Министерство образования Республики Беларусь

Учреждение образования

Белорусский государственный университет

информатики и радиоэлектроники

УДК 004.738.52

Антоненко
Александр Викторович

Многоагентная система поиска информации по заданным критериям в сети Интернет

АВТОРЕФЕРАТ

на соискание степени магистра технических наук

по специальности 1-40 80 03 Вычислительные машины и системы

_____

Научный руководитель

Самаль Дмитрий Иванович

к.т.н., доцент, доцент кафедры ЭВМ

_____

Минск 2016

ВВЕДЕНИЕ

В соответствии с темой диссертации разработанную систему можно рассматривать в двух ракурсах. В первую очередь это узкоспециализированная многоагентная метапоисковая система в сети Интернет, в качестве предмета специализации которой являются товары. Однако с точки зрения интерфейса пользователя система представляет собой агрегатор товаров.

Суть работы метапоисковых систем заключается в том, что при поисковом запросе параллельно опрашивается несколько независимых систем поиска и возвращаются их результаты одним, объединенным списком результатов без дублирования ссылок, улучшая частные результаты выдачи. Улучшение результатов достигается за счет того, что использование нескольких поисковых систем повышают вероятность обнаружения искомого документа, а также за счет обеспечения возможности выбора тех поисковых систем, которые лучше всего соответствуют текущим потребностям пользователя. Крупномасштабные поисковые системы, такие как Google или Yandex, не могут тратить много времени на обработку каждого отдельного запроса из-за их огромного количества. Метапоисковые системы не имеют такого ограничения и могут фокусироваться на решении специализированных задач поиска. Немаловажным моментом также является и то, что для реализации такой системы не нужно таких огромных ресурсов как для поисковых систем других типов.

Использование многоагентной архитектуры позволило минимизировать время, затрачиваемое на сбор информации в сети. А также способствовало обеспечению гибкости, расширяемости, упрощению решения задач распределения нагрузки между серверами.

Как было отмечено выше, с точки зрения интерфейса пользователя система представляет агрегатор товаров. Агрегатор товаров - это электронная площадка, позволяющая покупателям выбирать, сравнивать и покупать товары и услуги, представленные сразу несколькими интернет-магазинами. Функциональность отдельных агрегаторов товаров ограничена только выбором и сравнением товаров, при этом заказ и оформление покупки происходит на сайте продавца. В качестве примера существующих агрегаторов товаров можно привести market.yandex.com, torg.mail.ru, price.ru и др.

Уникальность разработанной системы заключается в том, что данные в систему попадают путем непосредственного их поиска в сети Интернет. В то время как в существующие агрегаторы данные попадают путем составления прайс-листа строго заданного формата и последующего его импорта в систему агрегатора. Не сложно заметить, что в существующих системах значительно более сложный для владельцев интернет-магазинов процесс обновления цен и добавления новых товаров. Для актуализации данных о товарах владелец

магазина должен обновить файл импорта и снова его импортировать, а до этого момента пользователи будут или видеть некорректные данные или вообще информация о товаре будет отсутствовать.

В разработанной системе информация о товарах собирается из поисковых систем. Процесс сбора информации осуществляется следующим образам. Сначала осуществляется запрос к поисковым системам, которые возвращают набор ссылок. Далее, так как запрос осуществляется к нескольким поисковым системам, то происходит отсеивание дублирующихся ссылок. После фильтрации ссылок происходит извлечение данных по полученным ссылкам и извлечение из них необходимой информации, в текущем прототипе это цена товара. Собранная информация сохраняется в базе данных. Визуализация осуществляется при помощи веб-приложения, которое, на основании собранных данных о товарах, предоставляет список по категориям.

На текущий момент разработан прототип, который осуществляет сбор информации из 2-х поисковых систем, это Google и Rambler. Прототип имеет две категории товаров, это мобильные телефоны и планшеты. В дальнейшем планируется расширить номенклатуру товаров и добавить также список услуг.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность

По результатам предоставленным компаний Yandex, в русском поисковом сегменте, количество запросов приходящихся на покупку/продажу товаров составляет около 1.5 %, при этом количество запросов на покупку в 6 раз больше, чем на продажу. Если учесть что месячная аудитория этой поисковой системы составляет приблизительно 60 млн. человек, то получается что около 800 тыс. человек ежемесячно находится в поиске товаров. Отсюда можно сделать вывод, что заинтересованность людей в удобной системе предоставления данных о товарах достаточно высока.

При сравнительно анализе поисковых систем, можно заметить, что поисковые системы по-прежнему не справляются с индексацией новых документов, так для таких популярных поисковых систем как Google и Yandex показатель степени индексации новых документов в течение 2-х дней составляет 30-40%. Поэтому создание узкоспециализированной поисковой системы, предоставляющей актуальную информацию о товарах более чем актуально.

