GoogleSummerOfCode PrecisionDataExtractor

Title :	GSOC 2016 Proposal	
Issue (Formerly):		NUTCH-987 - A Plugin for extracting certain element of a web page on html page parsing
Student :	Ammar Shadiq - ammar.shadiq@gmail.com	
Mentors :		

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 treated 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.

Mentor Comments

- Chris Mattmann - very cool! Why not just integrate Scrapy with Nutch as a plugin? Something like parse-scrapy?
- Ammar Shadiq - This means that the parse-scrapy plugin would communicate with external Python program, is it ok? Then, my idea so far would be to call scrapy program each time nutch encounters a URL and feed back the Scrapy output to parse-scrapy to be processed as a nutch format.

Introduction

To be added

Timeline:

To be added

Reference:

*[1] https://issues.apache.org/jira/browse/NUTCH-978

Reports

To be added

Documentation

To be added

Source Code

To be added

Jira Issues

To be added