Feeds. Estudio de características para su aplicación a la Web Semántica Geo-Espacial.

José Eduardo Córcoles
Universidad de Castilla-La Mancha - España
corcoles@dsi.uclm.es

Resumen
En este trabajo se pretenden estudiar las posibilidades de Google y su funcionamiento para ver la viabilidad de este motor de búsqueda como herramienta para las Web Semántica Geo-Espacial [Egenhofer, M][Córcoles, J.E. y González P.] . Para ello se comienza detallando la arquitectura del motor de búsqueda así como las estructuras de datos que emplea. Posteriormente, se describen los protocolos que emplean las soluciones empresariales que presenta Google, tanto para la búsqueda como para permitir el envío de información al índice que mantiene con todos los documentos que tiene indexados. En el último punto se esboza una arquitectura basada en los elementos básicos de Google para la búsqueda Geo-espacial con características semánticas.

1. ARQUITECTURA DE GOOGLE

El buscador de Google se compone de 14 módulos, estos módulos junto con sus relaciones se pueden ver en la figura 1, que representa la arquitectura de alto nivel de Google. A continuación se va a describir cada uno de los módulos de los que se compone el buscador.

- El módulo ***Crawler*** se compone de varios sistemas distribuidos.
- El ***URL Server*** es el encargado de enviar las URL al *Crawler* para que las analice.
- Las páginas que ya han sido analizadas se envían al ***Store Server*** para su almacenamiento.
- El *Store Server* comprime y almacena las páginas web para su almacenamiento en el ***Repository***. Además, cada página tiene asociado un identificador numérico llamado *docID*, el cual se asigna siempre que una nueva URL es analiza.

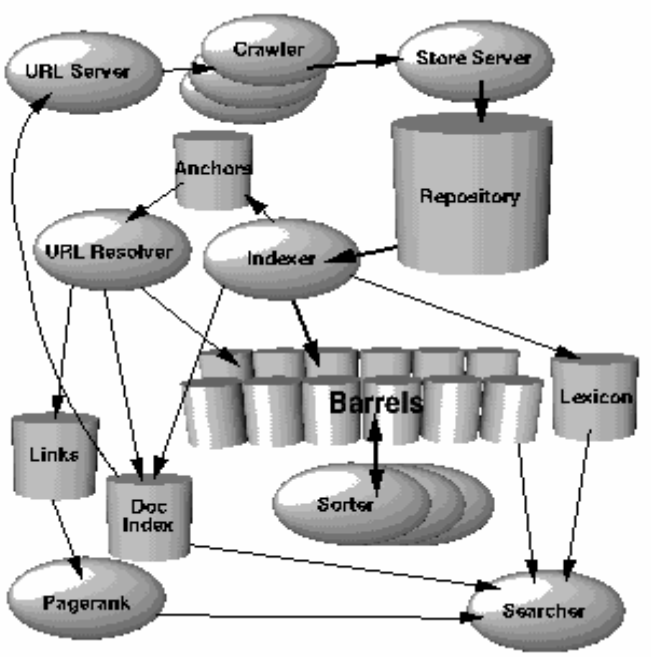

Figura 1. Arquitectura de alto nivel del buscador Google.

- La función de indexación es llevada a cabo por los módulos ***Indexer*** y ***Sorter***. El *indexer* realiza un gran número de funciones, entre ellas se encuentran, leer, descomprimir y analizar los documentos que se encuentran almacenados en el repositorio. Cada documento es convertido en un conjunto de ocurrencias de palabras que aparecen en él, llamadas *hits*. Estos *hits* almacenan la palabra y la posición en la que ésta se encuentra dentro del documento. El *indexer* distribuye estos *hits* en un conjunto de ***Barrels***, creando un índice ordenado.

  Además el módulo *indexer* realiza otras funciones importantes. Analiza todos los enlaces (***Links***) que contienen las páginas web y almacena, en un fichero llanada ***Anchors***, información importante sobre ellos. Este fichero contiene información suficiente para determinar donde se encuentran los enlaces, hacia donde apuntan y el texto que aparece en dichos enlaces.
- El ***URL Resolver*** lee el fichero que contiene la información de los enlaces y convierte las URL relativas en absolutas y los asigna un *docID*. Posteriormente guarda el fichero con la información de los enlaces en el índice, asociándolo con el identificador de la URL (*docID*). Los enlaces que se encuentran en la base de datos son utilizados para calcular el ***PageRank™*** de todos los documentos.

