GoogleSummerOfCode SitemapCrawler weeklyreport

Week : 1 (25 May 2015 - 31 May 2015)

Title : Sitemap url injection is done.

Sitemap list injection is provided on this week working. The url path of sitemap files wanted to be injected can be defined from seed file.

In addition, Some preparatory work related to following works is being done.

Example:

If you have two sitemap files for "http://www.example.com/" , you can define them in the seed file as follow:

- *http://www.example.com/ sitemaps: sitemap1.xml sitemap2.xml*
- _http://www.example2.com/
  *_ http://www.example3.com/"

Then you can run InjecterJob. So the sitemaps urls are injected to the db. The urls injected are signed as sitemap.

Week : 2 (1 June 2015 - 7 June 2015)

Title : Sitemap detection is done.

Robot.txt is a file on the website. The file has sitemap url list. So, sitemap url list of a website can be accessed from this file.

Nutch Project reads robot.txt file while fetcher job is running. The file is checked from new code block of sitemap crawler. If it has any sitemap urls, these are written to stm(sitemap) column in the webpage table on the database.

The stm(sitemap)column is added to webpage schema for sitemap crawler. The urls in stm column from db will be parsed at the next time.

Week : 3 & 4 (8 June 2015 - 21 June 2015)

Title : Sitemap parser plugin is developed.

A plugin to parse sitemap file is developed. The plugin make use of crawler commons library. The sitemap file is parsed by the parse plugin. Inlinks from sitemap file is written to db. The inlinks will be parsed at the next time.

Week : 5 (22 June 2015 - 28 June 2015)

Title : DbUpdater is updated

DbUpdaterJob is updated for sitemap. Detected sitemaps are written to crawldb as a new line. Then the sitemaps will be crawled at the new crawl cycle.

Week : 6 & 7 (29 June 2015 - 12 July 2015)

Title : Sitemap parse plugin was abondoned.

Parser plugin was abandoned after consultation with mentors. The parse process was embedded instead of plugin. Sitemap parser will be activated according to the parameters given as "sitemap". Also midterm report is prepared. Up to this stage, sitemap life cycle has been developed according to the outline. Sitemap crawler runs simply. The process until now and from now on have evaluated.

Week : 8 (13 July 2015 - 19 July 2015)

Title : Sitemap file detection

Sitemap file detection is implemented. The detection is activated according to the parameters given at instant of fetch.

Week : 9 (20 July 2015 - 26 July 2015)

Title : frequency & priority

Create processSitemapParse function on ParseUtil. Parser process is updated for sitemap. Fetch interval time is updated acording to frequency value from sitemap. Also priority field is added to crawldb for priority value from sitemap.

Week : 10 & 11 (27 July 2015 - 9 August 2015)

Title : Review & code cleaning

Some improvements were made according to the review of my mentor. Code cleaning is done. Sitemap score logic isn't developed, because current nutch score logic is affected. It can be done according to the evaluation about it later.

Week : 12 (10 August 2015 - 17 August 2015)

Title : Testing

Some of problems have been fixed in the nutch test classes. Sitemap Tests were prepared. Documents of sitemap crawler were prepared.

Week : 13 (18 August 2015 - 21 August 2015)

Title : Final evaluation

The final document were prepared. Nutch wiki is updated.