Instalación y configuración del buscador ht://Dig

Luis Llorente Campo
Universidad de León, España

luisllorente@luisllorente.com

Este documento muestra cómo instalar y configurar el buscador ht://Dig (http://www.htdig.org). Todo el proceso está pensado para la distribución Debian GNU/Linux. Pretende ser una guía que muestre el proceso de una forma genérica, teniendo que ser adaptado el proceso para cada situación específica.

Introducción

Se ha elegido el buscador ht://Dig (http://www.htdig.org/) por ser un buscador orientado a pequeñas empresas y sitios web de tamaño medio. Permite indexar todos los documentos que estén situados en el servidor, incluídos archivos PDF, PS y DOC.

Está dotado de una interfaz web con posibilidad de realizar búsquedas avanzadas (lógica borrosa). En los resultados de las búsquedas se muestra un gráfico de la calidad de las respuestas obtenidas.

Debido a la orientación que se le va a dar (indexación de la documentación generada por el proyecto), es más que suficiente para nuestros propósitos, y de ahí el motivo de su elección.

Instalación

Necesitamos instalar el paquete htdig, y opcionalmente (aunque muy recomendable), la documentación, que se encuentra en el paquete htdig-doc.

Para ello, ejecutaremos como es habitual en Debian el comando:

```
# apt-get install htdig htdig-doc
```

Una vez finalizada la instalación, pasaremos a adaptarlo al idioma español y a configurar los distintos parámetros del buscador.

Configuración

El fichero de configuración es:

- /etc/htdig/htdig.conf

Por defecto htdig viene preparado para la indexación y búsqueda de documentos en inglés, por lo que tendremos que adaptarlo al español. Esta adaptación requiere bastantes cambios en varios ficheros, por lo que hemos desarrollado un pequeño paquete con los ficheros ya modificados, junto con un pequeño script de instalación de dichas modificaciones.

Adaptación al español

El paquete contiene un diccionario de palabras y sinónimos en español, que son necesarios para una correcta indexación de archivos en este idioma. También incluye la traducción realizada por nosotros de la interfaz web de las búsquedas y resultados.

Está situado en el servidor web, y desde la página del grupo GSO podremos acceder a su descarga. El fichero es cuestión se llama "*htdig-3.5.1-GSO.tar.gz*".

Una vez lo hayamos descargado, lo descomprimiremos en algún lugar temporal como puede ser el directorio /tmp:

```
# cd /tmp
# tar zxvf /ruta_al_fichero.tar.gz
```

Uno de los ficheros descomprimidos es un pequeño script de instalación. Lo ejecutaremos mediante la orden:

```
# ./instala.sh
```

Una vez se hayan copiado todos los ficheros modificados a sus ubicaciones correctas, podremos seguir con la configuración del buscador.

Configuración del buscador

Veamos cuáles son los parámetros más importantes que debemos tener en cuenta en el fichero de configuración (/etc/htdig/htdig.conf).

El primero de ellos es la situación de la base de datos donde almacenará los índices que crea, que en este caso será en /var/lib/htdig:

```
database_dir: /var/lib/htdig
```

Deberemos indicar la ruta del servidor a partir del la cual queremos que realice la indexación de los ficheros. Esto se define mediante la directiva:

```
start_url: http://litio.sistemasop.ui/gso/
```

También podemos configurar las extensiones de los ficheros sobre los que no queremos que realice la indexación (ficheros binarios comprimidos, imágenes, etc...):

```
bad_extensions: .wav .gz .z .bz2 .sit .au .zip .tar .hqx .exe .com \
   .gif .jpg .jpeg .aiff .class .map .ram .tgz .bin .rpm .mpg .mov .avi .css
```

Y el tamaño máximo de los archivos indexados, que fijaremos en 200 KB:

```
max_doc_size: 200000
```

