

Cómo instalar Nutch 1.3 y Solr 3.3 en Ubuntu 10.04 (Lucid)

Juan-Antonio Martínez-Comeche

- **Problema:** Nutch es un programa de código libre diseñado especialmente para realizar las tareas de robot de búsqueda, esto es, para rastrear la web recuperando las páginas que componen la red a través de la estructura de enlaces existente entre ellas. Nutch crea una base de datos con todos los enlaces encontrados, al tiempo que guarda una copia de todas las páginas localizadas y el resultado del análisis de su contenido, pues incorpora parsers para muchos formatos, no solamente HTML. Sin embargo, las tareas de búsqueda y recuperación de dicha información no están incorporadas, dependiendo del programa Lucene para su indexación. De igual forma, tampoco incluye una interfaz web que facilite la administración y uso del programa, debiéndose instalar el programa Tomcat para dichas tareas. Sin embargo, la versión Nutch 1.3 ayuda a resolver en buena medida estos inconvenientes. Elimina la dependencia de Tomcat, incluyendo el contenedor de servlets Jetty, lo que implica disponer de una consola de administración web. Además, esta versión Nutch 1.3 incorpora comandos para la integración y funcionamiento simultáneo del programa Solr. Solr es el programa de código libre perteneciente al proyecto Apache que se ocupa de las tareas de búsqueda y recuperación de información. Solr emplea la librería Lucene para las tareas esenciales de indexación y recuperación, pero a través de Tomcat o Jetty permite una configuración externa integral que posibilita el desarrollo de cualquier aplicación de búsqueda y recuperación de manera relativamente sencilla. Sin embargo, es necesario instalar por separado Nutch y Solr, configurarlos correctamente, y posteriormente integrarlos de manera que las páginas obtenidas con Nutch puedan ser indexadas y recuperadas mediante Solr en una consola web.
- **Prerrequisitos:** Es preciso tener previamente instalado Java SDK 1.5 ó superior. Para ello, se puede consultar en este mismo sitio el artículo “Cómo instalar Java SDK en Ubuntu 9.10 (Karmic Koala)”.
- **Solución:** Para elaborar el procedimiento que figura a continuación de instalación y configuración de Nutch 1.3 y de Solr 3.3, logrando su correcto funcionamiento de manera integrada, se han consultado las siguientes páginas:
 - <http://ubuntuforums.org/showthread.php?t=1532230>
 - <http://lucene.apache.org/solr/tutorial.html>
 - <http://wiki.apache.org/nutch/NutchTutorial>
 - <http://wiki.apache.org/nutch/RunningNutchAndSolr>
 - <http://www.xing.net.au/blog/crawl-and-search-using-nutch>
 - <http://nutch.apache.org/>

Proceso de Instalación:

- **Paso 1:** Ir a <http://www.apache.org/dyn/closer.cgi/nutch/> donde nos sugerirán un repositorio desde el que descargar Nutch. En nuestro caso, nos remite a <http://apache.rediris.es//nutch>. Allí pinchamos en la versión más adecuada, en nuestro caso, apache-nutch-1.3-bin.tar.gz, y guardamos el archivo en el Escritorio.
- **Paso 2:** Mover el archivo al lugar donde vaya a instalarse el programa, en nuestro caso, /usr/local. Para ello tecleamos en un terminal:

```
$ mv /home/juan/Escritorio/apache-nutch-1.3-bin.tar.gz /usr/local/
```

- **Paso 3:** Descomprimir el archivo en su lugar de instalación. Para ello tecleamos en un terminal:

```
$ cd /usr/local  
$ tar -zxvf apache-nutch-1.3-bin.tar.gz
```

Ello creará el directorio “nutch-1.3” en /usr/local/

- **Paso 4:** Verificar la correcta instalación de Nutch. Para ello, ir a /usr/local/nutch-1.3/runtime/local tecleando en un terminal:

```
$ cd /usr/local/nutch-1.3/runtime/local
```

Una vez allí, permitiremos la ejecución del programa al usuario que habitualmente empleará el programa. Considerando que es el mismo que ha seguido los pasos hasta aquí, basta teclear en un terminal:

```
$ chmod +x bin/nutch
```

De igual forma, debemos comprobar que la variable de entorno JAVA_HOME está correctamente configurada. Para ello, consultar el tutorial “Cómo instalar Java SDK en Ubuntu 9.10 (Karmic Koala)” en este mismo sitio, en especial los pasos 15-20.

