Web Crawler

Crawler User Manual

David Tee

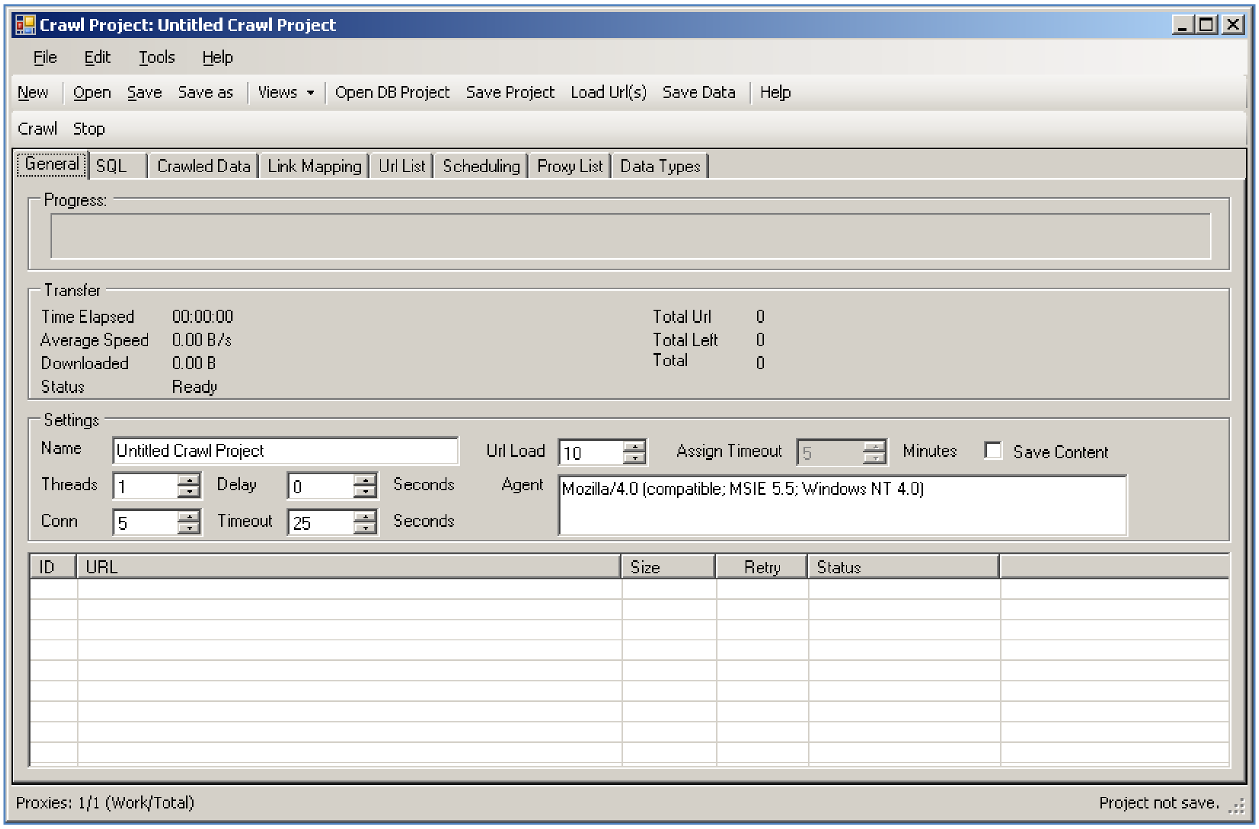
06

Contents

Overview

The crawler is program that allows the user to crawl any (non JavaScript) based websites.

Main Window



Time Elapsed 00:00:00

Average Speed 0.00 B/s

Downloaded 0.00 B

Status Ready

Total Url 0

Total Left 0

Total 0

Name Untitled Crawl Project

Url Load 10

Assign Timeout 5 Minutes

□ Save Content

Threads 1

Delay 0 Seconds

Agent Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 4.0)

Conn 5

Timeout 25 Seconds

Proxies: 1/1 (Work/Total)

Project not save.

Figure 1: Main Window

This Figure above is the main window of what the crawler looks like. In this screen, user can change tabs to configure a particular crawl project.

Menu

The main includes the follow the items

Menu	Description
File	Contains New, Open, Save, Save As Exit buttons
New	Create a new Crawl Project
Open	Open an existing Crawl project
Save	Save the crawl project to file (or to data base if it a database project)
Save As.	Save the crawl project to different file
Edit	Contains other Menu Items
Project Information	(Not yet implemented – If Implemented would be the same as the General Tab)
Connection	Change the data base connection for the Scraper to save the crawled data

	or crawl projects.
Tools	Contains Other Menu
Customize	(Not yet implemented – If implemented, would allow the user the change the default values of the fields, such as Agent, Connection, Threads, etc.
Options	Not yet implemented
Help	Contains other Menu Items
Online	(Not yet Implemented) Link to online manual for Web Crawler
Forums	(Not yet Implemented) Link to online forum dedicated to Web Crawler
Regular Expression	(Not yet Implemented) Link to regular expression reference on the website
About	About Window – Shows version number

Tool bar

The tool bar allows the user easier access to the menu item and also some new features. There are 3 different tool bars:

Main Tool bar

Button	Description
New	Create a new Crawl Project
Open	Open an existing crawl project
Save	Save the crawl project to file (or to data base if it a database project)
Save As.	Save the project to different file
View	Change the Tab group view
Open DB Project	Open a database project
Save Project	Save project to database
Load URL(s)	Refresh the URLs
Save Data	Save the project and the data crawled to database.
Help	(Not yet Implemented) Link to online manual for Web Crawler

Crawl Menu

Button	Description
Crawl	Start the crawling process.
Stop	Stop the crawling process.

