SecondReport

Google Summer of Code 2014 Report 2

Project Name: NUTCH-841 Create a Wicket-based Web Application for Nutch 2.X

Report date: 11th July 2014

Student Name: Fjodor Vershinin

Mentor Name: Lewis John McGibbney (lewismc)

Development Codebase: https://bitbucket.org/feodorv/uinutch/

Project description

Main goal of this project is to create an Apache Wicket-based Web GUI for Apache Nutch 2.X.

Review of Previous Actions

- ~~If possible create a graphic of the REST API as it exists in his proposed patch for NUTCH-1769-API-refactoring this should only include the information included in his above commentary on the topic.~~
- ~~Provide links to the **HTML Prototype**, I have not seen any of this code and therefore cannot assert that progress has been made as described above.~~
- ~~Provide links/patches for the **application skeleton** as stated above... I have yet to see any code.~~

Objectives

Crawling cycle

At previous week (13.07-19.07) I worked on the most challenging task, namely I'd tried to implement crawling cycle in GUI part. The most problematic was tasks status controlling, but I'd solved this issue with simple polling. Other option is to post whole batch of jobs to Nutch Server, and shift all the responsibility to server's side. You can see my pullrequest on bitbucket. Also, I would propose minor changes in API and created issue with a little patch about generate component. NUTCH-1819 19.07-27.07 I created page, which allows create and run remote crawls. Main issue was concerning asynchronous execution and displaying progress. I'd implemented this by using spring's @Async annotations and spring's executor. Progress reporting is made by polling mechanism, which can be replaced by wicket-atmosphere in future. Then html5 websockets can be used instead of polling. Also, some refactorings has been done and fixed bug in test execution process. pull request

Seed upload

Very important step to run application on VM. Concerning seed upload, I'd proposed not to upload files, but add ability to create seed lists on UI side, which can be uploaded by API, and nutch server will create seed file. This option can make management of seeding much easier. Current implementation creates directory in /tmp, which contains file with seed urls, which works fine for now.

Data store

Now UI app should store too much info in plaintext properties file. I'd proposed to take embedded H2 java database, then data management wouldn't be an issue. Current implementation has H2 database and Ormlite as persistence provider. It was hard decision to take ORM into this application, but SQL written in java looks even worse. With a ORM and database I'd got rid of ugly properties file parsing. Also, it gives ability to build more complex solutions for user and roles management in future.

Contributions to Nutch community

NUTCH-1819 Crawling cycle Crawls management

Future Actions

- Change entire build structure to Ant + Ivy as per existing 2.x codebase
- Implement seed information upload using REST API
- Create embedded database for storing crawls, user credentials, and so on
- Write tests and some documentation/javadoc

Mentors Comments

Fjodor's code has come on well since last reporting. We have been working to get a VM established on Apache Infrastructure with limited success, however I am going to host Fjodor's work at http://any23-vm.apache.org as an itermediate step to achieving the goals and aim of this GSoC project.

We have not been communicating very much which (whilst may suit Fjodor) does not necessarily suit myself. This being said, I do however very much appreciate the direction in which he is taking this project and of course the initiative he is showing by keeping the project moving at a reasonable pace.

Our next step will be to get the VM established, we can then patch up a local copy of Nutch 2.x, keeping it in Sync with the 2.X branch. I have a vision that we will establish an HBase server on the VM as well which we will simply truncate the WebPage tabel for on an hourly basis in an effort to prevent the VM for maxing out storage space. I will get this set up and provisioned once Fjodor is ready for his code to be displayed on a public level.

Overall, another good reporting period, however we still have a lot of work to do, including engaging in a close review of the codebase, based on user feedback from within the Nutch community.

Signed: Lewis John McGibbney (lewismc)