно также увидеть.

Система состоит из трех частей:

- Клиентской части поисковой системы
- Серверной части
- Системы сбора данных

Структурная схема системы изображена на рис.1

Клиентская часть поисковой системы предназначена для ввода ключевого слова. Для этого разработан удобный интерфейс на языке HTML (Hypertext Markup Language). HTML - язык, предназначенный для создания форматированного текста, который насыщен изображениями, звуком, анимацией и гипертекстовыми ссылками на другие документы, разбросанные по всему Web-пространству.

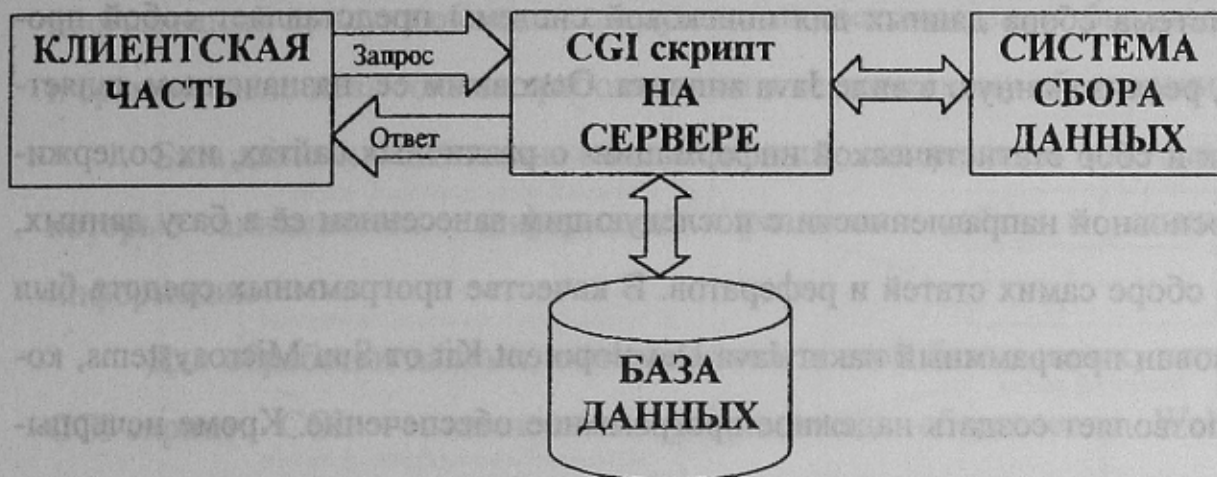


Рисунок 1 - Структурная схема системы поиска

Клиентская часть поисковой системы предназначена для ввода ключевого слова. Для этого разработан удобный интерфейс на языке HTML (Hypertext Markup Language). HTML - язык, предназначенный для создания форматированного текста, который насыщен изображениями, звуком, анимацией и гипертекстовыми ссылками на другие документы, разбросанные по всему Web-пространству.

Несмотря на некоторые ограничения, HTML обладает следующими преимуществами:

- Файлы HTML совместимы с любым браузером. Фактически, файлы HTML являются системно-независимыми и могут отображаться любым компьютером, способным поддерживать HTML.
- Файлы HTML имеют небольшой размер. Накладные расходы на использование тегов весьма не велики и, хотя при хранении HTML файлов не используется сжатие, оно не является по-настоящему необходимым.

Для данной системы поиска разработана программа-обработчик, которая сравнивает введенное ключевое слово с ключевыми словами, хранящимися в базе данных. Если идентичные ключевые слова были найдены, то в браузер из базы данных методом POST выводятся URL серверов Internet, содержащих сведения по данной теме и ссылки на эти статьи на институтском сервере.

Система сбора данных для поисковой системы представляет собой программу, реализованную в виде Java апплета. Основным ее назначением является поиск и сбор статистической информации о различных сайтах, их содержимом и основной направленности с последующим занесением её в базу данных, а также сборе самих статей и рефератов. В качестве программных средств был использован программный пакет Java Development Kit от Sun Microsystems, который позволяет создать надежное программное обеспечение. Кроме исчерпывающей проверки на этапе компиляции, системой предусматривается анализ на этапе выполнения. Компилятор Java производит байт-коды, т.е. модули приложения имеют архитектурно-независимый формат. Это позволяет работать на множестве разнообразных платформ.

Работа системы сбора данных происходит следующим образом: в окне апплета необходимо заполнить все ключевые поля, отвечающие за описание того или иного сайта, кроме этого можно установить с ним связь и одновременно протестировать корректность введенного адреса для подключения. После чего можно непосредственно перейти к передаче собранной информации для ее дальнейшей обработки и занесения в базу данных.

Разрабатываемая система после сбора данных устанавливает связь с находящейся на сервере программой (CGI скриптом) кодирует и передает методом POST всю собранную информацию. На стороне сервера происходит декодирование полученной информации и занесение ее в базу данных поисковой системы.

