Министерство образования и науки РФ
Санкт-Петербургский политехнический университет Петра Великого
Институт компьютерных наук и технологий
Высшая школа программной инженерии

Отчет по лабораторной работе
по дисциплине «Технологии разработки качественного программного обеспечения»

Реализация программы-паука для скачивания всех изображений с сайта

Выполнил
студент гр. в3530904/70321 А.В. Шипунов
Руководитель
доцент, к.т.н. Н.Г. Смирнов

Санкт-Петербург
2020

Содержание

Цель работы

Цель данной работы заключается в изучении принципа многопоточных программ и принципа работы протокола HTTP на языке программирования Java на примере разработки многопоточного паука для скачивания всех картинок с заданного сайта.

Необходимо:

1) Дать общий обзор предметной области;
2) Выполнить сравнение функционала аналогичных по назначению программ;
3) Реализовать многопоточного паука на языке программирования Java;
4) Описать архитектуру разработанного программного обеспечения.

Введение

Всем привычная технология работы в Интернет в режиме online далеко не всегда оправдана, поскольку требует значительного времени пребывания в сети (что может быть достаточно дорого, особенно в случае отсутствия доступа в интернет по широкополосному каналу) и всецело привязывает к скорости загрузки информации. Кроме того, в дневное время нагрузка на сеть интернет значительно выше, чем в ночное, что может привести к еще большему ухудшению скорости работы.

Работа же в режиме offline имеет свои неоспоримые преимущества: при просмотре тех же страниц тратится значительно меньше времени (особенно в случае большого количества медиа-файлов), экономится время за счет более быстрой загрузки, а самое главное – не используется подключение к интернету. Основным инструментом работы со страницами в отключенном от интернета режиме являются offline-браузеры. Сейчас их появилось достаточно большое количество. Однако большая часть таких программ – платная, что ограничивает их использование.

Сравнение похожих по функционалу программ

Webripper

Эта программа используется для загрузки с сайтов картинок, видео и других файлов. Например, можно использовать приложение, наподобие Webripper, которому поручить загрузку всех файлов с сайта, имеющих определенное расширение из категории видео. Программа, работая в фоновом режиме, через какое-то время выдает вам на жестком диске всю коллекцию.

К сожалению, программа не позволяет указывать расширения файлов вручную. Изображения – это JPEG, GIF, PNG и BMP, видео – MPEG, AVI, Quick Time, Real Video и Windows Media. Звук – это MP3, WAV, снова Windows Media, Real Audio и Midi, документы – MS Office (Word, Excel и Power Point), PDF, TXT и RTF. В качестве остальных документов предлагается EXE, ZIP, Macromedia Flash и Torrent. Все перечисленные файлы поддерживаются Webripper, их можно загружать с сайтов.