Ejemplo de fichero de configuración de ht://Dig

Este es el fichero de configuración del buscador instalado en uno de los servidores. Se incluyen los comentarios originales para una mejor comprensión del significado de algunas de las directivas de configuración.

```
#
# Example config file for ht://Dig.
#
# This configuration file is used by all the programs that make up ht://Dig.
# Please refer to the attribute reference manual for more details on what
# can be put into this file. (See http://www.htdig.org/confindex.html or
# locally at file:/usr/share/doc/htdig-doc/htmlconfindex.html)
# Note that most attributes have very reasonable default values so you
# really only have to add attributes here if you want to change the defaults.
#
# What follows are some of the common attributes you might want to change.
#

#
# Specify where the database files need to go.  Make sure that there is
# plenty of free disk space available for the databases.  They can get
# pretty big.
#
database_dir: /var/lib/htdig

#
# This specifies the URL where the robot (htdig) will start.  You can specify
# multiple URLs here.  Just separate them by some whitespace.
# The example here will cause the ht://Dig homepage and related pages to be
# indexed.
# You could also index all the URLs in a file like so:
# start_url:        `${common_dir}/start.url`
#

start_url: http://litio.sistemasop.ui/gso/

#
# This attribute limits the scope of the indexing process.  The default is to
# set it to the same as the start_url above.  This way only pages that are on
# the sites specified in the start_url attribute will be indexed and it will
# reject any URLs that go outside of those sites.
#
# Keep in mind that the value for this attribute is just a list of string
# patterns. As long as URLs contain at least one of the patterns it will be
# seen as part of the scope of the index.
#
limit_urls_to: ${start_url}
```

```
#
# If there are particular pages that you definitely do NOT want to index, you
# can use the exclude_urls attribute.  The value is a list of string patterns.
# If a URL matches any of the patterns, it will NOT be indexed.  This is
# useful to exclude things like virtual web trees or database accesses.  By
# default, all CGI URLs will be excluded.  (Note that the /cgi-bin/ convention
# may not work on your web server.  Check the  path prefix used on your web
# server.)
#
exclude_urls: /cgi-bin/ .cgi

#
# Since ht://Dig does not (and cannot) parse every document type, this
# attribute is a list of strings (extensions) that will be ignored during
# indexing. These are *only* checked at the end of a URL, whereas
# exclude_url patterns are matched anywhere.
#
bad_extensions: .wav .gz .z .bz2 .sit .au .zip .tar .hqx .exe .com \
   .gif .jpg .jpeg .aiff .class .map .ram .tgz .bin .rpm .mpg .mov .avi .css

#
# The string htdig will send in every request to identify the robot.  Change
# this to your email address.
#
maintainer: webmaster@litio.sistemasop.ui

#
# The excerpts that are displayed in long results rely on stored information
# in the index databases.  The compiled default only stores 512 characters of
# text from each document (this excludes any HTML markup...)  If you plan on
# using the excerpts you probably want to make this larger.  The only concern
# here is that more disk space is going to be needed to store the additional
# information.  Since disk space is cheap (! :-)) you might want to set this
# to a value so that a large percentage of the documents that you are going
# to be indexing are stored completely in the database.  At SDSU we found
# that by setting this value to about 50k the index would get 97% of all
# documents completely and only 3% was cut off at 50k.  You probably want to
# experiment with this value.
# Note that if you want to set this value low, you probably want to set the
# excerpt_show_top attribute to false so that the top excerpt_length characters
# of the document are always shown.
#
max_head_length: 10000

#
# To limit network connections, ht://Dig will only pull up to a certain limit
# of bytes. This prevents the indexing from dying because the server keeps
# sending information. However, several FAQs happen because people have files
# bigger than the default limit of 100KB. This sets the default a bit higher.
# (see <http://www.htdig.org/FAQ.html> for more)
#
max_doc_size: 200000
```

```
#
# Most people expect some sort of excerpt in results. By default, if the
# search words aren't found in context in the stored excerpt, htsearch shows
# the text defined in the no_excerpt_text attribute:
# (None of the search words were found in the top of this document.)
# This attribute instead will show the top of the excerpt.
#
no_excerpt_show_top: true

#
# Depending on your needs, you might want to enable some of the fuzzy search
# algorithms.  There are several to choose from and you can use them in any
# combination you feel comfortable with.  Each algorithm will get a weight
# assigned to it so that in combinations of algorithms, certain algorithms get
# preference over others.  Note that the weights only affect the ranking of
# the results, not the actual searching.
# The available algorithms are:
# accents
# exact
# endings
# metaphone
# prefix
# soundex
# substring
# synonyms
# By default only the "exact" algorithm is used with weight 1.
# Note that if you are going to use the endings, metaphone, soundex, accents,
# or synonyms algorithms, you will need to run htfuzzy to generate
# the databases they use.
#
search_algorithm: exact:1 synonyms:0.5 endings:0.1

