Nutch - Parse Metatags

Summary: When crawling HTML pages, it might be necessary to retrieve information which is stored in HTML Meta tags. This tutorial shows how to install the plugin and configure Nutch to parse meta tags into separate fields in the Solr index. Note that Nutch pushes the information to Solr, so this tutorial also includes the changes required to Solr.

Plugin Information

This plugin has been developed as patch for Nutch 1.3. It parses specified meta tags and stores them in separate fields in the Solr Index.

Prerequisites

Solr and Nutch should already be set up. `%NUTCH_HOME%` is used as reference to your Nutch installation directory.

Plugin Installation

There are two possibilities to install this plugin: by adding the relevant jar files to an existing Nutch installation or by applying a patch to the Nutch code and building Nutch completely new. In most use cases, you only need to copy the relevant files instead of building Nutch.

Option 1: Adding the relevant files to existing Nutch

1. Use the included plugin "index-metatags.zip".
2. Extract the zip file.
3. Put the folder `index-metatags` into `%NUTCH_HOME%/plugins`.

Option 2: Applying the patch to the code and build Nutch

1. Download the patch file "NUTCH-809_metatags_1.3.patch" from jira: https://issues.apache.org/jira/browse/NUTCH-809
2. Download the Nutch source code from here.
3. Apply the patch to the code - there is a new plugin called `index-metatags` available.
4. Build the Nutch tar by running the Ivy/Ant goals `runtime` and `tar`.
5. Set up Nutch.

Plugin Configuration

1. In the file `conf/nutch-site.xml`, edit the property `plugin.includes` to contain the following plugin: `|index-metatags`, so it looks like for example:

   nutch-site.xml

```
<property>
  <name>plugin.includes</name>
  <value>protocol-http|urlfilter-regex|parse-(html|tika|js|zip)|index-
(basic|anchor|metatags)|query-(basic|site|url)|response-
(json|xml)|summary-basic|scoring-opic|urlnormalizer-
(pass|regex|basic)</value>
</property>
```

2. In the file `conf/nutch-site.xml`, specify which metatags should be indexed. Either specify specific metatags you want to index, or you can index all metatags. To index all, provide a '*' for the value of the property "metatags.names", otherwise provide the list of names separated by ';'. For example, to only index the metatag 'role', add the following configuration to `conf/nutch-site.xml`:

   nutch-site.xml

```
<!-- Used only if plugin parse-metatags is enabled. -->
<property>
  <name>metatags.names</name>
  <value>role</value>
  <description>For plugin parse-metatags: Indicate here the name of the
html meta tag that should be
         parsed. Use a semicolon separated list if you want multiple
tags, or use '*' to index all.
         Example: description;keywords;role
  </description>
</property>
```

3. In order to have the specified metatags indexed by Solr, edit your Solr `schema.xml` (located in `$SOLR_HOME$/conf`) and include new fields for each metatag you want to indexed. For example for the field 'role', add the following lines:

   schema.xml

```
...
<fields>
  ....
  <!-- fields for the metatags plugin -->
  <field name="role" type="String" stored="true" indexed="true"/>
  ...
</fields>
```

4. Restart Solr to load the new configuration.
5. Re-index your pages by running Nutch again - the metatag should be available in the Solr index. Check the index with Luke to see if it is available as separate field.