- El ***Sorter*** toma los *Barrels*, que están almacenados según su valor de *docID* y los reordena según el valor del *wordID* para generar un índice invertido. Esta operación no consume demasiado tiempo.
- Un programa llamado *DumpLexicon* toma la lista junto con el ***Lexicon*** producido por el *indexer* y genera un nuevo *Lexicon* que pueda ser utilizado por el *Searcher*.
- El ***Searcher*** es ejecutado por un servidor web y utiliza el *Lexicon* construido por el programa *DumpLexicon* junto con el índice invertido y el *PageRank™* para responder a las consultas de los usuarios.

2. PROTOCOLO DE BÚSQUEDA

Google ha desarrollado un protocolo simple basado en HTTP para proporcionar los resultados de búsqueda. De esta forma los administradores tienen un control completo sobre como los resultados de búsqueda son solicitados y mostrados al usuario final.

Las soluciones empresariales de Google [Google GSA] aceptan peticiones de búsqueda como entradas y devuelven los resultados de las búsquedas como salida. Las peticiones de búsqueda, son simples peticiones HTTP al motor de búsqueda corporativo de Google debido a que normalmente los usuarios utilizan un formulario web para realizar dichas peticiones. Aunque también se pueden realizar peticiones desde otras aplicaciones haciendo la petición HTTP correspondiente. Por otro lado, los resultados de búsqueda pueden ser devueltos en dos formatos diferentes: HTML o XML.

El formato de salida HMTL puede ser mostrado directamente en un servidor web. El motor de búsqueda genera directamente la página HTML con los resultados aplicando una transformación XSLT, al aplicar una hoja de estilos XSL sobre el documento XML que posee los resultados. Por tanto, el administrador del equipo puede personalizar la página de resultados modificando esta hoja de estilos. Mientras que el formato de salida XML proporciona facilidad para procesar los resultados de salida en otras aplicaciones web.

Utilizar el protocolo de búsqueda de Google es tan simple como realizar una petición de una página a un servidor web. Para ello, se utiliza el comando GET de HTTP, el cual devuelve los resultados en formato XML o HTML. Google recomienda utilizar la versión 1.0 o posterior del comando GET.

3. PROTOCOLO FEED

Este es el protocolo utilizado para desarrollar conectores personalizados con el fin de enviar documentos al motor de búsqueda de Google para procesarlos, indexarlos y servirlos como resultados de búsqueda [WGoogleFeed].

Para crear *feed*, será necesario encapsular los datos dentro de un fichero XML. Una vez hecho esto, el fichero se enviará al motor de búsqueda utilizando el protocolo HTTP a través de un formulario web o un *script*. A estos *script* que crean los ficheros XML y los envían al motor de búsqueda se los denominan conectores personalizados.
Existen dos tipos de *feeds*:

- **Content feed:** este tipo de *feed* incluye la URL y el contenido de la URL, y todo ellos se encapsula dentro de un documento XML [WGoogleFeed]. Los documentos enviados a través de este método no los vuelve a procesar el rastreador, sin embargo, los enlaces que pueda contener si los encola el rastreador para su posterior procesamiento. Un ejemplo de este tipo de documento se puede ver en el listado 1.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE gsafeed PUBLIC "-//Google//DTD GSA Feeds//EN" "">
<gsafeed>
        <header>
                <datasource>pruebas</datasource>
                <feedtype>full</feedtype>
        </header>
        <group>
        <record url=http://217.15.35.247/db/fichero11089.xml mimetype="text/plain">
                <metadata>
                        <meta name="tipo" content="cartografia"/>
                        <meta name="nombre" content="Sierra Magina"/>
                </metadata>
                <content> <![CDATA[
                        <?xml version="1.0" encoding="ISO-8859-1"?>
                        <info>
                                <feature>
                                        <capa>PARQUE NATURAL</capa>
                                        <campo0>Sierra Magina</campo0>
                                        <campo1>Poligono</campo1>
                                </feature>
                                <operacion>contenido</operacion>
                                <feature>
                                        <capa>PROVINCIAS ANDALUCIA</capa>
                                        <campo0>JAEN</campo0>
                                        <campo1>Poligono</campo1>
                                </feature>
                        </info>]]>
                </content>
        </record>
        </group>
</gsafeed>
```

Listado 1. Ejemplo de fichero *content feed*.

- **Web feed:** son ficheros XML que contiene la URL de los documentos que se desean indexar, pero no su contenido [WGoogleFeed]. El rastreador encola las URL para posteriormente irlas recuperando con el

fin de realizar el proceso de indexación. Los documentos enviados de esta forma se los asigna el *PageRank*™ más bajo posible. Un ejemplo de documento *web feed* se puede ver en el listado 2.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE gsafeed PUBLIC "-//Google//DTD GSA Feeds//EN" "">
<gsafeed>
        <header>
                <datasource>web</datasource>
                <feedtype>incremental</feedtype>
        </header>
        <group>
                 <record url="http://www.albacete.org" mimetype="text/html"/>
        </group>
</gsafeed>
```

Listado 2. Ejemplo de fichero *web feed*.

Google es capaz de descubrir nuevos documentos que necesita indexar de forma automática, sin necesidad de comunicárselo, a través de los enlaces que poseen las páginas hacia ellos. Por tanto, es aconsejable enviar documentos mediante el protocolo *feed* únicamente en los siguientes casos:

- Documentos que no pueden ser recuperados por el rastreador. Por ejemplo, registros de una base de datos o documentos que no están publicados en la Web.
- Documentos que están publicados en Internet pero hacia los que no hay enlaces y, por tanto, no pueden ser recuperados por el rastreador
- Documentos que pueden ser recuperados por el rastreador, pero se prefieren enviar a través de este protocolo para evitar los continuos accesos por parte del rastreador, solo en el caso de *content feed*.
- Documentos que pueden ser recuperados por el rastreador, pero que resulta muchos más rápido enviarlos a través del *feed*, debido a problemas en el servidor web o problemas en la red.

Los parámetros, que se corresponden con etiquetas XML, que se deben especificar a la hora de construir un fichero para ser enviado a través del protocolo *feed*, son los siguientes:

- **Fuente de datos (datasource):** una fuente de datos es el nombre que se especifica cuando se envía un *feed* al motor de búsqueda. Éste mantiene una lista con todas las URLs que forman parte de una fuente de datos. Se especifica mediante la etiqueta *datasource* dentro del fichero XML. Para los ficheros de *content feed* se puede especi-

ficar cualquier nombre para la fuente de datos (en el ejemplo anterior se llama pruebas). Mientras que para los *web feed* es obligatorio poner como nombre de la fuente de datos *web*, como se puede observar en el ejemplo anterior de *web feed*.

- **Tipo de feed (feedtype):** determina el tipo de *feed* del documento. Puede tomar los valores *full* o *incremental*. En caso de tratarse de un *full feed* todos los elementos que existen en el motor de búsqueda asignados a la fuente de datos especificada en el documento, los elimina y asigna los documentos enviados en el fichero actual a la fuente de datos señalada. Por ejemplo, si la fuente de datos *pruebas* tiene asignado dos documentos, D1 y D2, y se envía un *full feed* sobre esa fuente de datos con los documentos D3, D4 y D5, la fuente de datos dejará de servir los documentos D1 y D2 para pasar a servir los documentos D3, D4 y D5.

  Mientras que si se trata de un *incremental feed* los documentos que se envían en el fichero se añaden a los que ya tenía asignado la fuente de datos. Por ejemplo, si para el caso anterior, en vez de enviar un *full feed* se envía un *incremental feed*, la fuente de datos servirá los documentos D1, D2, D3, D4 y D5.
- **Metadatos (metadata):** esta etiqueta es opcional. Los metadatos son etiquetas que poseen un nombre y un determinado valor, y se pueden añadir a los documentos enviados a través de un *feed*, o se pueden enviar de forma aislada, para su posterior recuperación y tratamiento en las aplicaciones personalizadas que se deseen desarrollar. Si se añaden a un documento de un *feed*, los metadatos y el documento se almacenarán en el índice como un único archivo. Mientras que si se envían de forma individual se almacenarán en archivos separados.