В сети Интернет существует множество различных агрегаторов прайс-листов, к примеру, yandex.market.*(различные доменные зоны), torg.mail.ru, price.ru, poisk-podbor.ru и т.д. Однако в эти сервисы данные попадают путем составления прайс-листа строго заданного формата и последующего его импорта, тем самым сильно усложняется процесс обновления цен, добавления новых товаров. Таким образом, владельцу интернет магазина необходимо ежедневно обновлять файл для импорта и импортировать его на сайте агрегатора. До этого времени пользователи будут или видеть некорректные данные или вообще информация о товаре будет отсутствовать, несмотря на то, что она содержится на сайте магазина. В худшем случае, владелец должен каждый раз вносить изменения через административную панель сайта вручную. В разработанной системе информация о товарах собирается из поисковых систем автоматически, по запуску планировщика. После получения ссылок с информацией из поисковых систем, происходит проверка их существования, категоризация по доменной зоне, типу товара, фирме производителя и после этого данные сохраняются в системе. Вследствие чего пользователь будет видеть найденные товары с их описанием и ценами и сможет выбрать магазин, который для него предпочтительнее. Использование многоагентной архитектуры позволило минимизировать время, затрачиваемое на сбор информации в сети. А также способствовало обеспечению гибкости,

расширяемости, упрощению решения задач распределения нагрузки между серверами. Таким образом, система обеспечивает актуальность данных для пользователей, а также значительно минимизируют время, затрачиваемое на ее администрирование.

Цель и задачи исследования

Целью диссертационной работы является анализ существующих подходов, алгоритмов и средств поиска информации в сети Интернет. Разработка узкоспециализированной многоагентной системы метапоиска информации в сети Интернет.

Достижение поставленной цели потребовало решения следующих основных задач:

1. Анализ компьютерных систем поиска и метапоиска информации в глобальной сети Интернет;
2. Анализ алгоритмов работы многоагентных систем поиска и обработки информации в сети Интернет;
3. Определение инструментария для программирования системы;
4. Реализация архитектуры и соответствующих алгоритмов и проведение экспериментальных исследований на реальных данных.

Объект исследования: многоагентные системы поиска информации; системы поиска в сети Интернет.

Методы исследования

Теоретические методы исследования основывались на методах поиска и метапоиска информации в глобальной сети Интернет.

Практическая часть основывалась на обработке данных из поисковых систем из сети Интернет и последующим ее ранжировании, разбиение в соответствии с категориями, сохранение и реализации визуального представления.

Для программной реализации разработанных алгоритмов использовались методы создания программных систем и программирование на языках высокого уровня.

Личный вклад соискателя

Основные результаты и положения, выносимые на защиту, получены лично автором. Научный руководитель принимал участие в постановке цели и

задач исследования, их предварительном анализе, планировании, а также в обсуждении полученных результатов.

Опубликованные результаты

По теме диссертационной работы опубликована 1 печатная работа. Из них 1 тезисы доклада на научной конференции.

Структура и объём диссертации

Диссертация изложена на 98 страницах. Она состоит из введения (2 стр.), общей характеристики работы (2 стр.), трёх глав (70 стр.). Работа содержит 26 иллюстраций (8 стр.), список использованных источников, состоящий из 34 наименований (3 стр.).

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Разработанная система представляет собой узкоспециализированную систему поиска информации. В качестве критерия поиска выступает товар. Система представлена тремя модулями. Первый модуль обеспечивает сбор информации из интернет-ресурсов с целью построения дерева товаров. Тем самым отпадает необходимость заполнять все товары вручную. К информация извлекаемой из ресурсов относятся: название товара, его краткое описание и изображение. В текущем прототипе информация собирается для двух категорий товаров это мобильные телефоны и планшеты.

Второй модуль на основании собранных данных первым модулем обращается к поисковым системам для получения информации о ценах на товар, собранные данные сохраняются в базе данных. Обращение к веб-серверам поисковых систем осуществляется параллельно, тем самым обеспечивая более рациональное использование ресурсов системы и ускоряя ее работу. Таким образом, учитывая, что для добавления новых поисковых систем достаточно добавить конфигурационный файл, можно прийти к выводу, что расширяемость системы обеспечивается достаточно просто. Запуск первых двух модулей осуществляется планировщиком задач. Не целесообразно запускать сбор информации чаще 2-х раз в сутки.

Третий модуль обеспечивает визуализацию собранных данных о товарах. Визуализация представляет собой веб-приложение. В качестве платформы был выбран ASP.NET MVC. На рисунке 1 изображен веб-интерфейс модуля визуализации. На рисунке 2 представлена полная информация по одному из товаров.

