Web Extractor

Version 2.0

Web spidering is a typical and common task for which each search engine has a solution. There are even free web spiders that make their source code available on the Internet. Thus, many would ask, "Why build another web spider?" The answer is that it is easy to build a web spider that goes to a web site and pulls down the text from the pages that is readily available, but it is something entirely different to create a web spider that extracts the valuable data from a website which is needed and only that data.

Most web spiders were built with the mind-set that it would never be possible to have a web spider pull out precise information from the almost endless variety of web sites on the Internet and most likely the hope that no one would ever look too closely at the large amount of garbage information retrieved and the lack of information that was intended to be retrieved. 30 Digits development team has seen too many projects where the off the shelf web spiders just did not meet the customer's needs. Thus, the 30 Digits Web Extractor was created.

The Web Extractor approaches spidering as an integral part of the data retrieval process. To this end, the spider has been designed to deal with various layouts of pages and structure of websites and to treat them individually. It can even parse a single page in different ways to extract multiple sections of valuable information. One example of this would be a page with information on a hotel and the associated reviews. The spider can retrieve all of the hotel data, and store that separately from each of the reviews which can also be split in to multiple documents even if they are all on one web page. The Web Extractor is essentially designed to cut out the particular areas of interest from a web site and deliver them in the appropriate text form whether that be for a search engine or a content platform that will analyze the data or reformat it for integrated display into another platform.

The 30 Digits development team has also built in a number of other features to make the spider stand out above the rest. Features like data qualifications, live spidering analysis, and regular expression based parsing are just the beginning.

Supported Web Sites

The Web Extractor uses the HTTP protocol. Thus, any site using that (regardless of port) can be spidered. This includes Internet web sites, company intranets, wikis, forums, and a number of other types of sites. What is unique about the 30 Digits Web Extractor is that it can be configured differently for each of these types to retrieve the relevant content. It can even handle sites designed with dynamic content.

Speed

The Web Extractor is extremely fast for two main reasons. First, because the configuration can be setup to only spider the parts of a website that are relevant, it avoids wasted spidering and download time. Secondly, it is multithreaded meaning that it can perform multiple tasks simultaneously. The number of threads can also be set higher to increase speed or lower to reduce the impact on the web server.

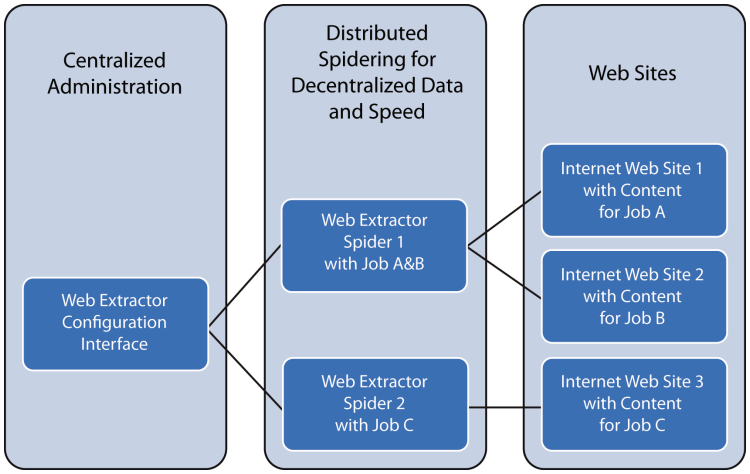
Localization

Web sites are getting more intelligent about recognizing a users location and settings to present them relevant content for their time, location, and language. The Web Extractor can be set to act as if it is from different users from different locals. Through this ability, it also can act anonymously.

On web sites, time is also represented on pages on multiple variations of formats and languages. The Extractor can take each of these in their native form and transform them into standard format that they can then be compared across sources.

