УДК 519.7

Р.В. Бобков, Е.А. Коновалов

Харьковский национальный университет радиоэлектроники, Харьков

ПОИСКОВАЯ СИСТЕМА WEB INDEXER

Рассматриваются вопросы разработки поисковой системы. Предлагаемая система является инструментом, обеспечивающим поиск на веб-страницах. Система предоставляет расширенный поиск в интернете, с возможностью поиска по заданным критериям, таким как: поиск по ссылкам, поиск по адресам, поиск по заголовкам и возможность поиска математических формул. А так же улучшенное быстродействие выполнения запросов пользователей, которое достигается путем создания индексации документов. Объем индексной информации, полученной из текста, может быть в два раза больше, чем сам текст.

Ключевые слова: *поисковая система, веб-страница, выполнение запросов пользователей, индексация документов.*

Постановка проблемы

Современный этап развития цивилизации характеризуется переходом наиболее развитой части человечества от индустриального общества к информационному. Одним из наиболее ярких явлений этого процесса является возникновение и развития глобальной информационной компьютерной сети.

С появлением сети Интернет актуальной становится проблема поиска. Интернет – всемирная компьютерная сеть, представляющая собой единую информационную среду и позволяющая получить информацию в любое время. Но с другой стороны в Интернете храниться очень много полезной информации, но для поиска её требуется затрачивать много времени. Эта проблема послужила поводом к появлению поисковых машин [1]. Существуют такие критерии классификации поисковых систем как функции и решаемые задачи. Справочные системы являются наиболее распространенным типом функций информационных систем, и заключается в предоставлении абонентам системы возможностей получения установочных данных на определённые классы объектов. Поисковые системы являются наиболее распространённым классом информационных систем. В общем, виде можно рассматривать как некое информационное пространство, задаваемое в терминах информационно – логического описания предметной области. Расчетные системы заключается в обработке информации, находящейся в системе, по определённым расчётным алгоритмам для различных целей. Технологические функции информационных систем заключаются в автоматизации всего технологического цикла или отдельных его компонентов, производственной или организационной структуры [2].

Разработка поисковой системы

Целью разработки системы является информационный поиск по веб-страницам, осуществляемый по заданным критериям: по заголовкам, по тексту, по ссылкам и поиск математических формул.

Данная система реализована с помощью: ASP.NET (Microsoft .NET Framework 2.0) [3], SQL Server 2005 [4].

Проект WebIndexer состоит из трех модулей. Взаимодействие модулей системы показаны на рис. 1. Модуль WebIndexer производит поиск на веб-страницах, модуль WebSpider ищет ссылки в интеренете и заносит их в базу данных, модуль DBConfigurator осуществляет соединение с базой данных.

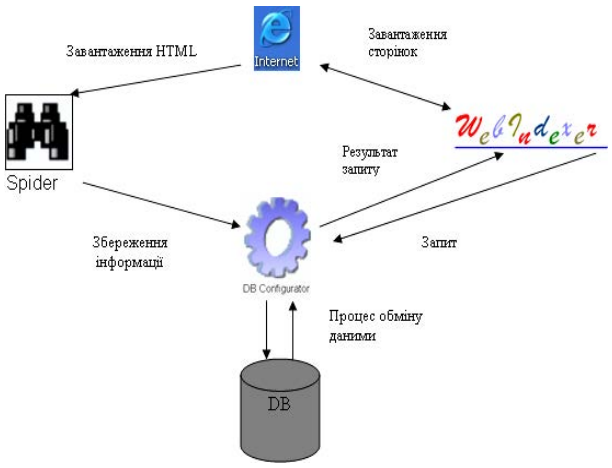

Рис. 1. Диаграмма взаимодействия модулей системы

Разработанная поисковая система предоставляет очень маленькое время выборки конкретной записи из миллионов других. Это достигается созданием так называемого индекса к таблице на какое-то поле из этой таблицы. Обычно индексы реализуются с применением алгоритма сбалансированного двоичного дерева. Индексация страниц происходит в несколько проходов робота. Полная индексация страниц сайта происходит на протяжении 1 недели до 2-3 месяцев в зависимости от организации сайта и его сложности.