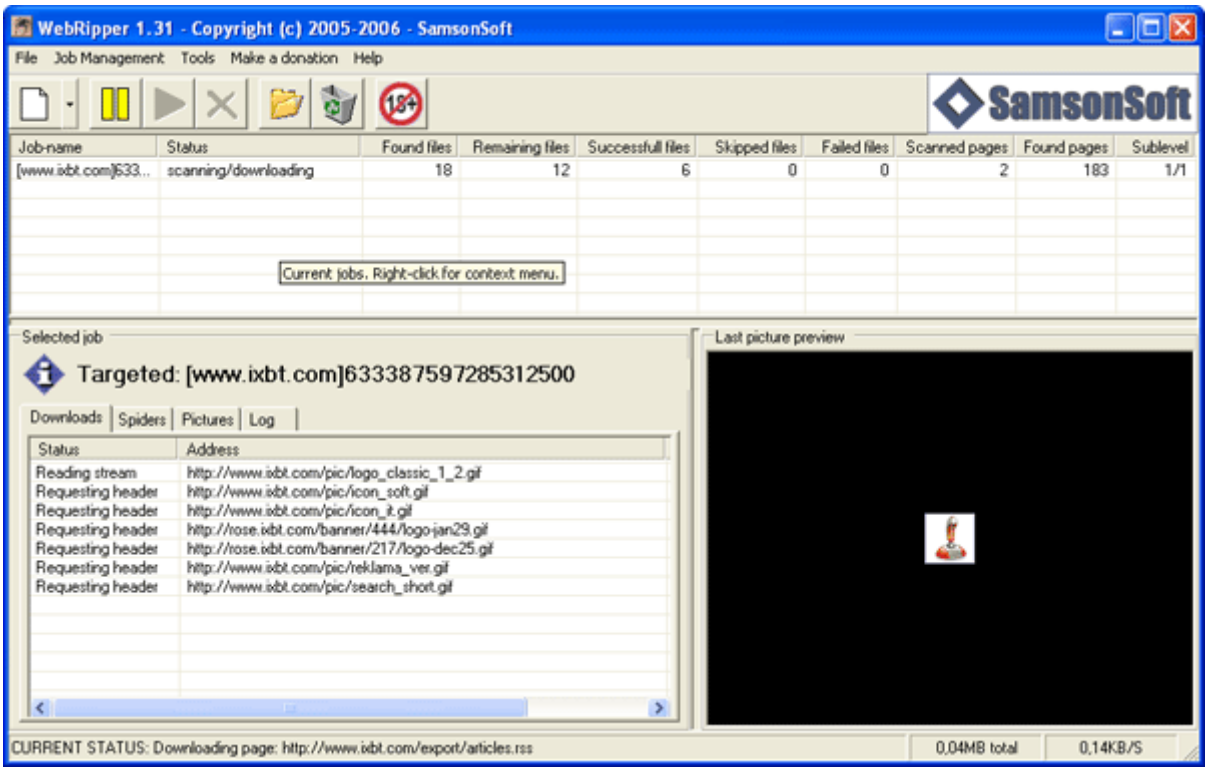

Рис. 1. Главное окно Webripper

Подобно классическим offline-браузерам, можно указывать глубину обработки ссылок. В качестве дополнительного ограничения выступает текущий домен, директории сервера, расположенные ниже от исходного адреса. Впрочем, можно не использовать никаких ограничений.

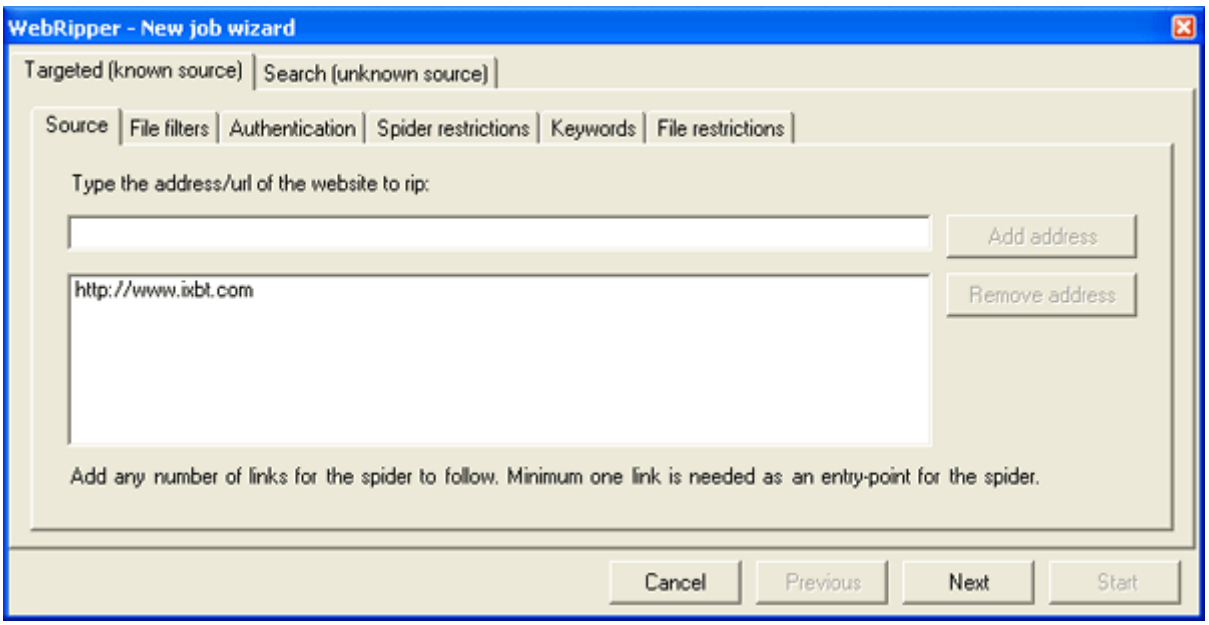

Рис. 2. Создание нового проекта в Webripper

Файлы могут загружаться не только с заранее определенных адресов, но и как результат поиска. Критериями поиска является URL и имя файла. Вы создаете список запросов, который делится на две категории – включение и исключение.

Дополнительные ограничения включают в себя указание минимального и максимального размера файлов. Применительно к изображениям, вы можете указывать диапазон их размеров. Например, при загрузке фотографий нет смысла получать маленькие картинки. Это, скорее всего, элементы дизайна сайта, а не высокохудожественные работы.

HTTrack

HTTrack – бесплатный offline-браузер, который распространяется с открытыми исходными текстами.

Левая боковая панель фактически является навигатором по файловой системе. Файлы показываются без значков. Правая же часть окна отводится, в первую очередь, под работу Мастера, обеспечивающего ввод данных проекта. Кроме того, во время загрузки сайта, в правой части окна отображается весь процесс передачи данных.

HTTrack не имеет встроенного браузера. Все ранее загруженные проекты запоминаются программой, а команда «просмотр сайтов» открывает браузер, установленный в системе по умолчанию. В нем открывается страница, которую генерирует HTTrack. Документ содержит список ссылок на все ранее загруженные сайты. Ссылки локальны. Библиотека сайтов имеет категории. Вы

можете просматривать не только линейный список всех проектов, но также открывать отдельные группы.