Помимо практической части была проведена работа по изучению и анализу различных систем поиска и метапоиска, что освещено в первой главе. Изучены виды поисковых систем, проблемы с которыми сталкиваются поисковые машины, уровень покрытия сети Интернет наиболее популярными поисковыми системами. А также рассмотрены некоторые алгоритмы их работы. Особое внимание было уделено метапоисковым системам, так как разработанная система использует метапоиск для получения необходимых данных о товарах.

Во второй главе был проведен анализ алгоритмов работы многоагентных систем и обработки информации в сети Интернет. Освещены основные признаки, которыми должны обладать многоагентные системы, типы архитектур, подробно рассмотрена структура типовой многоагентной поисковой системы, с описанием агентов и их обязанностей.

В третей главе представлена архитектура системы как с точки зрения многоагентной организации, так и с точки зрения взаимосвязи компонентов,

т.е. взаимосвязи уровней интерфейса пользователя, безнес-логики и доступа к данным. Подробно рассмотрена работа модулей. Приведены основные алгоритмы, описаны проекты, из которых состоит система и их взаимосвязь. Рассмотрены основные классы и интерфейсы. Детально рассмотрена структура конфигурационных файлов, определяющих базу знаний, которыми должны обладать поисковые агенты и агенты извлечения ссылок.

При проведении интеграционного тестирования критических недочетов не выявлено.

