Web Page Scrapper Parser Plugin

Ammar Shadiq
ammar.shadiq@gmail.com
Universitas Pendidikan Indonesia
http://www.upi.edu

Organization/Project: Apache Software Foundation / Nutch

Assigned Mentor: Chris A. Mattmann

Abstract

Nutch use parse-html plugin to parse web pages, it process the contents of the web page by removing html tags and component like javascript and css and leaving the extracted text to be stored on the index. Nutch by default doesn't have the capability to select certain atomic element on an html page, like certain tags, certain content, some part of the page, etc.

A html page have a tree-like xml pattern with html tag as its branch and text as its node. This branch and node could be extracted using XPath. XPath allowing us to select a certain branch or node of an XML and therefore could be used to extract certain information and treat it differently based on its content and the user requirements. Furthermore a web domain like news website usually have a same html code structure for storing the information on its web pages. This same html code structure could be parsed using the same XPath query and retrieve the same content information element. All of the XPath query for selecting various content could be stored on a XPath Configuration File.

The purpose of nutch are for various web source, not all of the web page retrieved from those various source have the same html code structure, thus have to be threated differently using the correct XPath Configuration. The selection of the correct XPath configuration could be done automatically using regex by matching the url of the web page with valid url pattern for that xpath configuration.

This automatic mechanism allow the user of nutch to process various web page and get only certain information that user wants therefore making the index more accurate and its content more flexible.

Proposal Timeline

April 25 - May 7:

- To familiarize myself with Nutch functionality and architecture.
- Familiarize with the code and the community, the version control system, the documentation and test system used, and the new Nutch 2.0 version.

May 7 - May 23 (Before the official coding time):

- To do self coding with Nutch 2.0 to improve my further understanding of the new nutch version and plugin system. Including a mechanism for converting html page to XHTML page so it could be
- During this period I will remain in constant touch with my mentor and the Nutch community. I will remain active on IRC and Mailling lists to discuss the functionality and architecture of nutch. Thus with the help of my mentor I will become absolutely clear about my future goals,the final implementations that need to be done.

May 23 - June 6 (Official coding period starts):

- Define all requirements of the screen scrapper plugin.
- Define how the Configuration File would be interacting with nutch system.
- This will help in testing of the proper working of the entire basic code changes that we will later on incorporate in Nutch Source code (if any).

June 6 - July 4:

- Bringing the decided functionality for the plugin.
- Testing the overall working of each and every function and capability of the plugin.

July 4 - July 11:

- Testing the overall working of each and every function and capability of the plugin more thoroughly including test for various url seeds and condition.

JULY 11th MID TERM EVALUATION

July 11 - August 8:

- To be in constant touch with the Nutch's developers and to let them know about our progress.
- Making further changes in the code to improve the Functionality, Exception handling, Bug Removal.

A Buffer of two weeks has been kept for any unpredictable delay.

Additional Information:

The component for this idea have been tested on nutch 1.2 for selecting certain elements on various news website for the purpose of document clustering. This includes a Configuration Editor Application build using NetBeans 6.9 Application Framework. though its need a few debugging.

http://dl.dropbox.com/u/2642087/For_GSoC/for_GSoc.zip