Серверная часть решает две задачи:

1. Прием и обработка информации, полученной от системы сбора данных с последующим занесением ее в базу.
2. Обработка запроса от клиентской части, формирование и вывод результатов.

На сервере хранится база данных, реализованная с помощью SQL-сервера MySQL. По данным тестирования независимых экспертов этот

программный продукт обладает следующими достоинствами: быстродействием и сравнительно невысокими требованиями к техническому обеспечению.

База данных реализована в виде таблицы, содержащей ключевые поля, в которые записывается информация о различных сайтах и ссылки на эту информацию.

Для обработки данных, содержащихся в этой базе данных, используется CGI-скрипт. CGI — компонент программного обеспечения Web-сервера, который может взаимодействовать с другими программами или базами, находящимися на этом сервере. С помощью CGI Web-сервер может вызвать внешнюю программу и передать в нее пользовательские данные (например, информацию о том, с какой хост-машины пользователь установил соединение, или данные, введенные пользователем в HTML-форму). Эта программа затем обрабатывает полученные данные, и сервер передает результаты ее работы обратно в Web-браузер.

Методика организации взаимодействия приложений CGI скриптов и расширений сервера Web основана на применении классов URL и URLConnection. Приложение Java, желающее работать с расширением сервера Web, создает объект класса URL для связи с исполняемым модулем CGI. Далее приложение получает ссылку на канал передачи данных и пользуясь методами `getOutputStream` и `getInputStream` из класса URLConnection, создает выходной и входной канал передачи данных. Когда данные переданы в выходной канал, созданный подобным образом, они попадают в стандартный поток ввода приложения CGI, как будто бы пришли методом POST из формы, определенной в документе HTML. Обработав полученные данные, расширение CGI записывает их в стандартный выходной поток, после чего они становятся доступны приложениям через входной поток.

На рис. 2. показаны потоки данных для описанной выше схемы взаимодействия CGI скриптов с приложениями.

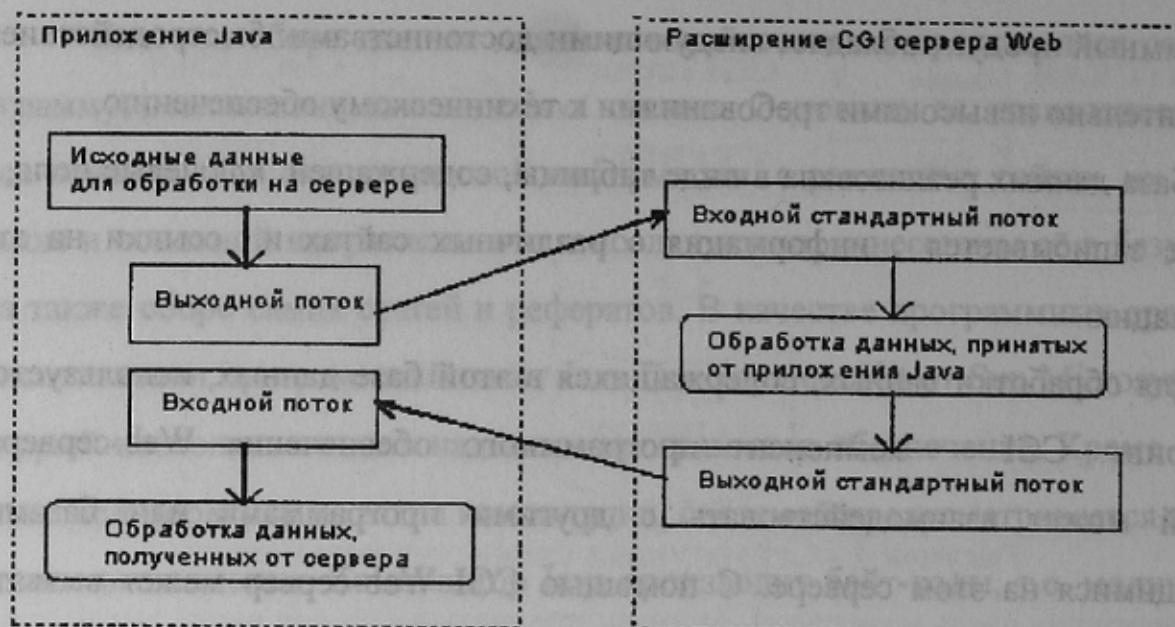


Рисунок 2 - Взаимодействие CGI скриптов с приложениями.

Создание такой системы упростит работу студентов и преподавателей в поиске необходимой научной информации и значительно уменьшит временные затраты на поиск.

В настоящее время эта система находится на стадии завершения. Существует готовые к использованию база данных и клиентская часть. Система сбора данных и серверная часть находятся на стадии завершения.

Литература

1. Пол Дж. Перри. Секреты Word Wide Web
2. Павел Храпцов. Лабиринты Internet. Практическое руководство.
3. С. Спейнаур, В. Каурсиа. Справочник Web-мастера.