Для ускорения индексации страниц возможно подать запрос поисковой системе о том, что страница существует. Для ускорения индексации много поисковых систем имеют специальную форму «До-

бавить URL». Некоторые поисковые системы принимают запросы только на главные страницы сайта, некоторые на все ссылки. Можно также создать страницу с перечнем ссылок для более быстрой индексации. Схематично процесс индексации страницы изображен на рис. 2.

Рис. 2. Схема процесса индексации веб страниц

Общее представление механизма взаимодействия с поисковой системой представлено на рис. 3:

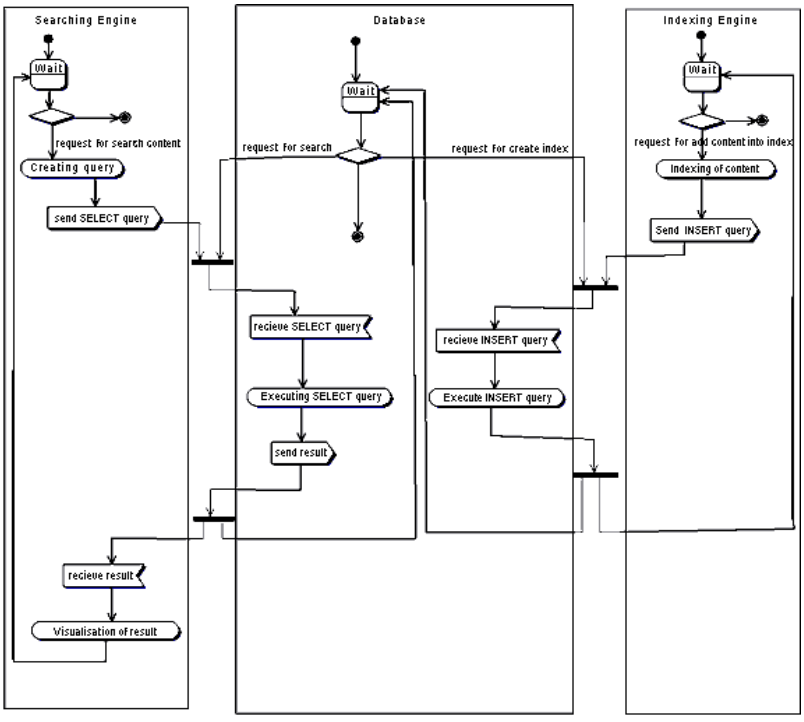

Рис. 3. Механизм взаимодействия с поисковой системой

Как видно из рис. 3, существует три потока управления. Первый обслуживает запросы пользователя, второй выполняет поисковые запросы, а третий занимается индексированием новых документов поступающих в систему. Первый поток – это скрипт на Perl, Servlet, ASP или PHP, который из ключевых слов пользователя формирует поисковые SQL запросы. Второй поток – это СУ базой данных, которая поддерживает целостность данных, индексный механизм и обслуживает SQL запросы. Третий поток – это тоже скрипт, который работает с новыми документами, индексирует их и посылает запросы в базу данных на внесения новой индексной информации.

База данных поисковой системы состоит из четырех таблиц: T_ShortSentens, T_Titles, T_Links, T_Body. Таблица T_ShortSentens содержит короткое описание страниц, T_Titles содержит заголовки страниц, T_Links содержит ссылки на страницы, а таблица T_Body содержит тело страниц. На рис. 4 показаны соединения таблиц в третьей нормальной форме:

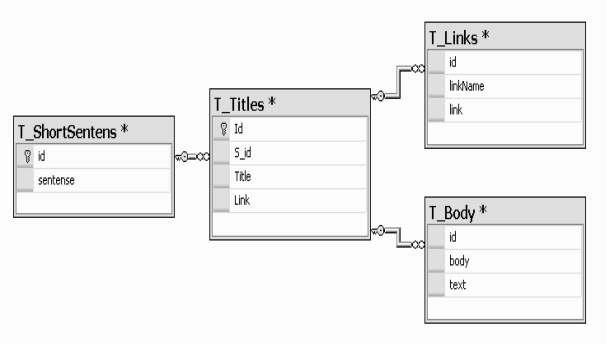

Рис. 4. Связи между таблицами

Диаграмма классов программы WebSpider. Диаграмма классов необходима для наглядного изображения взаимодействия классов в системе. Каждый класс – это набор функциональности с полями данных, которые обрабатываются внутренней логикой классов. Каждый класс имеет наследника или потомка, использование свойств потомка называется наследованием. Расширения возможностей классов за счет использования наследников – полиморфизмом. Данные преимущества объектно-ориентированного программирования позволяют конструировать гибкие системы, возможности к расширению которых неограниченны.

Программа WebSpider является клиент-серверным приложением, реализованным средствами объектно-ориентированного программирования. Диаграмма классов показана на рис. 5.

Класс DBWork необходим для реализации соединения локального приложения с базой данных. Он обрабатывает события соединения с базой данных и закрытия соединения. Для формирования строки соединения наследует класс Regedit, который берет переменные из реестра, заранее сформированные программой DBConfigurator.

Класс Regedit – класс обращения к реестру для считывания параметров соединения с базой данных.

Класс SpiderInterface – класс, который является промежуточным между пользовательским интерфейсом и базой данных. Он наследует класс Spider и класс DBWork. Он необходим для формирования списка

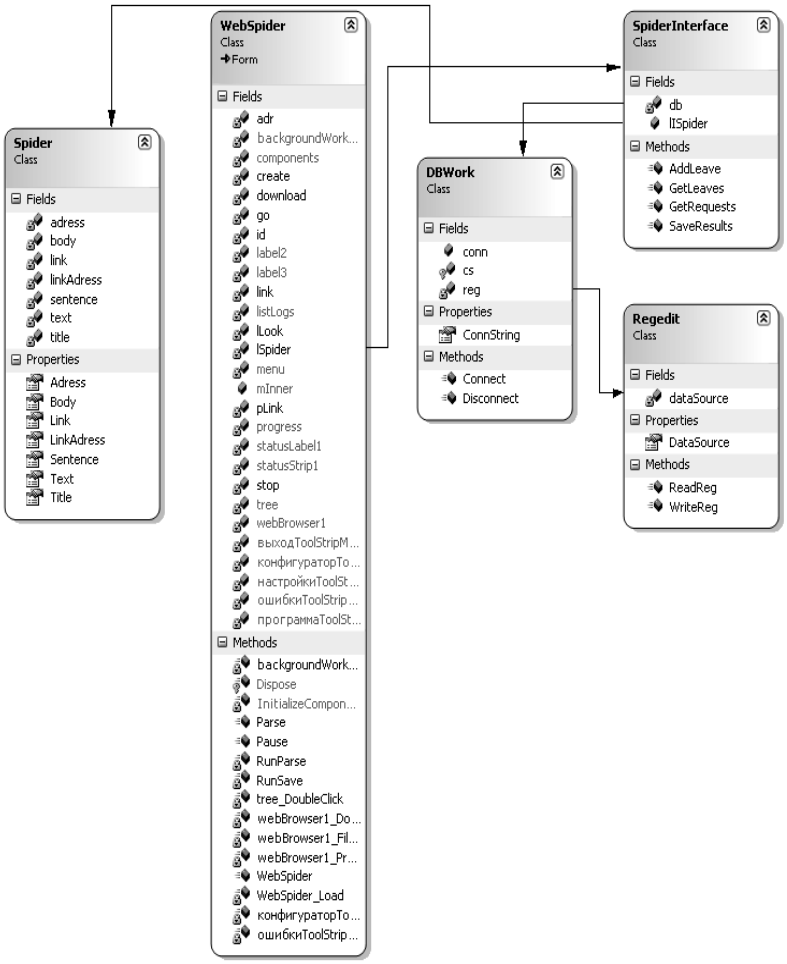

Рис. 5. Диаграмма классов программы WebSpider

считываемых данных и данных для записи, для сохранения промежуточных данных в памяти машины.