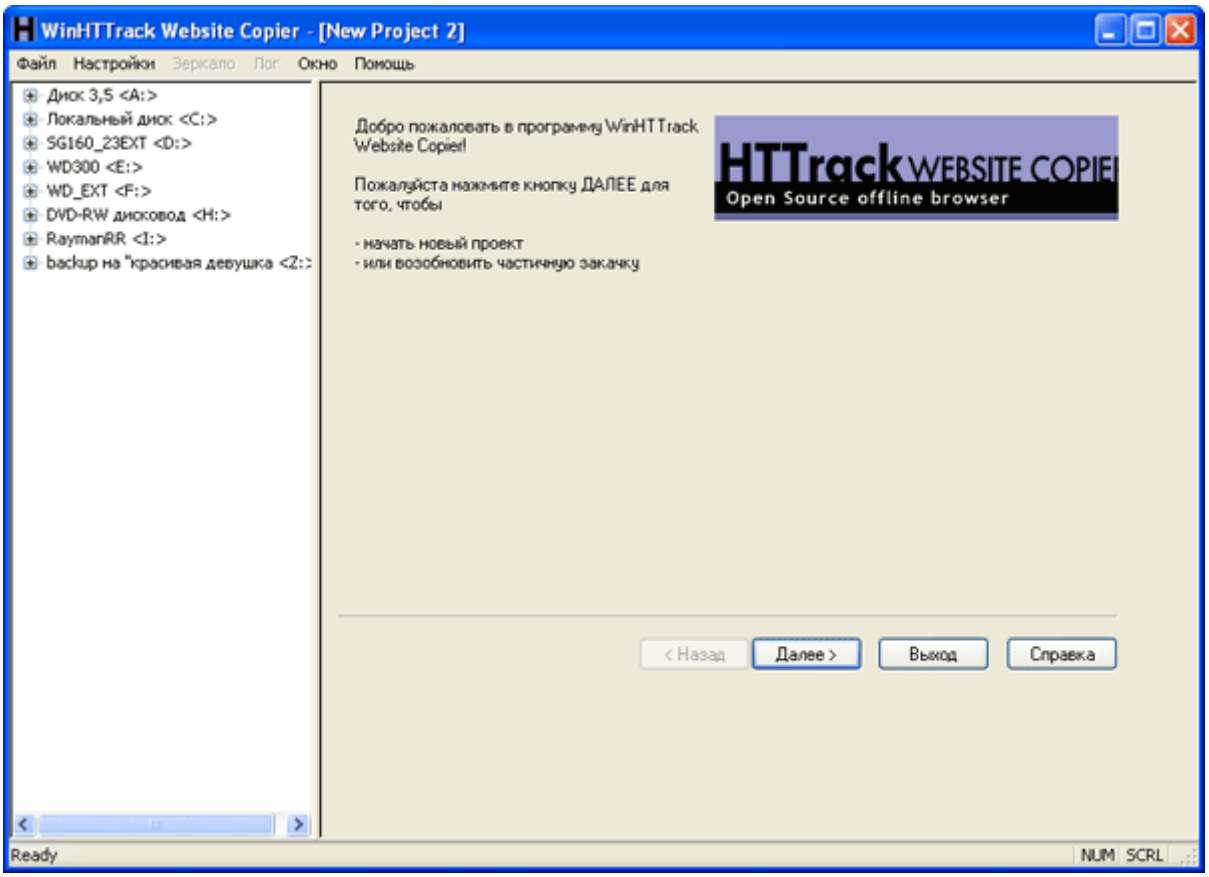

Рис. 3. Стартовое окно HTTrack

Еще одно необычное свойство программного продукта заключается в доступе к настройкам проекта. HTTrack позволяет создавать проекты только через Мастера, и полный список опций вызывается именно оттуда. Первый шаг заключается во вводе имени проекта, а также в указании его категории. Если уже существуют готовые примеры, то можно выбрать соответствующий пункт выпадающего меню. Указав каталог сохранения сайта, можно переходить к следующему шагу.

Первым делом, необходимо указать тип проекта. Можно начать загрузку сайта, продолжить ее с прерванного места, обновить сайт, проверить ссылки на актуальность. Проект может состоять из нескольких URL. Иными словами, имеется возможность загрузки нескольких сайтов с едиными настройками. Мало того, адреса даже не обязательно вводить вручную – достаточно указать файл их списком, URL List.

Первым делом, необходимо указать тип проекта. Можно начать загрузку сайта, продолжить ее с прерванного места, обновить сайт, проверить ссылки на

актуальность. Проект может состоять из нескольких URL. Иными словами, имеется возможность загрузки нескольких сайтов с едиными настройками. Мало того, адреса даже не обязательно вводить вручную – достаточно указать файл их списком, URL List.