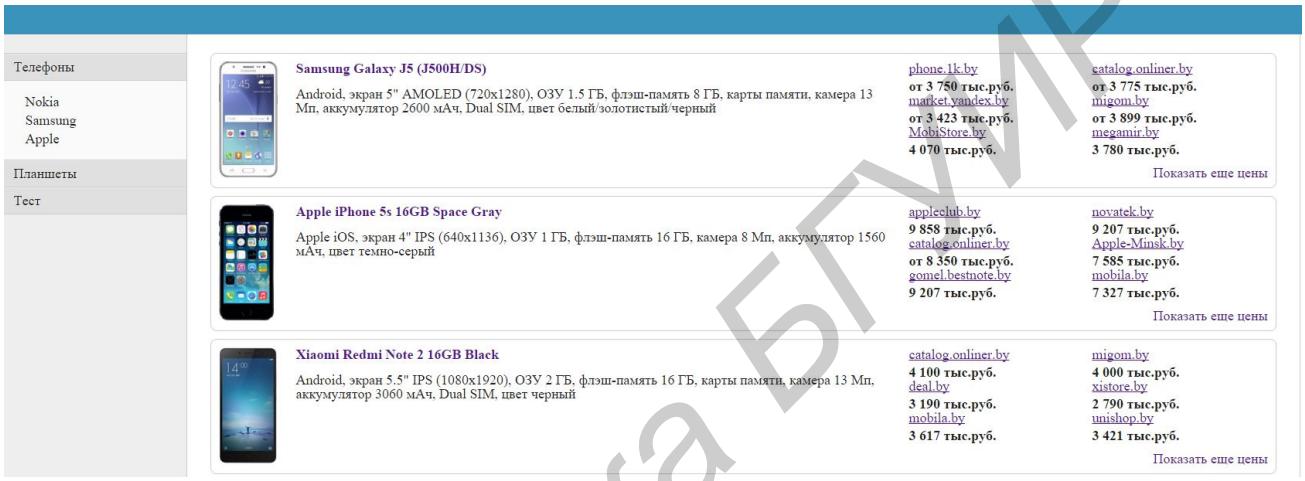

Рисунок 1 – Интерфейс веб-приложения

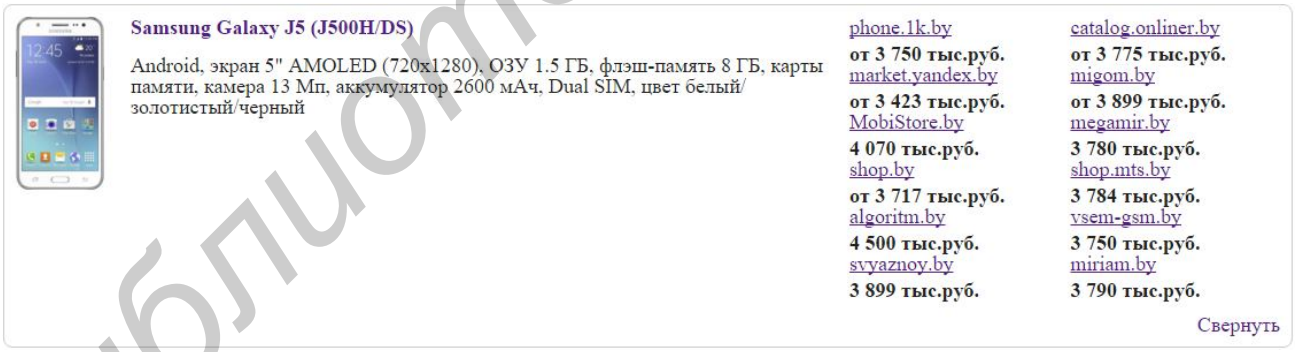

Рисунок 2 – Информация о товаре с полным списком ссылок на интернет-магазины

ЗАКЛЮЧЕНИЕ

Как уже отмечалось ранее, уже существуют агрегаторы товаров, но данный проект уникален тем, что добавление товаров происходит автоматически, при помощи сбора информации из поисковых систем, что значительно упрощает администрирование системы и позволяет уменьшить время, затрачиваемое на сопровождение такого рода системы. При этом система собирает данные, в том числе и из уже существующих агрегаторов.

Система представляет собой три модуля:

– модуль визуализации;

– модуль построения дерева товаров;

– модуль извлечения данных из поисковых систем.

Модуль построения дерева товаров позволяет заполнить каталог товаров на основании данных полученных из существующих интернет магазинов. После извлечения данные сохраняются в базу данных.

Модуль извлечения данных из поисковых систем, на основании сформированного дерева товаров опрашивает поисковые системы, используя названия товаров полученные модулем построения дерева товаров. По ссылкам полученным из поисковых систем осуществляется сбор информации о стоимости товара, после чего информация сохраняется в базу данных.

Модули построения дерева товаров и модуль извлечения данных из поисковых систем запускаются из консольного приложения при помощи планировщика задач.

Для визуализации был использован ASP.NET MVC Framework, который способствует расширяемости системы, а также значительно облегчает процесс тестирования.

Спроектированная архитектура представляет собой набор модулей, каждый из которых может быть заменен другой реализаций без изменений остальных, что обусловлено способом построения архитектуры, а также использованием DI контейнера.

В силу того, что система является многоагентной, многие операции, в частности сбор данных из поисковых система, сбор данных из интернет-ресурсов для создания дерева товаров, проверка существования страниц, осуществляется параллельно и не зависимо друг от друга, что соответствует современным тенденциям развития многоядерных процессоров. Это также позволило минимизировать время, затрачиваемое на сбор информации в сети. А также способствовало обеспечению гибкости, расширяемости, упрощению решения задач распределения нагрузки между серверами. Таким образом, система обеспечивает актуальность данных для пользователей, а также значительно минимизируют время, затрачиваемое на ее администрирование.

Помимо того, что выполнение основных операций осуществляется параллельно, система обладает еще рядом достоинств. Так, время сбора информации из разных поисковых систем различно, это связано со скоростью ответа серверов поисковых систем, форматом возвращаемых данных, и как следствие способом анализа данных и прочих факторов. Несмотря на это потоки завершившие выполнение операции не блокируются, так как они извлекаются из пула потоков и после завершения операции снова возвращаются в пул и могут использоваться для выполнения других операций. Использование пула способствует тому, что нет необходимости создавать дополнительные потоки, затрачивая на это процессорное время, к тому же наличие дополнительных потоков увеличивает частоту переключения контекста процессоров, что также отрицательно влияет на скорость выполнения программы.

Для обеспечения корректности работы в случае непредвиденных ситуаций, так как операция сбора данных осуществляется параллельно, предусмотрено ряд механизмов отслеживания. Во-первых, на различных этапах работы выводятся сообщения в консоль и в файл логирования. Во-вторых, предусмотрен механизм прекращения выполнения работы потоками по таймауту. В-третьих, пользователь может ввести команду остановки.

Причина длительности сбора информации, для одной категории товара заключается в следующем. К примеру, для категории мобильные телефоны существует примерно 1100 моделей. Как отмечалось, поиск информации для каждой из поисковых систем осуществляется параллельно, т.е. для каждой поисковой системы для категории мобильные телефоны происходит 1100 запросов. Для того чтобы поисковые системы не посчитали чрезмерной нагрузку с одного IP адреса, между запросами предусмотрена установка задержки, которая выставляется в конфигурационном файле. Проверить минимальную величину задержки можно опытным путем, на данный момент в тестовых целях была выставлена задержка в 1 секунду. В итоге сбор информации по одной категории составляет примерно 25 минут. Для одновременного сбора информации для разных категорий существует множество способов динамически менять IP адрес.

Стоит отметить, что на данный момент создан прототип системы, в дальнейшем планируется собирать информацию из большого количества поисковых систем, что увеличит релевантность получаем данных. В прототипе данные собираются по двум категориям товаров – мобильные телефоны и планшеты, в дальнейшем планируется расширить номенклатуру товаров и добавить также список услуг.