Manual of the Web Spider for Fish4Knowledge Project

Bas Boom

March 14, 2011

Abstract

This is a short manual for the webspider developed for the Fish4Knowledge project. The purpose of this webspider is to collect data on the fishes in the Taiwan Fish Database (http://fishdb.sinica.edu.tw/). Information, which can be interested for the Fish4Knowledge project, is the links to other websites and URNs which uniquely identify the species. From a computer vision point, we need examples of the different fish species to be able to recognise these species when they appear in the camera footages. Other information like the habitats, habitats depth and size of the fish allows us to filter species that are not likely to appear in front of the camera.

1 Short Manual

The webspider is a python program written in eclipse. It can basically run once you have set the global save_path variable in the code. The save_path variable is the directory where the program saves all the output. The program will generate the following output: 3 xml files + directories containing all download images. The spider is able to save all the data that it find into an xml structure. This xml tree structure which is similar to the tree provided on the Taiwan Fish Database (http://fishdb.sinica.edu.tw/ajaxtree/tree_t_e.php). For all the items in this tree (class, order, family, genus, species, etc), we can add attributes. An example is that we provide a family image by the item family. On the species level, we added most of the attributes (habitat, habitats depth and size) and added subitems containing all the images that we found on both the Taiwan Fish Database and fishbase.org. The xml tree structure is saved in species_urn_catalogue_of_life.xml. It sometimes happens that the spider crashes, due to other libraries that the spider uses or url that are unavailable. For this reason, we save temporary results in species_urn_catalogue_of_life_temp2.xml. If a crash occurs, the species, which are in the temporary xml file, can be added in the program by renaming the species_urn_catalogue_of_life_temp2.xml to species_urn_catalogue_of_life_temp.xml. In this case, it will not search all the internet sites but will directly retrieve the information that it has obtain in the previous search from the xml file. The images are save in the same directory structure as is used by the websites from which we retrieved them. The directories are create automatically if they do not exist.

2 Data Structures

The program is basically a recursive loop to search through a tree structure. We added some special function to search in google, fishbase and catalogueoflife. We also have a special function to retrieve images and interped the information in the html pages In the spider program are two classes:

ProcessInfo: The class ProcessInfo allows us to keep count of the data that we have retrieved. It will sum up the number of fish, urns and images. We can also count for instance where the images are from, like Taiwan Fish Database or fishbase.org. Furthermore, we can define counter on the fishes that might be interesting to us, for instance by counting the fish that swim upto a depth of 10 meters.

CollectedData: The class CollectedData allows us to save all the information of a certain object in the tree structure (class, order, family, genus, species, etc). It can save and retrieves itself to an XML file. In this case, we assume that the Latin name is unique. We developed the function in this way that it will only save non-empty attributes. It also allows us to save multiple images at the species level. New attributes can be include by other partners and will be automatically save if they are not empty.

3 Extension

The spider can be easily extended to retrieve also other information, if you are interested in other information from the database of taiwan, checkout the function get_tag_info_taiwan, this allows you to give the fieldname of the information and it will save the value. If you want to link to other websites, checkout the function search_synonyms, which allows you to search for other names biologist also use for certain species. If there are any questions, just mail me (bboom@inf.ed.ac.uk). If you have developed an improvement or find a bug, I am also happy to hear it.