Класс WebSpider – основной класс, который наследует класс SpiderInterface, соответственно наследует все остальные классы, т.е. имеет всю функциональность всех вышеперечисленных классов.

Во всех классах используется наследование has-a, т.е. класс предок объявляется как переменная, как поле, в классе-наследнике.

Интерфейс поисковой системы. Когда пользователь обращается с запросом к поисковой машине, он хочет найти то, что ему нужно, максимально быстро и просто. Получая результат, он оценивает работу системы, руководствуясь несколькими основными параметрами. Пользователь должен найти то, что искал, а если не нашел, то он должен переформулировать запрос, чтобы найти искомое. Также учитывается актуальность информации и то, насколько быстро обрабатывает запрос поисковая машина, и насколько должно быть удобно представление результатов поиска.

На странице расширенного поиска находятся поля для поиска со всеми словами, с точной фразой, с любым из слов. Полнота – это одна из основных характеристик поисковой системы, которая представляет собой отношение количества найденных по запросу документов к общему числу документов в Интернете, удовлетворяющих данному запросу. Полнота поиска в большой мере зависит от работы системы сбора и обработки информации. В связи с постоянным ростом количества документов в сети, эта система в первую очередь должна быть масштабируемой.

На странице результата выводится результат поиска по заданному запросу. Вывод результата осуществляется с помощью компонента DataList. Пример работы вывода результат показан на рис. 6.

Скорость поиска тесно связана с его устойчивостью к нагрузкам. На сегодняшний день в рабочие часы к поисковой машине Рамблер приходит около 60 запросов в секунду. Такая загруженность требует сокращения времени обработки отдельного запроса. Здесь интересы пользователя и поисковой системы совпадают: посетитель хочет получить результаты как можно быстрее, а поисковая машина должна отрабатывать запрос максимально оперативно, чтобы не тормозить вычисление следующих.

Сбором информации занимается робот-паук, который обходит страницы с заданными URL и скачивает их в базу данных, а затем архивирует и перекладывает в хранилище суточными порциями. Робот размещается на нескольких машинах, и каждая из них выполняет свое задание. Так, робот на одной машине может качать новые страницы, которые еще не были известны поисковой системе, а на другой – страницы, которые ранее уже были скачаны не менее месяца, но и не более года назад.

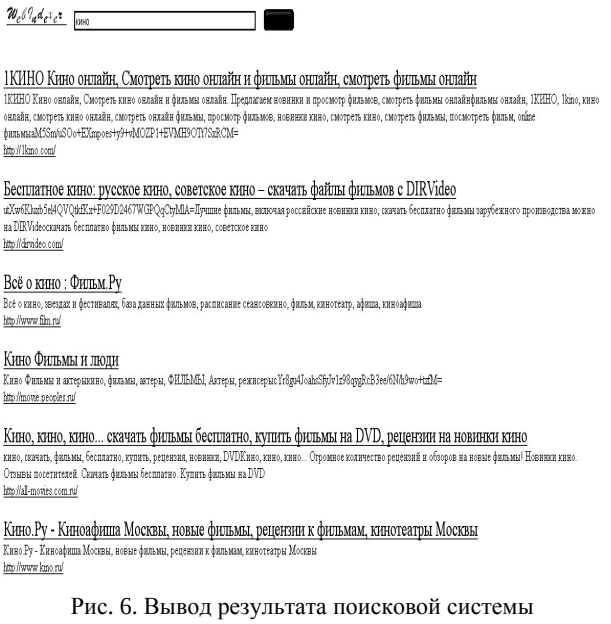

Рис. 6. Вывод результата поисковой системы

Хранилище у всех машин едино. При необходимости работу можно распределить другим способом, например, разбив список URL на 10 частей и раздав их 10 машинам. Параллельная работа программы позволяет легко выдерживать дополнительную нагрузку: при увеличении количества страниц, которые нужно обойти роботу, достаточно просто распределить задачу на большее число машин. Пример работы программы WebSpider приведен на рис. 7.