Efectuadas estas comprobaciones, ejecutamos el programa Nutch tecleando:

```
$ bin/nutch
```

La instalación es correcta si se observan las siguientes líneas:

```
Usage: nutch [-core] COMMAND  
where COMMAND is one of:
```

```
.....
```

```
NOTE: this works only for jobs executed in 'local' mode
```

- **Paso 5:** Configurar inicialmente el programa para poder realizar rastreos de la web. En primer lugar, debemos incluir el nombre del agente -nosotros, nuestro equipo de investigación, la Facultad, etc.- que llevará a cabo el crawling. Esta información sobre quién lleva a cabo las tareas de rastreo debe añadirse a uno de los archivos de configuración del programa. Suponiendo que el agente es “Mi robot Nutch”, primeramente nos situamos en:

```
$ cd /usr/local/nutch-1.3/runtime/local/conf
```

A continuación, para conservar la versión inicial del archivo nutch-site.xml, lo renombramos:

```
$ mv nutch-site.xml nutch-site-old.xml
```

Luego hacemos una copia del archivo nutch-default.xml para emplearlo como archivo nutch-site.xml, que a su vez modificaremos posteriormente para configurar el programa adaptándolo a nuestras necesidades:

```
$ cp nutch-default.xml nutch-site.xml
```

Ahora editamos el archivo nutch-site.xml:

```
$ gedit nutch-site.xml
```

Y modificamos la propiedad “agent.name” de manera que quede:

```
<property>
  <name>http.agent.name</name>
  <value>Mi robot Nutch</value>
</property>
```

Guardamos los cambios efectuados en el archivo.

- **Paso 6:** Crear un archivo de texto plano con la/las url/urls inicial/iniciales que se emplearán a modo de semillas (seeds) para rastrear la web. Por ejemplo, para rastrear el sitio web de Nutch, debemos crear un archivo denominado “lista” con la siguiente línea:

<http://nutch.apache.org/>

Para ello, creamos un directorio denominado “urls” directamente bajo nutch-1.3/runtime/local:

```
$ mkdir /usr/local/nutch-1.3/runtime/local/urls
```

Creamos el archivo “lista”:

```
$ gedit /usr/local/nutch-1.3/runtime/local/urls/lista
```

Una vez hayamos tecleado la línea "<http://nutch.apache.org/>", cerramos el archivo (Ctrl-S).

- **Paso 7:** Estamos en condiciones de efectuar un primer rastreo del sitio web de Nutch. Para ello nos situamos en:

```
$ cd /usr/local/nutch-1.3/runtime/local  
$ bin/nutch crawl urls -dir crawl -depth 3 -topN 5
```

Se habrá creado el directorio `nutch-1.3/runtime/local/crawl` con los siguientes subdirectorios: `crawl`, `linkdb` y `segments`, donde se hallan los resultados del rastreo del sitio web de Nutch.

- **Paso 8:** Para comprobar cuántas urls hay en la base de datos y cuántas han sido rastreadas, ejecutamos el siguiente comando en un terminal:

```
$ bin/nutch readdb /usr/local/nutch-1.3/runtime/local/crawl/crawl -stats
```

Se mostrará en pantalla una información semejante a esta:

```
TOTAL urls: 413  
.....  
.....  
status 1 (db_unfetched): 402  
status 2 (db_fetched): 11  
CrawlDb statistics: done
```

- **Paso 9:** Para utilizar Solr es preciso disponer de una versión de Java SDK 1.5 ó superior, como se señaló en Prerrequisitos. Una vez confirmado este extremo, ir a <http://www.apache.org/dyn/closer.cgi/lucene/solr/> donde se nos sugiere utilizar el repertorio de rediris para descargar el programa. Pinchamos sucesivamente en <http://apache.rediris.es//lucene/solr/>, en la versión más reciente, en nuestro caso la 3.3.0, y finalmente en `apache-solr-3.3.0.zip`, descargando el archivo en el Escritorio.

