МУЛЬТИПОТОКОВЫЙ ЗАГРУЗЧИК

Рябченко Д.В., Аноприенко А.Я.
Донецкий национальный технический университет

Почти любую информацию можно найти в Internet. Зачастую в Internet информация представляет собой совокупность страниц гипертекста, содержащих текстовые данные и ссылки на другие страницы, и файлы изображений, звуков, видео. Для просмотра информации пользователь загружает индексную страницу, содержащую ссылки на другие страницы. По клику на ссылку загружается для отображения другая страница. Следовательно, если размер интересующих нас страниц составляет, к примеру, 45 кб, то каждый раз, при переходе от одной страницы к другой броузер будет загружать страницу заново, увеличивая объем траффика.

Данная проблема может быть решена только при наличии кэширующего прокси-сервера. Если нет прокси-сервера, или загружаемые страницы генерируются динамически, то при каждом просмотре страницы пользователь будет терять драгоценный объем трафика.

В таком случае целесообразно создавать копию интересующих страниц на локальном диске. Для этого удобно использовать программу Teleport Pro. Данная программа позволяет скопировать на локальный диск указанную страницу с ресурсами, а так-же страницы, на которые ссылается данная страница. Таким образом Teleport Pro позволяет сохранять на локальном диске совокупность связанных гипертекстовых документов со всеми встроенными ресурсами. Полученая копия будет доступна для просмотра при отсутствии доступа к Internet.

К сожалению, данный метод широко известен и порталу, предоставляющему интересующие данные. Размер прибыли, получаемой провайдером зависит от объма загруженных пользователем данных и длительности соединения. Поэтому некоторые порталы устанавливают на своих серверах специальные программы, предотвращающие загрузку страниц посредством Teleport Pro и его аналогами, которые загружают страницы несколькими потоками параллельно. Если на сервер, с одного и того же IP-адреса одновременно поступают несколько заявок на разные страницы, то вместа ожидаемых страниц сервер страницы с сообщением типа “Copying denied”.

Этого можно было бы избежать, если уменьшить число потоков до одного, но Teleport Pro использует десять потоков для загрузки файлов и пользователь не может изменить этот параметр. Более того, на данный момент не существуют портативного ПО (динамчески-подключаемых библиотек), которое может быть использованно разработчиками в своих программах для загрузки связных документов.

На основе анализа проблеммы повторной загрузки была поставлена задача разработки динамически-подключаемой библиотеки средствами IDE Microsoft Visual C++ 7.1 для загрузки связных страниц гипертекста.

Библиотека позволяет пользователю задавать количество потоков для загрузки и глубину обхода страниц. Страницы и встроеные ресурсы загружаются без повторений и только в том случае, если они расположены в пределах одного сайта.

Механизм взаимодействия потоков и распределения задач обеспечивает равномерное оптимальное использование потоков для обеспечения максимальной скорости загрузки.

Логика взаимодействия потоков и библиотеки построено таким образом, что поставленная задача равномерно распределяется между потоками,

исключая их простаивание, что позволяет выполнить задачу за минимальное время (рисунок 1).

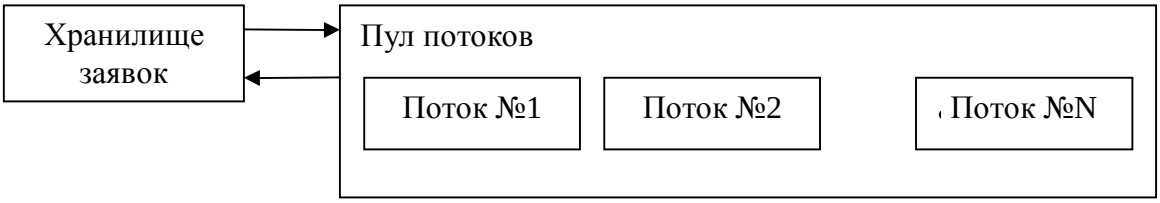

Рис. 1. Схема взаимодействия модулей

Хранилище представляет собой стек заявок. Методы доступа к данным синхронизированы при помощи очереди сообщений, что исключает одновременный доступ к данным хранилища. Пул представляет собой вектор указатель на обьекты класса потока-загрузчика. Сначала создаются хранилище и пул, затем в хранилище помещается первая заявка, а в пул добавляется требуемое количество потоков. Каждый поток запускается сразу после создания и пытается получить заявку из хранилища. Метод извлечения заявок из хранилища построен таким образом, что если хранилище пусто, то поток приостанавливается, пока не появится заявка. Все заявки, за исключением первой, обрабатываются параллельно.

Поток, получивший заявку, загружает гипертекстовый документ по адресу заявки, сохраняет его на жестком диске, и, если позволяет оставшаяся глубина, сканирует его на предмет ссылок на страницы гипертекста и изображений, добавляя их в хранилище. Парсер HTML представляет собой одноленточную машину Тюринга, распознающую ссылки на изображения, звуки, видео и другие HTML-страницы. Загрузка считается завершенной когда в хранилище нет ни одной заявки и все потоки ожидают в очереди хранилища.

Поскольку для загрузки URL-файла используется класс асинхронного моникера, поставляемого библиотекой MFC, загрузка осуществляется системными средствами, с учетом установленных параметров соединения (адрес прокси-сервера, параметры кеширования и т.д.).

В библиотеке реализована возвожность использования callback-интерфейса, для нотификации приложения о таких событиях, как начало и конец загрузки файла, получение ссылки из хранилища, добавление ссылки в хранилище.

Реализованная библиотека не требует инсталляции, дополнительных библиотек и приложений. Размер dll-файла составляет 52 кб, тогда, как, размер установленного пакета программы Teleport – порядка 1,5 мб. Нужно заметить, что данная программа уступает существующим аналогам в гибкости настройки. Библиотека разработана в среде MS Visual C++ 7.1. Откомпилированная программа представляет собой dll-файл, использующий системные библиотеки, поставляемой вместе с операционными системами семейства Microsoft Windows.

Литература

[1] А. Пол. *Объектно-ориентированное программирование на C++,2-е изд.*/Пер. с англ. – СПб.; М.: «Невский диалект» - «Издательство БИНОМ», 1999 г. – 462 с., ил.

[2] Круглински Д., Уингоу С., Шеферд Дж. *Программирование на Microsoft Visual C++ 6.0 для профессионалов, 5-е изд.*/Пер. с англ. – СПб: Питер; М.: Издательско-торговый дом “Русская Редакция”, 2004. – 861 с.: ил.