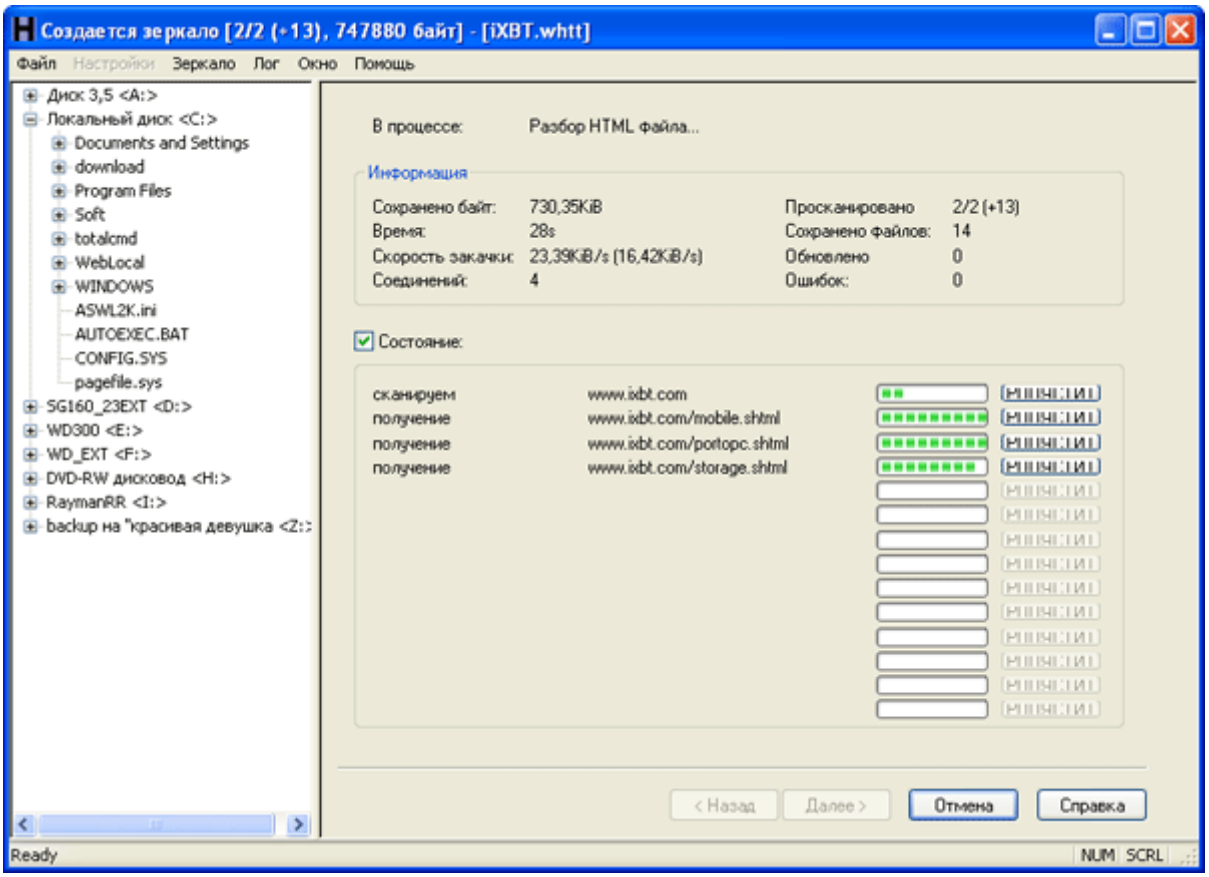

Рис. 4. Загрузка проекта с помощью HTTrack

Фильтрация файлов осуществляется на основе списков включения и исключения. Предлагаются три предустановленные группы – графика, архивы и мультимедиа. В качестве фильтров можно использовать расширения файлов, а также маски адресов.

Глубина обработки ссылок задается раздельно для внутренних и внешних страниц (с других сайтов). Ограничения на размер файла задаются отдельно для HTML и всех остальных типов. Имеется возможность также установить максимальное время загрузки файла. Программа позволяет ограничивать скорость передачи данных, а также количество новых соединений в секунду. Впрочем, можно использовать и традиционную опцию – ограничение максимального числа соединений.

Еще одна возможность HTTrack – управление структурой зеркала сайта. Это означает, что вы можете помещать HTML в одну папку, а картинки в другое.

Программа предлагает пятнадцать предустановленных вариантов распределения файлов. Кроме того, можно использовать собственные варианты, вводя управляющую строку с переменными, описание которых дается в том же диалоговом окне.