- **Registro (record):** en esta etiqueta es donde se encuentra el documento que se pretende enviar, ya sea la URL o la URL y el contenido. Además puede tener una serie de atributos:
  - **URL (url):** indica donde se encuentra el documento, para que una vez mostrados los resultados, Google pueda enlazar con él.

- **Acción (action):** determina la acción que se va a realizar con el documento, puede ser *delete* o *add*. Si no se indica nada la acción por defecto es *add*. Si se desea borrar un documento del índice se puede enviar un *feed* del documento con la acción borrar.
- **Bloqueo (lock):** cuando el motor de búsqueda alcanza el límite máximo de documentos a indexar, los ficheros de menor *PageRank™* son eliminados del índice, con la finalidad de dejar espacio para documentos con mayor *PageRank™*. Pero mediante este parámetro se puede indicar que se queden bloqueados y no se eliminen. Pude tomar dos valores *0* (documento no bloqueado) y *1* (documento bloqueado).
- **Tipo de documento (mimetype):** este parámetro es obligatorio, indica el tipo de documento que va en el registro (pdf, html, texto,…).
- **Última modificación (last-modified):** determina la fecha de la última modificación del documento, en caso de no especificarla, el motor de búsqueda le asigna la fecha en la que se envió el documento. El formato de la fecha es el especificado en la RFC822 (ejemplo, Mon, 15 oct 2006 10:20:06 GMT).
- **Método de autentificación (authmethod):** indica los usuarios que tienen autorización para poder ver los documentos en los resultados.

- **Contenido (content):** determina el contenido del registro, en caso de tratarse de un *content feed*. Esta etiqueta puede tener un atributo:
  - **Codificación (encoding):** se utiliza para documentos que no son de texto plano como pdf, doc, etc. Estos documentos deben ser codificados utilizando base64, por lo que este parámetro debe tomar el valor *base64binary*. Esto aumenta el tamaño del documento una tercera parte, por lo que es más conveniente enviar este tipo de ficheros como *web feed* mandando la URL donde se encuentra el documento.

Un ejemplo de la utilización de las etiquetas, arriba descritas, dentro de los ficheros *feed* puede ser el que se muestra en el listado 3.

```
<?xml version="1.0" encoding="UTF8"?>
<!DOCTYPE gsafeed PUBLIC "-//Google//DTD GSA Feeds//EN" "">
<gsafeed>
        <header>
                <datasource>pruebas</datasource>
                <feedtype>incremental</feedtype>
        </header>
        <group>
        <record url="http://www.corp.enterprise.com/hola01" mimetyp="text/plain"
        last-modified="Tue, 15 Nov 1994 12:45:26 GMT">
                <content> Esto es hola01 </content>
        </record>
        <record url="http://www.corp.enterprise.com/hola02" mimetype="text/plain"
                  lock="true">
      <content> Esto es hola02 </content>
        </record>
        <record url="http://www.corp.enterprise.com/hola03" mimetype="text/html">

                <metadata>
                        <meta name="tipo" content="pagina web"/>
                        <meta name="nombre" content="hola03"/>
                </metadata>
                <content><![CDATA[
                        <html>
                             <title>Hola03</title>
                             <body>esto es hola03</body>
                        </html>]]>
              </content>
        </record>
        <record url=http://www.corp.enterprise.com/hola04.pdf action="delete"
                  mimetype="text/pdf">
                <content encoding="base64binary">Zm9vIGJhcgo</content>
        </record>
        </group>
</gsafeed>
```

Listado 3. Ejemplo de utilización de los parámetros *feed*.

