IndexMetatags

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. This article relates to the parse{{-metatags}} plugin, provided in jira: NUTCH-809

{{{#!wiki solid This plugin is not included in 2.x series (it will be included in 2.3). Please check NUTCH-1478 for patch, and also NUTCH-1827.
}}}

Plugin Information

This plugin parses specified meta tags and relies on the `index-metadata` plugin. It has been included in Nutch 1.5. With Nutch 1.7 all values of multi-valued metatags are added (see NUTCH-1467), with Nutch 1.9 the configuration is simplified (NUTCH-1561).

Plugin Configuration

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

```
<property>
<name>plugin.includes</name>
<value>protocol-http|urlfilter-regex|parse-(html|tika|metatags)|index-(basic|anchor|metadata)|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:

```
<!-- Used only if plugin parse-metatags is enabled. -->
<property>
<name>metatags.names</name>
<value>description,keywords</value>
<description> Names of the metatags to extract, separated by ','.
  Use '*' to extract all metatags. Prefixes the names with 'metatag.'
  in the parse-metadata. For instance to index description and keywords,
  you need to activate the plugin index-metadata and set the value of the
  parameter 'index.parse.md' to 'metatag.description,metatag.keywords'.
</description>
</property>
```

3. In the same file you need to configure the `index-metadata` plugin. The values are stored in the parse metadata so we need to specify the property `index.parse.md`:

```
<property>
  <name>index.parse.md</name>
  <value>metatag.description,metatag.keywords</value>
  <description>
  Comma-separated list of keys to be taken from the parse metadata to generate fields.
  Can be used e.g. for 'description' or 'keywords' provided that these values are generated
  by a parser (see parse-metatags plugin)
  </description>
</property>
```

CAUTION (1.x only): the names of the fields must be prefixed with 'metatag.'!
For **2.x** enter comma-separated metatags (without any prefix) which should be indexed to the property `index.metadata`:

```
<property>
  <name>index.metadata</name>
  <value>description,keywords</value>
  <description>
  Comma-separated list of keys to be taken from the metadata to generate fields.
  Can be used e.g. for 'description' or 'keywords' provided that these values are generated
  by a parser (see parse-metatags plugin), and property 'metatags.names'.
  </description>
</property>
```

4. You can test that the fields are generated correctly by using the bin/nutch indexchecker command
5. 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:

```
...
<fields>
....
<!-- fields for the metatags plugin -->
<field name="metatag.description" type="text" stored="true" indexed="true"/>
<field name="metatag.keywords" type="text" stored="true" indexed="true"/>
...
</fields>
```

   Note : you can use the file_ solrindex-mapping.xml_ to rename the fields e.g. *<field dest="description" source="metatag.description"/>*
6. Restart Solr to load the new configuration.
7. Re-index your pages by running Nutch again - the metatags should be available in the Solr index. Check the index with Luke (http://code.google.com/p/luke) to see if it is available as separate field.