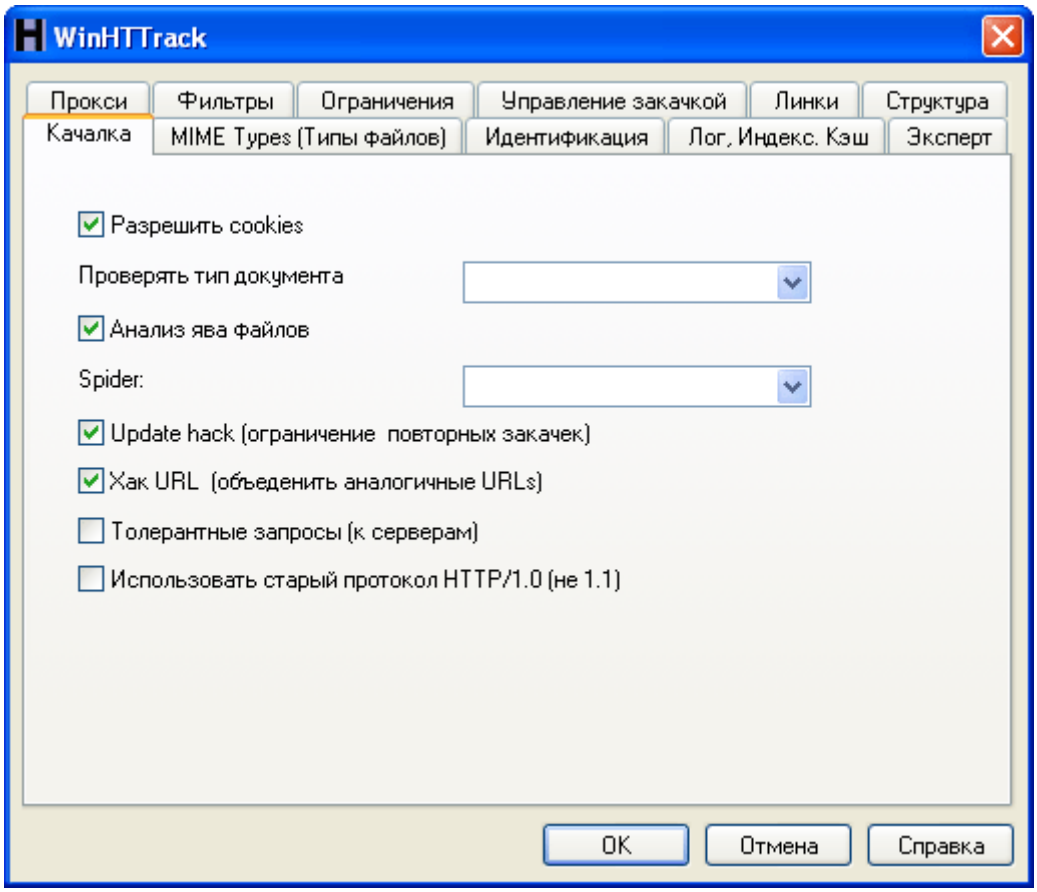

Рис. 5. Свойства проекта HTTrack

Программа может представляться как множество современных браузеров. Всего имеется тридцать вариантов строки идентификатора. Загружаемые файлы могут дополнительно записываться в кэш HTTrack.

На последнем этапе работы Мастера вы указываете тип соединения с Сетью, после чего можно приступать к непосредственному получению данных.

HTTrack – очень необычная программа, обладающая массой оригинальных функций. Имеется русификация. Язык выбирается при первом старте. Качество русского перевода очень низкое. Впрочем, смысл опций, в любом случае, легко угадывается.

Teleport Pro

Teleport Pro является классическим примером минимализма, но не в ущерб функциональности. Программа лишена красот и многих второстепенных возможностей, но создавать локальные копии сайтов ей под силу.

Рабочее окно приложения напоминает Проводник. Левая боковая панель предназначается для того, чтобы просматривать структуру загружаемого сайта. Справа отображается список файлов текущей директории. Панель инструментов, помимо традиционного назначения, связанного с размещением кнопок популярных функций, является своеобразным индикатором сетевой активности приложения. Количество точек равняется числу соединений в данный момент.