4. ARQUITECTURA BASADA EN FEEDS

La figura 1.1 muestra un esbozo de la arquitectura diseñada para el motor de búsqueda geo-espacial. En ella interviene: (i) un **Mediador**, encargado de atender las consultas del usuario, procesarlas y ejecutarlas; (ii). un **sistema de base de datos espacial**, en donde se guardan la información espacial sobre la cual se aplicará los criterios espaciales; (iii) y por último, la **GSA Google** encargada de indexar y devolver los recurso web que cumplan un cierto criterio.

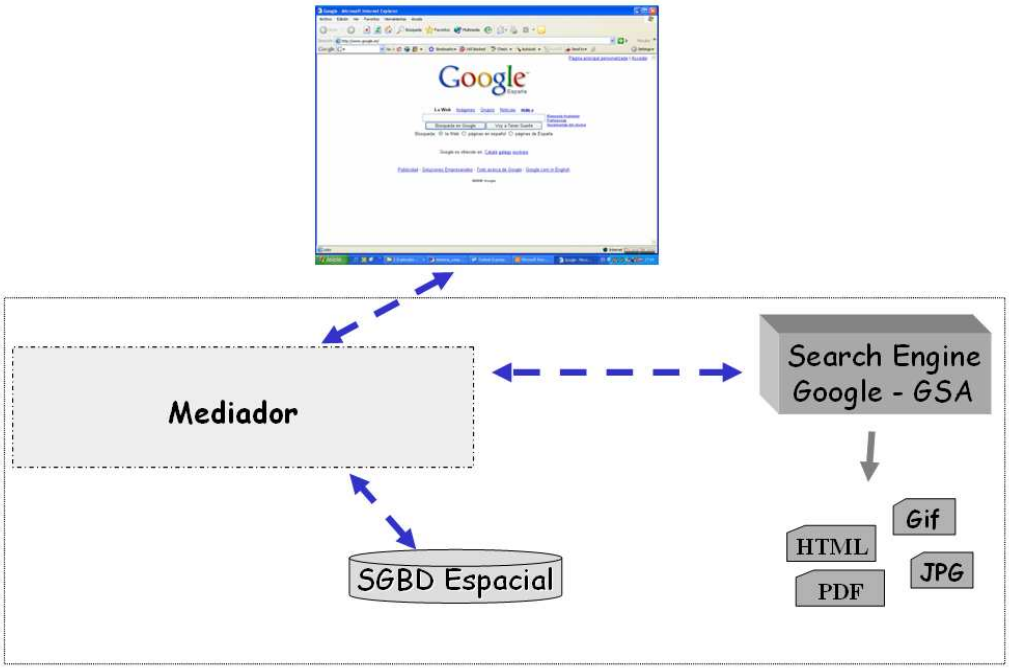

Figura 2. Esbozo de la arquitectura diseñada.

Así, en un escenario de ejecución, el usuario escribe las consultas desde el entorno del buscador (para ello, necesita usar un lenguaje de consulta, que será una extensión del lenguaje usado por Google). La consulta es tratada por el mediador quien satisface las características espaciales buscando en el índice del motor de búsqueda. Posteriormente, el mediador envía una segunda consulta a la GSA para obtener los recursos web (html, gif, pdf) relacionados con los datos espaciales obtenidos.

5. Bibliografía

Córcoles, J.E. y González P. "Integrating gml resources and other web resources". 1st International Workshop on Geographic Information Management (GIM'04) in Conjuntion with DEXA'04. Zaragoza. Spain. IEEE Computer Society Press (2004)

Egenhofer, M. "Toward the Semantic Geospatial Web". In Proc. 10th ACM International Symposium on Advances in Geographic Information System (ACM-GIS 2002). Washington (USA). 2002.

Google GSA. "Web con información corporativa de Google". WGoogleCorp. http://www.google.es/intl/es/corporate/index.html
WGoogleFeed. "Web del protocolo feed de Google."
http://code.google.com/enterprise/documentation/feedsguide.html

SOCIEDAD DE LA INFORMACION

www.sociedadelainformacion.com

Edita:

Director: José Ángel Ruiz Felipe
Jefe de publicaciones: Antero Soria Luján
D.L.: AB 293-2001
ISSN: 1578-326x