#
# The following are the templates used in the builtin search results
# The default is to use compiled versions of these files, which produces
# slightly faster results. However, uncommenting these lines makes it
# very easy to change the format of search results.
# See <http://www.htdig.org/hts_templates.html> for more details.
#
# template_map: Long long ${common_dir}/long.html \
# Short short ${common_dir}/short.html
# template_name: long

#
# The following are used to change the text for the page index.
# The defaults are just boring text numbers.  These images spice
# up the result pages quite a bit.  (Feel free to do whatever, though)
#
next_page_text: <img src="/htdig/buttonr.gif" border="0" align="middle" width="30" height
no_next_page_text:
prev_page_text: <img src="/htdig/buttonl.gif" border="0" align="middle" width="30" height
no_prev_page_text:
```

```
page_number_text: ’<img src="/htdig/button1.gif" border="0" align="middle" width="30" hei
’<img src="/htdig/button2.gif" border="0" align="middle" width="30" height="30" alt="2">
’<img src="/htdig/button3.gif" border="0" align="middle" width="30" height="30" alt="3">
’<img src="/htdig/button4.gif" border="0" align="middle" width="30" height="30" alt="4">
’<img src="/htdig/button5.gif" border="0" align="middle" width="30" height="30" alt="5">
’<img src="/htdig/button6.gif" border="0" align="middle" width="30" height="30" alt="6">
’<img src="/htdig/button7.gif" border="0" align="middle" width="30" height="30" alt="7">
’<img src="/htdig/button8.gif" border="0" align="middle" width="30" height="30" alt="8">
’<img src="/htdig/button9.gif" border="0" align="middle" width="30" height="30" alt="9">
’<img src="/htdig/button10.gif" border="0" align="middle" width="30" height="30" alt="10"
#
# To make the current page stand out, we will put a border around the
# image for that page.
#
no_page_number_text: ’<img src="/htdig/button1.gif" border="2" align="middle" width="30"
’<img src="/htdig/button2.gif" border="2" align="middle" width="30" height="30" alt="2">
’<img src="/htdig/button3.gif" border="2" align="middle" width="30" height="30" alt="3">
’<img src="/htdig/button4.gif" border="2" align="middle" width="30" height="30" alt="4">
’<img src="/htdig/button5.gif" border="2" align="middle" width="30" height="30" alt="5">
’<img src="/htdig/button6.gif" border="2" align="middle" width="30" height="30" alt="6">
’<img src="/htdig/button7.gif" border="2" align="middle" width="30" height="30" alt="7">
’<img src="/htdig/button8.gif" border="2" align="middle" width="30" height="30" alt="8">
’<img src="/htdig/button9.gif" border="2" align="middle" width="30" height="30" alt="9">
’<img src="/htdig/button10.gif" border="2" align="middle" width="30" height="30" alt="10"

#
# These files don’t belong to /etc/htdig
#
synonym_db: /usr/lib/htdig/synonyms.db
endings_root2word_db: /usr/lib/htdig/root2word.db
endings_word2root_db: /usr/lib/htdig/word2root.db

#
# image_url_prefix is the relative url for the images htdig uses.
# This defaults to "/doc/htdig/images".  To get the images from another dir
# if you’ve turned off the /doc alias, you’ll need to modify the .html files
# in /etc/htdig as well as this setting.
#
image_url_prefix: /htdig

#
# external parser for msword, postscript and pdf files
#
# If you really want to parse pdf, ps and doc files, make sure that the
# max_doc_size value above is set to a larger value than your largest file
# of this type. If it is smaller, the indexing might hang
#
#external_parsers: application/msword /usr/share/htdig/parse_doc.pl \
# application/postscript /usr/share/htdig/parse_doc.pl \
# application/pdf /usr/share/htdig/parse_doc.pl

#
# special Debian variable for PDF documents
```

```
#
# If you want to parse PDF documents, set the (htdig variable) pdf_parser
# to the wrapper script /usr/bin/htdig-pdfparser (which is default). Using
# the debian_pdf_parser variable the actual pdf parser can be controlled.
# Make sure you set the external_parsers option too.
# Recognized options are: acrobat, xpdf
#
#debian_pdf_parser: acrobat

# local variables:
#
# Definición del idioma español

locale:                         es_ES
endings_dictionary:             ${common_dir}/espa~nol.0
endings_affix_file:             ${common_dir}/espa~nol.aff
bad_word_list:                  ${common_dir}/bad_words.es
synonym_dictionary:             ${common_dir}/synonyms.es

#
# Otros valores comunes.
#
method_names:                   and Todas or Cualquier boolean Booleana
sort_names:                     score Conteo time Fecha title Título revscore \
                'Conteo Inverso' revtime 'Fecha Inversa' revtitle \
                'Título Inverso'
template_map:                   Largo builtin-long builtin-long Corto builtin-sh
ort \
                builtin-short
# matches_per_page:             100
max_stars:                      5
# use_star_image:                       false
no_excerpt_text:                <em>(No hay texto en el documento)</em>
page_list_header:               <hr noshade size=2>P&aacute;ginas:<br>
```