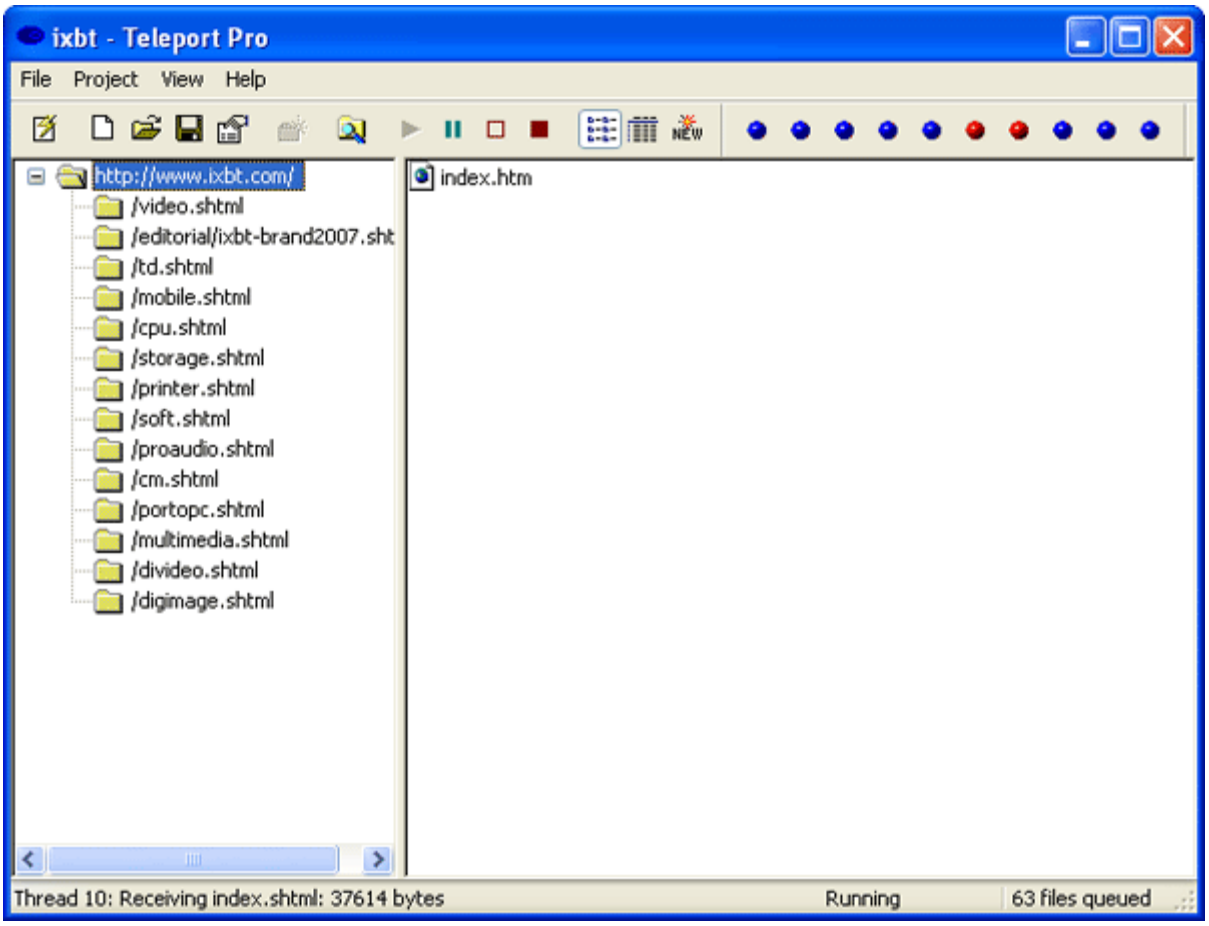

Рис. 6. Главное окно Teleport Pro

Создание нового проекта возможно в двух режимах. Один из них использует настройки по умолчанию – нужно просто указать адрес сайта, после чего сразу запустится процесс скачивания. Более широкие возможности копирования открываются при использовании мастера. Он предлагает серию заранее определенных типов операций. Можно создать копию сайта на жестком диске с целью локального просмотра, полностью дублировать сайт, осуществить

поиск файлов на сайте, загрузить все сайты по ссылкам с указанного документа, загрузить все файлы с указанного сайта, а также осуществить поиск сайта по ключевым словам. Во многом перечисленные режимы пересекаются. Если бы дело обстояло иначе, проще было бы создать несколько программных продуктов, вместо интеграции всех режимов в едином мастере. Режимы, конечно, имеют отличия, но их суть едина – загружается информация с сайта с целью ее локального использования.

Сначала нужно ввести адрес сайта и глубину обрабатываемых ссылок. Далее выбирается группа загружаемых файлов – текст, графика, звук и все остальное. На этом, по сути, и заканчивается работа мастера.

Тонкая настройка проекта производится из его свойств. Здесь можно уже указывать диапазон размеров файлов для каждого расширения по отдельности. Фильтрация основывается на вводе ключевых слов, которые встречаются на страницах. Присутствует лишь список исключения.

