Aproximación a la búsqueda basada en términos sobre conjuntos de datos medioambientales

David Álvarez-Castro, José R.R. Viqueira, and Alberto Bugarín

Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela
Santiago de Compostela, Spain
david.alvarez.castro@rai.usc.es, {jrr.viqueira, alberto.bugarin.diz.}@usc.es

Resumen En este artículo se discuten los trabajos, actualmente en curso, de diseño e implementación de un sistema de búsqueda por términos sobre fuentes de datos medioambientales, entre las que se incluyen fuentes de entidades geográficas y arrays que almacenan la variación espacio-temporal de distintas variables geo-físicas. Este tipo de sistemas facilitan el descubrimiento y el acceso a fuentes de datos de naturaleza científica a usuarios no expertos, que pueden utilizarlas en aplicaciones de muy diverso tipo.

Keywords: Búsqueda por términos, Datos Medioambientales, Recuperación de Información, Búsqueda Geoespacial

1. Introducción

Muchas disciplinas científicas necesitan para sus estudios datos sobre las condiciones cambiantes del medio ambiente. Un ejemplo de esta necesidad, en el área de la salud pública, puede ser el análisis del riesgo de aparición de epidemias de cólera [1], donde se plantea la localización de, por ejemplo, "Zonas de alta temperatura del agua de mar y elevada precipitación". Datos de este tipo, también combinados con otros datos de naturaleza geo-espacial, son de gran importancia para la toma de decisiones en muchas otras áreas como el turismo, en el que una necesidad de información posible sería: "Playas con poco oleaje y temperatura agradable cerca de algún punto de interés cultural". En la actualidad, sin embargo, la resolución de este tipo de necesidades de información requieren la intervención de expertos que conozcan la existencia, ubicación, disponibilidad y características detalladas de cada fuente de datos, así como los medios para acceder a los mismos.

Por su parte, los sistemas de búsqueda por términos han sido implementados con éxito para descubrir y acceder a fuentes de datos no estructuradas como la web. Recientemente, algunas aproximaciones proponen soluciones para la búsqueda por términos sobre fuentes de datos estructuradas, tanto sobre fuentes relacionales como de datos enlazados (Linked Data) [2,3]. Una gran parte de los

datos medioambientales disponibles no encajan en modelos basados en entidades como los anteriores, sino en modelos basados en estructuras de arrays multidimensionales con dimensiones espacio-temporales (datos raster). Por último, en algunas Infraestructuras de Datos Espaciales se proporcionan buscadores basados en términos sobre catálogos de metadatos [4]. Sin embargo, estos buscadores no permiten resolver las necesidades de información descritas anteriormente.

En este artículo se describen los trabajos, actualmente en curso, de diseño e implementación de una primera solución de búsqueda basada en términos para fuentes de datos medioambientales. Los términos que podrá usar el usuario en el sistema de búsqueda incluyen nombres de propiedades espaciales (*temperatura*, *oleaje*, *precipitación*), valores o términos lingüísticos imprecisos (*alto*, *bajo*, *poco*, *agradable*), nombres de entidades geográficas (*Santiago*, *Monasterio de Caaveiro*), nombres de tipos de entidades (*Hotel*, *Monasterio*, *Población*), relaciones espaciales (*cerca de*) y/o referencias a instantes e intervalos de tiempo.

2. Arquitectura del sistema

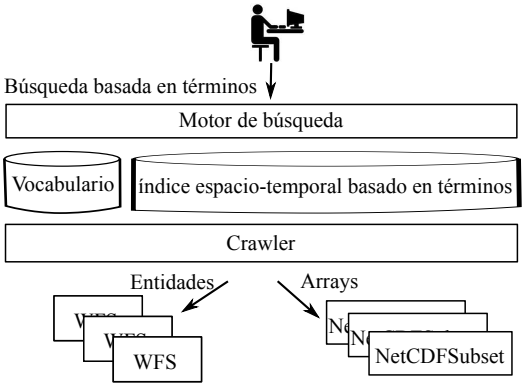

Figura 1. Arquitectura del sistema.

La arquitectura del sistema está basada en la arquitectura típica utilizada por los motores de búsqueda, tal y como se muestra en la Fig. 1. En la parte inferior de la figura se muestran los dos grandes tipos de fuentes de datos (entidades y arrays), que serán accedidos mediante estándares Web Feature Service (WFS) [5] y NetCDFSubset[1], bien conocidos y ampliamente utilizados. El *Crawler* descubre y accede a las fuentes de datos para actualizar el índice espacio-temporal basado en términos que permite responder a las búsquedas. Finalmente, en la

[1] https://www.unidata.ucar.edu/software/thredds/current/tds/reference/NetcdfSubsetServiceReference.html

parte superior de la arquitectura, el motor de búsqueda recibe las consultas basadas en términos (que pueden ser imprecisos) y utiliza el índice para generar la respuesta. Esta respuesta definirá de forma difusa las zonas del espacio y tiempo en las que se cumplen las condiciones especificadas, el grado de dicho cumplimiento (valor real en [0,1]) e incluirá referencias a las fuentes de datos utilizadas para su elaboración.