Architecture

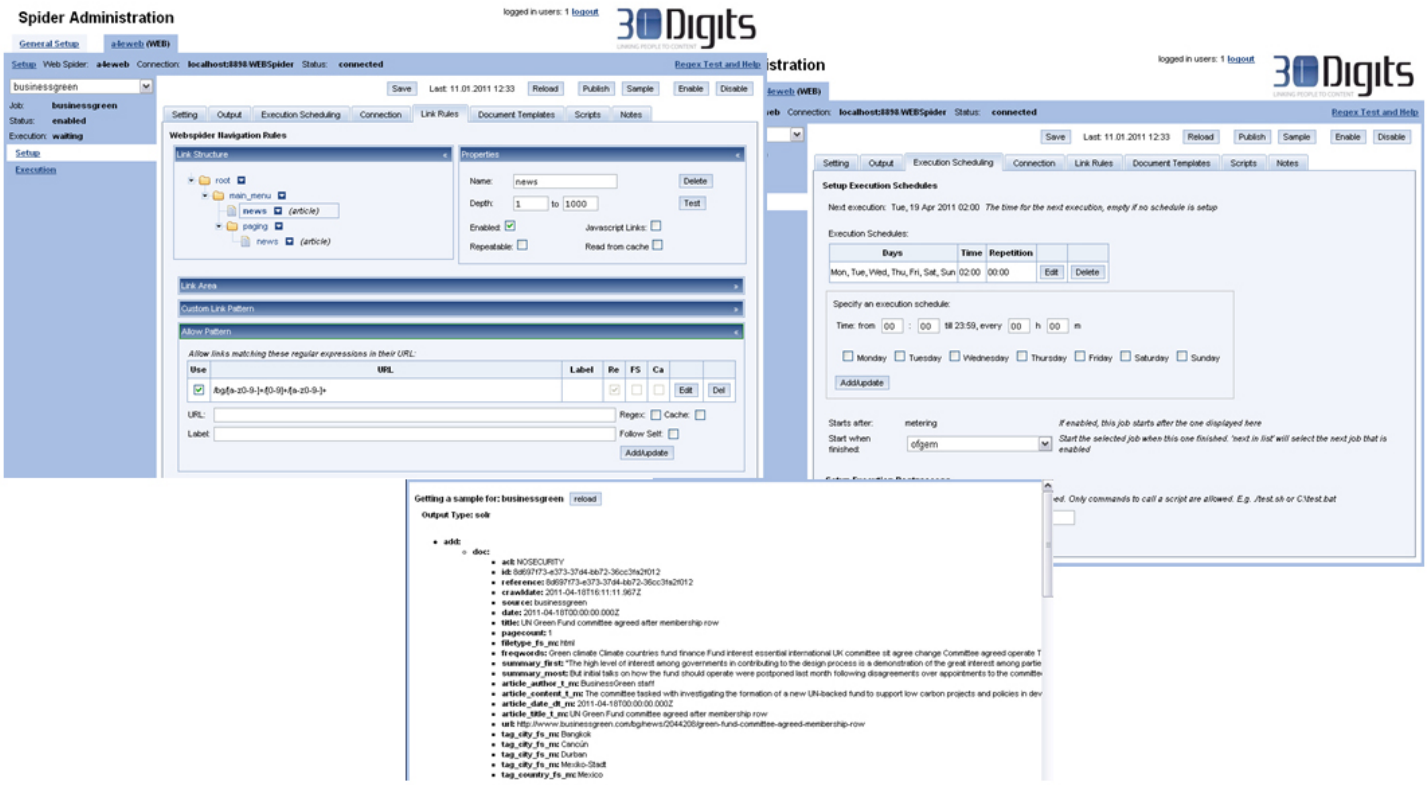
The architecture has been elegantly designed that it has a single point of administration while allowing multiple instances of the spider component to be run on several different machines. This allows scaling for large implementations even across multiple geographic locations while maintaining accuracy and ease of administration.



Setup and Administration

An often overlooked aspect of such a system is the effort involved in setting it up and administering it. The development team at 30 Digits have seen how complex and manual most of the solutions on the market are and have thus taken great effort to create a method for extracting content that is easy and intuitive while remaining flexible and powerful. Some of the main ways this has been achieved are by providing the following:

- Graphical User Interface (GUI) available via web browser for all of the configuration and administration options instead of textual configuration files which are often nearly impossible to decipher and lack documentation of how to setup and no central way to administrate them.
- The system uses AJAX extensively to load lists in the background to help fill out information.
- Simple field property modification like changing the title name of a field or even more complex operations like switching field variables from names to numbers for search engine optimization.
- Advanced features are available when an administrator wants to go deeper but not necessary for beginners or basic setup.
- Regular polling times for checking new data can be set very granularly and flexibly (i.e. every Monday, Wednesday, and Friday at 0, 6, 12, and 18 hours).
- When information is changed or deleted, the system can be set to either replace or delete the information in standard mode or keep time stamped versions of the document in the versioning mode.
- Regular Expressions can be used both in setting up the paths to follow and in defining what information to extract from a page.
- The Web Extractor can also login to web sites requiring authentication or requiring cookies.
- Sampling of spider output before jobs are run. This can save tremendous amounts of time as different configurations can be tried before committing jobs that could have long run times. It also allows the capturing of errors at the beginning instead of waiting to analyze a log file after hours of spidering.



File types and Meta-Data extraction

Web sites often have files for downloading which can be as important as the information in the web pages. The Web Extractor can also download these, and extract not only the text from the content but also the meta-data associate with these files.

The Web Extractor handles a large assortment of file types. Here is a list of some of the major ones: Microsoft Word, Excel, PowerPoint, PDF, HTML, TXT, RTF, MSG, XML, and ZIP (which are handled recursively to extract out the relevant data from each file within the ZIP). New file types are constantly being added to the import process to assure all kinds of files can be processed.

Contact

For more information or to schedule a demo, contact us at one of the following:

Tel: +49 89 45 23 89 66

Fax: +49 89 45 23 89 70

contact@30digits.com

30 Digits GmbH

Agnes-Pockels-Bogen 1

80992 München

Germany