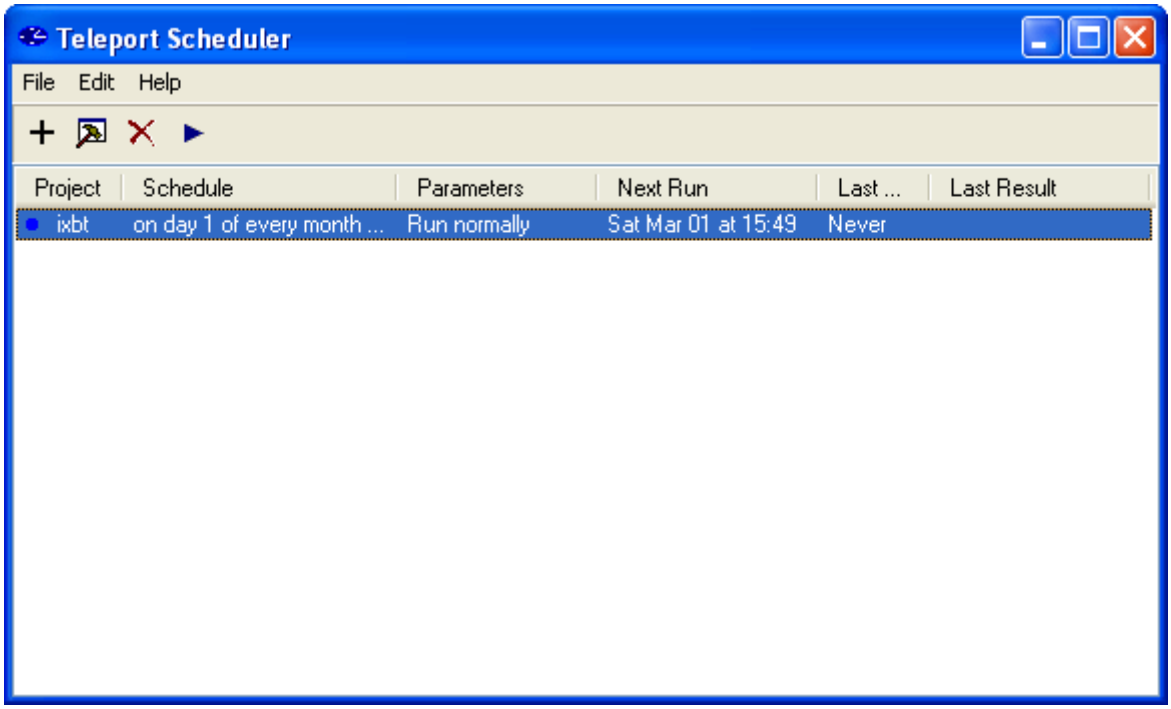

Рис. 7. Планировщик Teleport Pro

В состав Teleport Pro включен планировщик. Он позволяет запускать проекты при старте системы, ежечасно, ежедневно, ежемесячно и один раз, в определенное время. После завершения задачи можно разорвать сетевое соединение.

Описание и обоснование архитектуры приложения

Данное программное обеспечение построено по модульному принципу с использованием принципов объектно-ориентированного программирования. Для этого у всех основных компонентов системы выделены соответствующие интерфейсы, использование которых осуществляется с помощью Dependency Injection.

Загрузка HTML-файлов с указанного сайта и последующая их обработка с целю поиска URL-адресов осуществляется с помощью соответствующего сервиса, после чего возвращается список всех найденных на веб-странице URL-адресов.

В общей для всех рабочих потоков потокобезопасной коллекции хранится актуальный список найденных и еще не просмотренных URL-адресов, при этом в момент запуска программы в нее добавляется указанная пользователем начальная страница. Каждый рабочий поток сначала берет следующую не обработанную страницу, загружает найденные на ней картинки и добавляет в общую коллекцию новые найденные URL-адреса, после чего поток возвращается к выбору следующей не обработанной веб-страницы. Это позволяет максимально равномерно загружать все рабочие потоки.

При запуске приложения стартует указанное пользователем количество потоков. Критерием окончания работы программы служат отсутствие необработанных URL-адресов и завершение всеми рабочими потоками загрузки веб-страниц. Информация о текущем прогрессе выводится отдельным потоком по таймеру каждые три секунды. Для сохранения всех ошибок подключен логгер.

Пользователь может указать настройки программы (начальная страница, количество потоков, минимальный размер файла и т.д.) в конфигурационном файле.

При этом хочется отметить, что многопоточное приложение работает значительно быстрее однопоточного варианта. В частности, удалось увеличить скорость работы более чем в 15 раз по сравнению с однопоточным вариантом.

Табл. 1. Объем загруженных картинок за 10 минут работы программы

Кол-во потоков	1	2	3	4	8	16	32	64
Скачано МБ за 10 мин.	62	115	148	208	510	718	951	910

