IntranetDocumentSearch

This wiki is to hopefully get others an easier start into indexing and searching local intranet documents typically found in an enterprise file share. These would include Microsoft Office and PDF documents, text files and digital assets.

It draws upon various and sparse sources of information found online on the topic and will try to make their suggestions and changes related to later versions of the required software.

Pre-requisites and Assumptions

This tutorial assumes you are using the following software and configurations

- Apache Nutch 2.x HEAD
- Apache Solr 4.4.0
- Solr server will be addressed at http://localhost:8983/solr

This tutorial will most likely work with other versions of the above software out-of-the-box.

Configuration

Apache Solr

Apache Solr configuration will not be covered here in depth. However, there are some things that should be noted when setting up Solr to receive data from a Nutch crawler. There is a *schema.xml* file located in the Nutch *conf* directory which contains a Solr schema that Nutch utilises and expects to be present when posting data. A recommended course of action would be to use this schema in it's own core instance in Solr. In this example, it is assumed you have a core named *nutch* with this schema.

When configured correctly, there should be a core located at _http://localhost:8983/solr/nutch_. You can test this by accessing the administration page at _http://localhost:8983/solr/nutch/admin_ where you can also verify that the schema is being correctly loaded.

Apache Nutch

Put the Nutch package *apache-nutch-${version}.tar.gz or .zip into /opt and e_xtract Nutch using the command _tar ~~xvf apache-nutch~~${version}.tar.gz or .zip* . There is an example runtime setup under the directory */opt/nutch-${version}/runtime/local*. Change into this directory. All configuration files are referenced relative to this path.

File: conf/regex-urlfilter.txt

Make the following changes:

Comment out the lines preventing file:// URLs from being handled

```
# skip file: ftp: and mailto: urls
#-^(file|ftp|mailto):
```

Add the folowing lines to skip HTTP(S), FTP and MailTo URLs

```
# skip http: https: ftp: and mailto: urls
-^(http|ftp|mailto|https):
```

In this case the documents to be indexed are under */srv/samba/files**. Please use an appropriate regular expression for the paths you would like to index.

```
# accept anything else
+^file://srv/samba/files
#-.
```

Now create a directory called *urls*. This will hold text files containing lists of URLs for the crawler to process.

```
> mkdir urls
> touch urls/local-fs
```

File: urls/local-fs

Example content for urls/local-fs is. The priory configured URL filter regular expression should allow these paths to be accepted.

```
file://srv/samba/files/
```

It should be noted that directory paths which include spaces e.g. /media/my external disk/dir, should remain in thier native form. There is no need to escape characters.

File: conf/nutch-site.xml

This file configures the plugins that Nutch will use. It requires configuration to allow the file:// protocol plugin to be loaded. There are many configuration options that can be done here but the following should suffice to get a working local crawler going.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>http.agent.name</name>
    <value>My Nutch Spider</value>
  </property>
  <property>
    <name>plugin.includes</name>
    <value>protocol-file|urlfilter-regex|parse-(html|tika)|index-(basic|anchor)|scoring-opic|urlnormalizer-
(pass|regex|basic)</value>
    <description>Regular expression naming plugin directory names to
      include.  Any plugin not matching this expression is excluded.
      In any case you need at least include the nutch-extensionpoints plugin. By
      default Nutch includes crawling just HTML and plain text via HTTP,
      and basic indexing and search plugins. In order to use HTTPS please enable
      protocol-httpclient, but be aware of possible intermittent problems with the
      underlying commons-httpclient library.
   </description>
 </property>
</configuration>
```

Running the Crawler

The following command is used to start crawling.

```
> ./bin/crawl urls crawlId http://localhost:8983/solr/nutch 2
```

Command Breakdown

./bin/crawl - The Nutch crawl script which chains together individual crawl commands

urls - The urls directory we created earlier with our URL lists

crawlId - An identifier to associate this crawl with. Extremely useful for debugging crawls and tracking crawl progress in highly concurrent environments where many crawls may overlap.

http://localhost:8983/solr/nutch - Tells Nutch where the Solr server is located to upload index data after crawling has finished

2 - Is essentially a number of rounds (sometimes referred to as depth). In this instance (merely for demo) we set this to 2, otherwise crawling can take a very long time. This will test recusive retrieval without going too deep. You can amend this value you want to do a full crawl and index

Notes

MichaelAlleblas: I am very new to Solr and Nutch myself and therefore this is by no means a comprehensive or completely accurate guide. I just hope it can be a starting point for others to pool together their collective knowledge to help this capability of Nutch to be exploited more often and allow newcomers like myself to get things up and running as simply and fast as possible.