Prueba

Una vez esté todo configurado, deberemos realizar la indexación de los ficheros. La primera vez que se realiza tarda bastante tiempo, aunque depende de la cantidad de documentos a indexar. Esto es debido a que esta primera vez, debe construir el índice de palabras y sinónimos, proceso que es bastante lento. En las siguientes indexaciones para actualizar la base de datos, este proceso no será necesario y se realizará de una forma muchísimo más rápida.

La indexación se realiza mediante el comando **rundig**. Para ello debemos ejecutarlo:

```
# rundig
```

Una vez finalizado el proceso, ya estará todo listo para poder realizar búsquedas.

Accederemos a la interfaz web del buscador, que estará situada en la dirección:

```
http://nombre.del.servidor/search.html
```

desde la cual podremos realizar cualquier búsqueda que queramos y veremos los resultados obtenidos.

> **Importante:** La actualización de los índices se realiza de forma automática una vez al día, aunque puede forzarse su realización en cualquier momento mediante la ejecución del comando **rundig** mencionado anteriormente.

Más información

Se recomienda la lectura de toda la documentación que acompaña al buscador, estando ésta situada en /usr/share/doc/htdig-doc/.

En dicha documentación veremos entre otras muchas cosas los pasos que hay que seguir para poder activar la búsqueda e indexación en archivos PDF, PS, DOC, RTF, etc, así como las explicaciones de los distintos tipos de indexación y búsqueda que puede realizar; ya que lo aquí visto es sólo una pequeña parte de las posibilidades que ofrece ht://Dig.

Otra fuente de información puede ser la página web oficial: http://www.htdig.org/ (http://www.htdig.org), en la cual se encuentra la última versión de la documentación disponible, así como varias secciones de interés como pueden ser las FAQ (preguntas frecuentes), y las listas de correo de usuarios de ht://Dig.

Sobre este documento

Este documento ha sido escrito en formato XML utilizando la DTD de DocBook (http://www.docbook.org). Mediante este sistema, puede ser fácilmente transformado a múltiples formatos (HTML, TXT, PDF, PostScript, LaTeX, DVI, ...). Se recomienda su utilización como herramienta de documentación potente y libre.