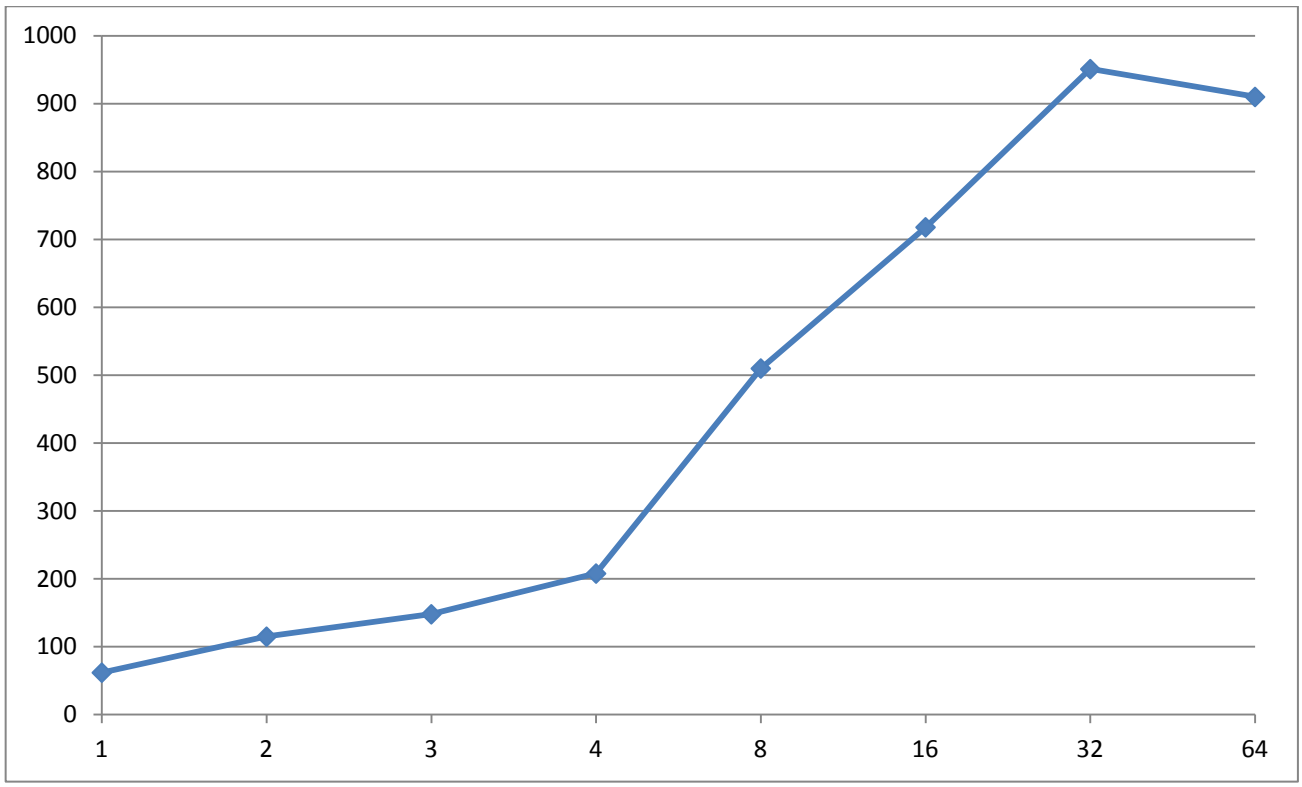

Рис. 8. Зависимость объема загруженных картинок от количества потоков

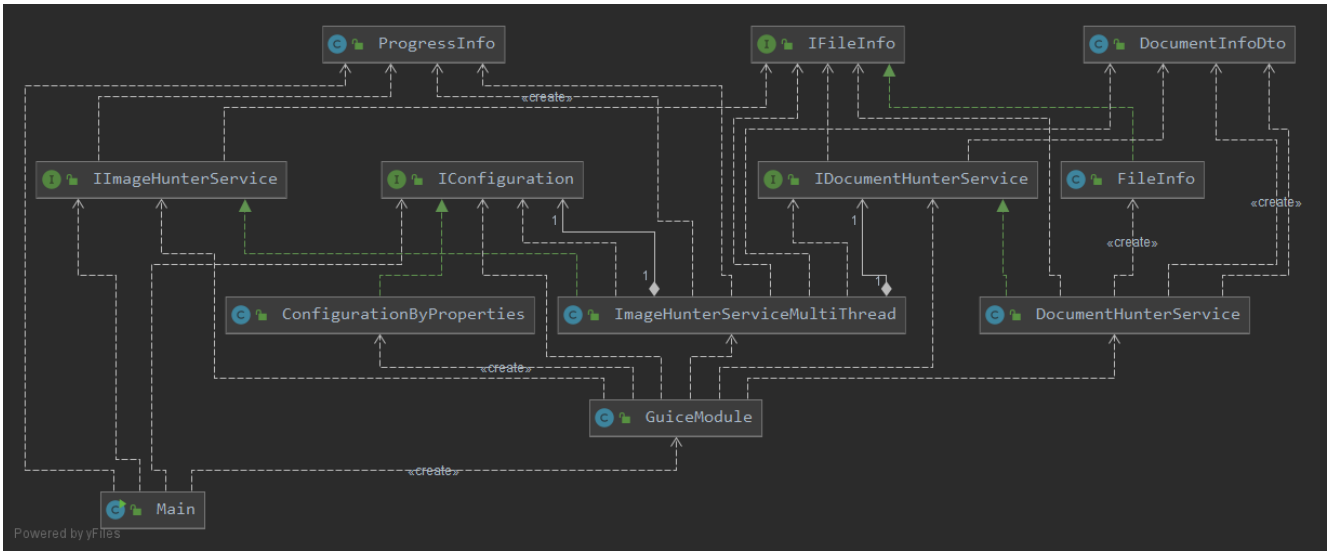

Рис. 9. UML-диаграмма классов

Выводы

В ходе выполнения данной работы была успешно разработана программа многопоточного паука для скачивания всех картинок с заданного сайта. При этом программная реализация удовлетворяет всем поставленным требованиям, а именно:

- Возможность выбора начального URL-адреса
- Возможность поиска картинок на дочерних страницах сайта
- Возможность выбора папки для сохранения картинок
- Возможность указания минимального размера загружаемой картинки
- Возможность параллельной работы прогаммы

Не смотря на реализацию всех поставленных требований, разработанное программное обеспечение уступает существующим аналогам. В частности, программа HTTrack позволяет полностью скачать сайт (при этом можно указать разные папки для сохранения изображений и веб-страниц), а программа Webripper изначально разрабатывалось именно для загрузки различных файлов с сайтов (в том числе и картинок). При этом хочется отметить, что обе эти программы бесплатные, что значительно расширяет возможности по их использованию.

Библиография

1) http://visualwebripper.com
2) https://www.httrack.com
3) http://www.tenmax.com/teleport/home.htm
4) https://www.ixbt.com/soft/offline-browsers-2.shtml
5) https://www.ixbt.com/soft/offline-browsers-4.shtml
6) https://www.ixbt.com/soft/offline-browsers-6.shtml