Universidad Complutense de Madrid

Facultad de Ciencias de la Documentación

Instalación y primera preparación de Nutch1.0 en Ubuntu 9.10

Manuel Blázquez Ochando

Madrid, 2009

Objetivo:

El presente manual trata de ayudar en la instalación de programa **Nutch**. Se trata de un webcrawler de código abierto, de gran relevancia por ser desarrollado por el Apache Software Foundation y utilizado ampliamente en investigaciones de la web. A la par que su reputación, se encuentra su manejo, que puede resultar complejo si no se dominan sus comandos esenciales. Con el objetivo de facilitar su uso e instalación se propone la presente guía de iniciación.

Guía:

Paso1: Instalar Java JDK 1.6 o SDK con synaptic.
Nutch es un programa desarrollado en el lenguaje de programación JAVA. Este requerimiento implica la instalación del paquete JAVA JDK 1.6 o SDK preferiblemente. El método de instalación más sencillo es acudiendo a la herramienta de instalación de paquetes de Ubuntu, también denominada SYNAPTIC. Se encuentra en la sección de herramientas del sistema y permite la búsqueda, selección y marcado para instalación de los paquetes, módulos y extensiones del sistema operativo Ubuntu. Entre ellas se encuentra el paquete JAVA señalado anteriormente. La peculiaridad de dicha versión de Java es su consideración para el profesional y desarrollador de aplicaciones, lo que garantiza que cualquier programa que utilice librerías Java, pueda encontrarlas con mayor seguridad. Sobre la instalación de Java SDK, también puede revisarse un excelente manual elaborado por el profesor Juan Antonio Martínez Comeche, que al igual que el presente, expone paso a paso su instalación particular.

Véase:

- *MARTÍNEZ COMECHE, J.A. 2009. Cómo instalar Java SDK en Ubuntu 9.10. Disponible en: http://www.comeche.es/documents/Software/como_instalar_java_sdk_ubuntu_910.pdf*

Paso2: Instalar Apache Tomcat 6.
El programa Tomcat es un servidor especializado en servicios JAVA. Esto significa que su empleo se centra en servir de plataforma web para las aplicaciones desarrolladas en alguna de las versiones de Java, dotándolas de un interfaz más amigable para el usuario. Aun no siendo condición necesaria para el funcionamiento de Nutch, sí resulta interesante su instalación de cara a terceras pruebas y consultas en un entorno visual. En este punto remito al manual elaborado por el profesor Juan Antonio Martínez Comeche, en el que expone los pasos a dar para su instalación.

Véase:

- *MARTÍNEZ COMECHE, J.A. 2009. Cómo instalar Tomcat en Ubuntu 9.10. Disponible en: http://www.comeche.es/documents/Software/como_instalar_tomcat_ubuntu.pdf*

Paso3: Descargar Nutch 1.0.tar.gz
Habiendo instalado JAVA SDK y Apache Tomcat, se está en disposición de instalar el programa Nutch. Para ello es necesario descargarlo de su sitio web oficial http://lucene.apache.org/nutch/ , accediendo a su sección "resources→downloads" apareciendo un portal de descargas hasta localizar el archivo en cuestión http://apache.rediris.es/lucene/nutch/nutch-1.0.tar.gz

Paso4: Descomprimir Nutch
Descomprímase el archivo *Nutch 1.0.tar.gz* descargado. Obtendrá una carpeta descomprimida denominada "nutch". Cópiela o muévala en algún directorio de fácil acceso para su explotación, como por ejemplo el correspondiente a un disco duro o dispositivo de almacenamiento "media/disk/nutch". Llegados a este punto el programa Nutch ya ha sido correctamente instalado. No obstante eso no significa que sea operativo, para ello se siguen los siguientes pasos.

Paso5: Crear una carpeta "urls" en /media/disk/nutch/urls
Para operar correctamente con el programa Nutch, se necesitan efectuar algunos cambios, como por ejemplo crear una carpeta "urls" dentro de la carpeta "nutch" de la siguiente forma "/media/disk/nutch/urls". Dicha carpeta contendrá la información correspondiente a las URLs de los sitios web que se desean explotar.

Paso6: Crear un archivo de sites.txt en /media/disk/nutch/urls/sites.txt
Dentro de la carpeta "urls" es necesario crear un archivo con la lista de sitios web de explotación. Esto también es considerado como la semilla que utilizará el crawler para efectuar la recopilación de información. El contenido del archivo sites.xml serán por tanto las URLs de cada sitio web, separadas por cambio de línea.

sites.txt

http://www.ucm.es/
http://www.ucm.es/BUCM/
http://eprints.ucm.es/
http://www.ucm.es/centros/webs/ebiblio/
….