Рис. 7. Обработка запросов на индексацию в программе WebSpider

WebSpider – администраторская прикладная программа, используется для выполнения запросов на индексацию веб-страниц и для изменения конфигурации налаживания соединения с базой данных для всей системы.

Процесс обработки данных в программе WebSpider заключается в том, что на вход в программу поступает url адрес, по которому происходит загрузка HTML страницы и её анализ и разбивка на составляющие.

Далее проанализированные данные сохраняются в базу данных системы. Весь процесс требует наличие Интернета и много времени, т.к. скорость процесса зависит от скорости Интернета, которая, к сожалению, довольно низка в нашей стране и намного ниже скорости обработки данных в процессоре компьютера.

Выводы

В ходе выполнения работы была разработана программа настройки соединения с установленной базой данных, которая позволяет легко настроить соединение, не изменяя программный код класса работы с базой данных.

Также была разработана программа для обхода веб-страни, их анализа, сохранения результатов этого анализа в системе. И разработана поисковая система, которая выполняет следующие запросы, как поиск по адресу, по ссылкам и поиск математических формул.

Список литературы

1. Таненбаум Э. Компьютерные сети / Э. Таненбаум. – СПб.: Питер, 2002.

2. Справочная информация по сетям ЭВМ и телекоммуникациям [Электронный ресурс]. – Режим доступа к ресурсу: www.index.com.

3. Мэтью Мак-Дональд. Microsoft ASP.NET с примерами на C# 2005 для профессионалов: пер. с англ. / Мэтью Мак-Дональд, Марио Шпуштра. – М.:ООО И.Д. Вильямс, 2006. – 1408 с.

4. Хэндерсон К. Профессиональное руководство по SQL Server: хранимые процедуры, XML, HTML / К. Хэндерсон. – СПб.: Питер, 2005. – 620 с.

5. Дейт К. Введение в системы базы данных, 8–е издание: пер. с англ. [Текст] / К. Дейт. – М.: Издательский дом «Вильямс», 2005. – 1328 с.

6. Сеппа Д. Программирование на Microsoft ADO.NET 2.0. Мастер класс: пер. с англ. [Текст] / Д. Сеппа. – М.: Издательский дом «Русская редакция»; СПб.: Питер, 2007. – 784 с.

Поступила в редколлегию 11.06.2009

Рецензент: д-р техн. наук, проф. Г.Г. Четвериков, Харьковский национальный университет радиоэлектроники, Харьков.

ПОШУКОВА СИСТЕМА WEB INDEXER

Р.В. Бобков, Є.О. Коновалов

Розглядаються питання розробки пошукової системи. Пропонована система є інструментом, що забезпечує пошук на веб-сторінках. Система надає розширений пошук в інтернеті, з можливістю пошуку за заданими критеріями, таким як: пошук по посиланнях, пошук по адресам, пошук за назвою і можливість пошуку математичних формул. А також покращене швидкодію виконання запитів користувачів, що досягається шляхом створення індексації документів. Обсяг індексної інформації, отриманої з тексту, може бути в два рази більше ніж сам текст.

Ключові слова: *пошукова система, веб-сторінка, виконання запитів користувачів, індексація документів.*

SEARCH ENGINE WEB INDEXER

R.V. Bobkov, E.A. Konovalov

The problems of developing a search engine. The proposed system is a tool for searching web pages. The system provides advanced search on the Internet with a search on specific criteria such as: search for the links, search by address, search by title and the opportunity to search for mathematical formulas. As well as improved speed performance of user requests, which is achieved through the creation of indexed documents. Volume index information obtained from the text, can be twice more than the text itself.

Keywords: *searching system, web page, implementation of queries of users, indexation of documents.*