[Scrape Project] Menu

Button	Description
New	Add a new [Scrape] Project
Wizard	Create a new project using wizard (wizard will generate the tag library – which does not work for most websites.)
Edit	Edit the [Scrape] Project information. (Currently project name, and the sample content for the proxy to search)
Remove	Remove the [Scrape] Project from list. (You may not remove the last [Scrape] Project; a crawl project requires at least one [scrape] project.

Configuration Tabs

The configuration tabs allow the user to focus on the specific configuration of the crawler.

General Tab

In this tab, user can set the general crawl project properties and can also view the progress of the crawl project. The progress is live and the screen will refresh as the crawl progresses. The screen is divided into a three portions: Progress, Transfer and Settings.

Progress

View the overall progress of the crawl. This progress is an estimate and may not be accurate as more and more URLs are discovered during the crawl session.

Transfer

View the data transfer details.

Settings

Property	Description
Name	The name of the crawl project. This is to distinguish a craw project from another. (This is necessary if the crawl projects are saved to database.)
Threads	Total downloads threads allowed. (This allows the crawler to download web pages concurrently to increase the performance)
Conn (Connections per IP)	This is a feature that's not yet integrated. This feature allows the crawler to limit connection per IP address (per proxy listed).
Delay (seconds)	Download delay in between each download. This allows the crawler to not hammer websites. This is a must have feature as some websites limit number of pages downloaded per second.
Timeout (seconds)	Total time given for the downloader to download the website.
URL Load	This is a feature is fully integrated and allows the crawler to check out a set amount of URL from database. This way, the crawler would only crawl a set amount of URLs and update it back to the database. This allows efficient distributed crawling.
Save context (check box)	This enables the crawler to save the content of the URL downloaded to database, into the URL data table.
Assign Timeout (minutes)	Assign Timeout allows the database to give a particular crawler client a limited time to crawl the given URLs. If the URLs are not crawled and updated back to data within time, the URLs are reassigned to another crawler that's requesting URLs.
Agent	The downloader sends the agent string as part of the html request string to the web server when requesting a page to download. This is how the crawler will identify itself to the web server.

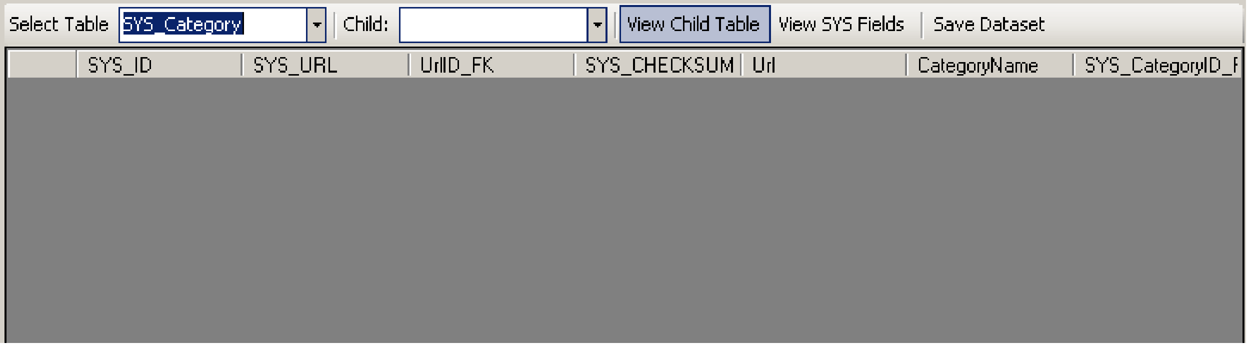
Progress List Box

SQL

In is tab, you can view the SQL code generated to install the database schema. The group box contains one text box which contains SQL commands for Microsoft SQL Server which generates the data tables necessary for saving crawl projects to database.

Crawled Data

In this tab, you can view the data crawled during a crawl session. This tab contains a view that allows the user to select tables and the data scraped for the table. (The functionality is the as "Result View Window" in Scraper).