Paso7: Editar el archivo crawl-urlfilter.txt en /media/disk/nutch/conf/crawl-urlfilter.txt
Para posibilitar la labor extendida del crawler, conviene determinar un filtro de búsqueda. Generalmente está configurado mediante una expresión regular, lo cual dificulta su edición. En esta guía se propone reemplazar la línea de código correspondiente a la búsqueda de dominios, por aquel que el usuario esté más interesado.

crawl-urfilter.txt

…

modificar el valor +^http://([a-z0-9]*\.)*MY.DOMAIN.NAME/ por +^http://([a-z0-9]*\.)*ucm.es/ por ejemplo.

….

Paso8: Editar el archivo regex-urlfilter.txt en /media/disk/nutch/conf/regex-urlfilter.txt
Para permitir la ejecución de expresiones regulares como la anteriormente expresada es necesario editar el archivo reges-urlfilter.txt. Si no se sigue este paso, Nutch no funcionará correctamente.

regex-urfilter.txt

…

modificar el valor # accept anything else +. por # accept anything else -.

….

Paso9: editar el archivo nutch-default.xml en /media/disk/nutch/conf/nutch-default.xml
Otro punto fundamental en la configuración de Nutch es la introducción de un nombre de agente. No debe olvidarse que un crawler o agente, actúa en la web y deja tras de sí una huella con su información básica, entre tales datos se encuentra su nombre y por ende, su descripción y administrador. Al igual que los pasos anteriores es fundamental la edición de la siguiente línea en el archivo.

nutch-default.xml

```
…

<property>
 <name>http.agent.name</name>
 <value>nombre de su agente o crawler</value>

….
```

Paso10: Activar JAVA.
Si se han seguido todos los pasos anteriores, Nutch está correctamente instalado y configurado. Para hacerlo funcionar, es necesario activar JAVA o comprobar que la variable de entorno de Java, la que emplea el sistema operativo Ubuntu para acceder a la versión de Java instalada es correcta. De lo contrario no funcionará. Una forma de hacerlo efectivo es tecleando la siguiente rutina en una de las sesiones abiertas en el terminal de Ubuntu.

Terminal de ubuntu

```
export JAVA_HOME='/usr/lib/jvm/java-6-sun-1.6.0.13'
```

La ruta marcada variará dependiendo de la versión de JAVA SDK instalada. No obstante, las carpetas señaladas en dicha ruta suelen ser invariables.

Paso11: Comprobar que se ha instalado Nutch
Una vez ejecutada la sentencia anterior suele comprobarse el correcto funcionamiento del programa Nutch. Para ello basta con introducir la siguiente ruta al archivo nutch.sh en el terminal de Ubuntu.

Terminal de ubuntu

```
cd /media/disk/nutch/bin/nutch.sh
```

Acto seguido, después de ejecutar dicha sentencia, aparecerá una lista de comandos de Nutch. Esto significará que se ha instalado y ejecutado correctamente quedando disponible para las primeras tareas de crawling.

Resultado de ejecución

```
Usage: nutch [-core] COMMAND
where COMMAND is one of:
 crawl             one-step crawler for intranets
 readdb            read / dump crawl db
 convdb            convert crawl db from pre-0.9 format
 mergedb           merge crawldb-s, with optional filtering
 readlinkdb        read / dump link db
 inject            inject new urls into the database
 generate          generate new segments to fetch from crawl db
 freegen           generate new segments to fetch from text files
 fetch             fetch a segment's pages
 parse             parse a segment's pages
 readseg           read / dump segment data
```

```
 mergesegs       merge several segments, with optional filtering and slicing
 updatedb        update crawl db from segments after fetching
 invertlinks     create a linkdb from parsed segments
 mergelinkdb     merge linkdb-s, with optional filtering
 index           run the indexer on parsed segments and linkdb
 solrindex       run the solr indexer on parsed segments and linkdb
 merge           merge several segment indexes
 dedup           remove duplicates from a set of segment indexes
 solrdedup       remove duplicates from solr
 plugin          load a plugin and run one of its classes main()
 server          run a search server
or
 CLASSNAME       run the class named CLASSNAME
Most commands print help when invoked w/o parameters.

Expert: -core option is for developers only. It avoids building the job jar,
      instead it simply includes classes compiled with ant compile-core.
      NOTE: this works only for jobs executed in 'local' mode
```

Paso12: Inicia el crawling de la carpeta urls

Una de las primeras pruebas de crawling consiste en utilizar el archivo sites.txt que se ha creado anteriormente dentro de la carpeta “urls”. Para ello se utiliza la siguiente sentencia.