- **Paso 10:** Mover el archivo al lugar donde vaya a instalarse, en nuestro caso, `/usr/local`. Para ello tecleamos en un terminal:

```
$ mv /home/juan/Escritorio/apache-solr-3.3.0.zip /usr/local/
```

- **Paso 11:** Descomprimir el archivo en el lugar donde se vaya a instalar Solr. Para ello, tecleamos en un terminal:

```
$ cd /usr/local  
$ unzip apache-solr-3.3.0.zip
```

Se creará el directorio `/usr/local/apache-solr-3.3.0`

- **Paso 12:** Solr puede funcionar con cualquier contenedor de servlets, como Tomcat, pero es suficiente con emplear Jetty, cuya instalación es muy simple. Para ello teclear en un terminal:

```
$ cd /usr/local/apache-solr-3.3.0/example
$ java -jar start.jar
```

Observaremos un mensaje parecido a este en pantalla:

```
08-jul-2011 19:26:47 org.apache.solr.core.SolrCore registerSearcher
INFO: [ ] Registered new searcher Searcher@18efaea main
.....
.....
2011-07-08 19:26:47.488:INFO:Started SocketConnector@0.0.0.0:8983
```

No debemos dar a ninguna tecla, aunque el cursor esté parpadeando. De hecho, con este último comando hemos iniciado Jetty en el puerto 8983, y está en funcionamiento. Nos limitamos, pues, a minimizar el terminal, de manera que siga funcionando Jetty, y abrimos el navegador introduciendo la dirección:

```
http://localhost:8983/solr/admin/
```

Si el proceso se ha efectuado correctamente, observaremos la pantalla inicial de la consola de administración del programa Solr (que se ha iniciado también junto con Jetty). De igual forma, si introducimos la dirección:

```
http://localhost:8983/solr/admin/stats.jsp
```

tendremos una pantalla con toda la información acerca del programa Solr y de la colección cargada (la primera vez, lógicamente, indica numDocs:0, esto es, que no hay ningún documento cargado en el sistema). Para cerrar la pantalla y consiguientemente el programa Solr, basta teclear Ctrl-C en el terminal que tenemos minimizado.

- **Paso 13:** Una vez que tanto Nutch como Solr se han instalado y configurado correctamente, debemos integrarlos de manera que las urls obtenidas con Nutch puedan ser recuperadas mediante Solr. Para ello, tecleamos en un terminal:

```
$ cp /usr/local/nutch-1.3/runtime/local/conf/schema.xml /usr/local/apache-solr-3.3.0/example/solr/conf/
```

de manera que se reemplace el archivo schema.xml por defecto de Solr con el de Nutch.

- **Paso 14:** Reiniciar Solr (junto con la consola web facilitada por Jetty) con el comando empleado en el Paso 12:

```
$ cd /usr/local/apache-solr-3.3.0/example
```

```
$ java -jar start.jar
```

Minimizamos el terminal para seguir teniendo acceso a la consola web.

- **Paso 15:** Abrir otro terminal, dejando el anterior minimizado, e introducir el comando de Nutch que efectúa la indexación de las urls rastreadas en el Paso 7 anterior. Para ello, teclear en el nuevo terminal:

```
$ cd /usr/local/nutch-1.3/runtime/local  
$ bin/nutch solrindex http://127.0.0.1:8983/solr/ crawl/crawlddb crawl/linkdb  
crawl/segments/*
```

Este comando consigue que Solr indexe todos los datos del rastreo efectuado en el Paso 7. En pantalla observaremos un mensaje semejante a este:

```
SolrIndexer: starting at 2011-07-08 20:13:24  
SolrIndexer: finished at 2011-07-08 20:13:26 elapsed: 00:00:02
```

- **Paso 16:** Si el proceso se ha efectuado correctamente, podemos empezar a realizar búsquedas sobre esas páginas. Para ello, abrimos el navegador e introducimos la dirección:

<http://localhost:8983/solr/admin/>

En Query String podemos introducir, por ejemplo, la búsqueda del término “nutch”:

Query String: +nutch

Hacemos clic en el botón “Search” y obtendremos un fichero en XML con los 8 documentos (páginas web rastreadas) que satisfacen esa consulta.