Select Table: SYS_Category

	SYS_ID	SYS_URL	UrlID_FK	SYS_CHECKSUM	Url	CategoryName	SYS_CategoryID_F

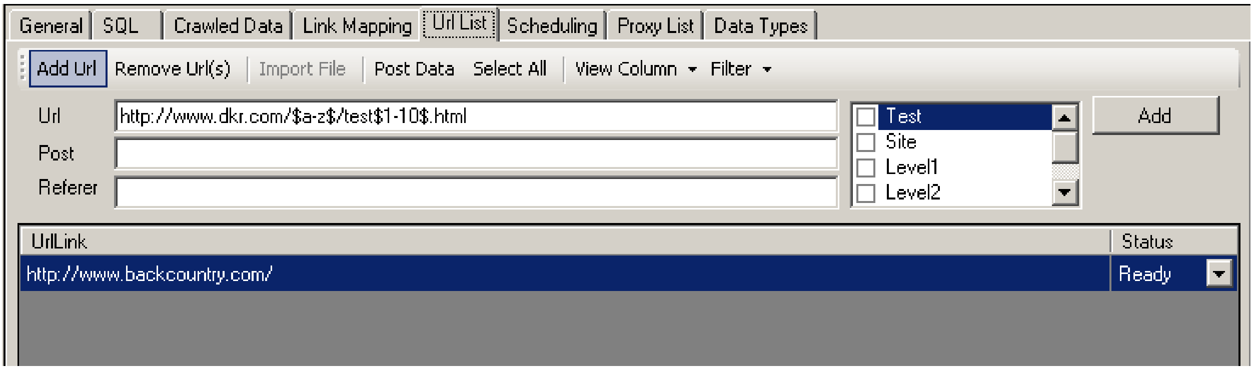
Figure 2: Crawled Data

Link Mapping

Link Mapping is used to configure how the crawler should navigate the website for information. This is where the user can specify which extracted data should be added to the URL list to be crawled next. The crawler also uses this tree to figure out the data relationships.

URL List

In this screen, user can view the URLs that are automatically by crawling process or manually added by user. User can manually modify the status of the URLs listed.



Url: http://www.dkr.com/$a-z$/test$1-10$.html

Post:

Referer:

□ Test
□ Site
□ Level1
□ Level2

UrlLink	Status
http://www.backcountry.com/	Ready

Figure 3: URL List

Scheduling

In this screen, user can define the schedule of when the crawling should begin. This is will be used when the when the crawler is separated into two parts, Crawl Edit, and Crawl server. The crawl server will use this scheduling configuration to run the crawl project on the given schedule. The schedule is implemented using UNIX's Cron Job design.

Proxy List

The crawler in the future would support the use of anonymous public proxy to crawl a website. User will be able to add a proxy and the crawler would figure out which proxy to using when downloading a URL. The use of proxy (if implemented) would only work on websites that do not keep track of user state using cookie or IP.

Data Types

Data type allows the crawler to systematically format the extracted data by allowing the user to write codes that formats the extracted data.

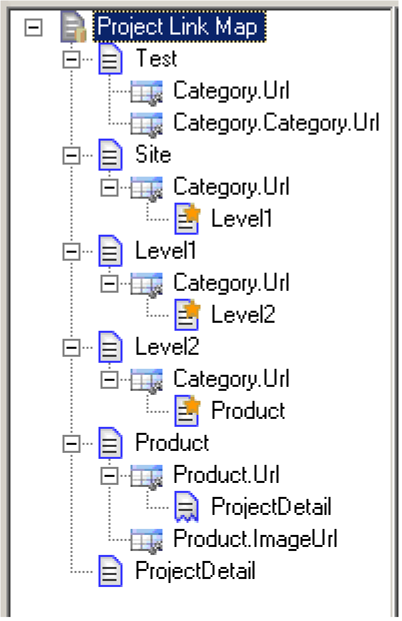
[Scrape] Project Tab

This tab can be accessed by changing the tab group view (located in the Main Tool Bar – Views button). In this tab, user can edit the configuration for data extraction from a particular webpage. For more information on the Scrape Project, see the Scraper Manual.

Sample Project: BackCountry.com

Back country.com Crawl project crawls http://www.backountry.com for their category and product information. This project demonstrates how various scraper projects are tied together using Link mapping. Link mapping allows the crawler to figure out which [scraper] project(s) should be used to scrape the URLs extracted from the content downloaded.

Here's is the tree of the Link Mapping in the backcountry.com crawl project.



```
Project Link Map
    Test
        Category.Url
        Category.Category.Url
    Site
        Category.Url
            Level1
    Level1
        Category.Url
            Level2
    Level2
        Category.Url
            Product
    Product
        Product.Url
            ProjectDetail
        Product.ImageUrl
    ProjectDetail
```

Figure 4: Link Map

Link Map

The link Map in Figure 2 lists all the scrape projects and the URL tag nodes that it contains. You can add or remove scraper project(s) to the URL nodes listed in the tree.

URL List

Backcountry.com crawl project starts from one URL that serves as the entry point for navigating and crawling the site for its context. If you view the URL list, you will see http://www.backcountry.com URL listed there with the status Ready, meaning this URL is ready be downloaded and scraped.

Sample Crawl Project: Lyrics007.com

Lyrcis007.com Crawl Project crawls http://www.lyrics007.com for lyric information. It will extract the information and save the information to Artist and Lyric data table. To view this project, open “lyrics007.com.sproj”.