Terminal de ubuntu

```
bin/nutch crawl urls -dir crawl -depth 3 -topN 50
```

Al ejecutar dicha sentencia, se observará una importante actividad en el terminal de Ubuntu, visualizándose en tiempo real el proceso de crawling y su progreso.

Paso13: Invertir links

Una forma de comprobar que Nutch ha efectuado correctamente el proceso de crawling es efectuando una consulta sobre los sitios web que ha recopilado. Esto es posible, mediante el proceso de inversión de links, que permite generar un fichero índice de todos los contenidos recopilados. Para ejecutar dicha función se utiliza la siguiente sentencia en el terminal.

Terminal de ubuntu

```
bin/nutch invertlinks crawl/linkdb -dir crawl/segments
```

Paso14: Efectuar una consulta

Para efectuar una consulta desde el terminal, se emplea una librería interna de Nutch, denominada NutchBean. La sintaxis de consulta para esta función sería la siguiente:

Terminal de ubuntu

```
bin/nutch org.apache.nutch.searcher.NutchBean Universidad Complutense
```

En amarillo está señalada la función de búsqueda NutchBean y en verde los términos de la consulta. El

resultado de ejecutar ésta sentencia es la obtención de un listado de páginas web que cumplen los criterios de la consulta planteada.

Paso15: Cargar Nutch en Apache Tomcat
Como se ha señalado anteriormente, Apache Tomcat es un elemento interesante para comprobar y explotar el resultado del crawling de Nutch desde un entorno web y con un interfaz gráfico. Para ello es necesario abrir apache Tomcat. Esto se consigue abriendo una sesión en el navegador web y tecleando la siguiente dirección. http://localhost:8080/ (URL por defecto para acceder siempre a un servidor Tomcat)

Una vez cargada la página de bienvenida, se accede al panel de gestión de Tomcat y se comprobará la existencia de un botón examinar. Esta opción sirve para cargar paquetes de aplicaciones en java y por ende poderlas ejecutar desde el navegador. El archivo que se pretende examinar y desplegar se encuentra en la ruta **/media/disk/nutch/bin/nutch.war**

Una vez se acceda a dicho archivo, se hace clic en el botón desplegar, que permitirá cargar el módulo de nutch para Apache Tomcat, que carga automáticamente un buscador con todos los datos indexados en el proceso de inversión de links.

También se puede efectuar el proceso de carga del módulo de Nutch en Apache Tomcat mediante los siguientes comandos:

Terminal de ubuntu

```
/media/disk/nutch$ rm -rf /media/disk/tomcat6/webapps/ROOT*
/media/disk/nutch$ cp nutch*.war /media/disk/tomcat6/webapps/ROOT.war
```

Paso16: Para obtener estadísticas básicas del directorio crawldb
Otras operaciones útiles para con Nutch es la obtención de estadísticas del proceso de crawling. Son útiles para conocer el progreso y funcionamiento del trabajo de recopilación e indexación. Para acceder a ellas se emplea la siguiente sentencia.

Terminal de ubuntu

```
bin/nutch readdb crawl/crawldb -stats
```

Y se obtienen resultados parecidos a los expuestos a continuación:

Terminal de ubuntu

```
CrawlDb statistics start: crawl/crawldb
Statistics for CrawlDb: crawl/crawldb
TOTAL urls:     600
retry 0: 595
retry 1: 5
min score:      0.0
avg score:      0.034855
max score:      1.711
status 1 (db_unfetched):        510
status 2 (db_fetched):  67
status 3 (db_gone):     11
status 4 (db_redir_temp):       1
```

```
status 5 (db_redir_perm):        11
CrawlDb statistics: done
```

Paso17: Para obtener estadísticas de segments
Por cada proceso de crawling ejecutado, se crean en Nutch, secciones o tal como lo denomina, “segmentos” que tienen sus correspondientes estadísticas, nombres y códigos distintivos. Su visualización se logra utilizando la siguiente sentencia:

Terminal de ubuntu

```
/media/disk/nutch$ bin/nutch readseg -list -dir crawl/segments/
```

Los resultados obtenidos son similares a los expuestos a continuación:

Terminal de ubuntu

```
NAME  - GENERATED - FETCHER START - FETCHER END – FETCHED - PARSED
20090626123457      2             2009-06-26T12:35:04   2009-06-26T12:35:05   2       2
20090626123513      44            2009-06-26T12:35:19   2009-06-26T12:36:07   50      35
20090626123617      50            2009-06-26T12:36:28   2009-06-26T12:36:56   56      25
```

Otros manuales y guías importantes:

- Nutch tutorial. 2009. Disponible en: http://wiki.apache.org/nutch/NutchTutorial
- Introduction to nutch1. 2006. Disponible en: http://today.java.net/pub/a/today/2006/01/10/introduction-to-nutch-1.html