3. Subsistema de indexación

Para resolver expresiones del tipo "temperatura del agua moderada", el índice debe almacenar el grado de cumplimiento de dicha expresión en cada punto del espacio y del tiempo. Esta información se almacena de forma distinta en función de cómo dicha propiedad cambia en el espacio y tiempo.

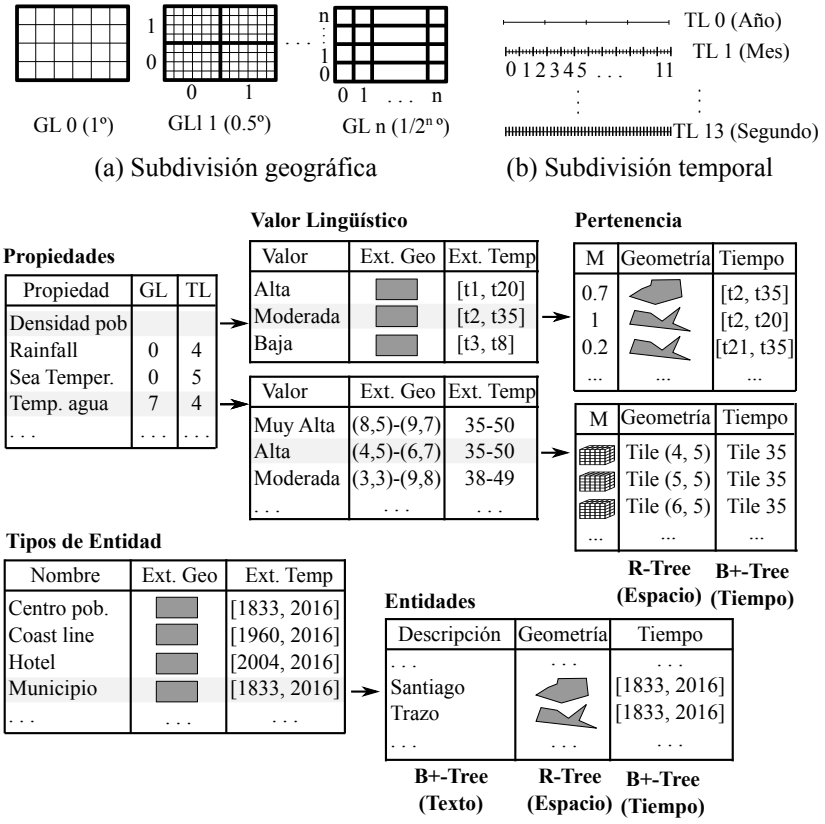

Figura 2. Índice espacio-temporal y baseado en términos

Si la propiedad cambia de forma continua, el cumplimiento para cada término difuso de la propiedad y para cada punto del espacio y tiempo se almacena en arrays tridimensionales. Para lograr una representación alineada en el espacio y el tiempo entre los datos generados de distintas fuentes se definen respectivas subdivisiones jerárquicas del espacio (ver Fig. 2(a)) y del tiempo (ver Fig. 2(b)). Como puede verse en la Fig. 2(c) para la propiedad "Temp. agua", se almacena el nivel en la jerarquía geográfica y temporal en la que se generaron los *tiles* de valores de pertenencia. Para cada etiqueta lingüística de cada propiedad se almacena el rango de *tiles* geográficos y temporales. Cada *tile* tendrá un array tridimensional de valores de pertenencia.

Si la propiedad cambia de forma discreta en el espacio y en el tiempo, se utiliza una representación espacial vectorial para las pertenencias, tal y como se puede ver en la Fig. 2(c) para la propiedad "Densidad pob".

Además de expresiones basadas en propiedades, el sistema permite también expresiones basadas en entidades geográficas y tipos, como por ejemplo "Cerca de Santiago" o "Lejos de un hotel". Para poder resolver estas expresiones, el índice almacena tanto tipos de entidades como entidades. Para cada entidad, además de su descripción textual, se guarda su geometría y tiempo de validez.

Para mejorar la eficiencia de acceso al disco, se utilizan estructuras de indexación para los textos, geometrías y marcas temporales.

4. Trabajo futuro

El trabajo futuro inmediato tiene que ver con la finalización de la implementación del primer prototipo y de su evaluación. A más largo plazo deberán abordarse nuevos retos relacionados con la mejora de la funcionalidad del sistema, su eficacia y su eficiencia. En este último caso será fundamental la incorporación de una arquitectura paralela de altas prestaciones.

Referencias

1. Baker-Austin, C., Trinanes, J.A., Taylor, N.G., Hartnell, R., Siitonen, A., Martinez-Urtaza, J.: Emerging vibrio risk at high latitudes in response to ocean warming. Nature Clim. Change 3(1), 73–77 (2013)
2. Bergamaschi, S., Guerra, F., Interlandi, M., Trillo-Lado, R., Velegrakis, Y.: Quest: A keyword search system for relational data based on semantic and machine learning techniques. Proc. VLDB Endow. 6(12), 1222–1225 (Aug 2013)
3. Demidova, E., Zhou, X., Nejdl, W.: A probabilistic scheme for keyword-based incremental query construction. IEEE Transactions on Knowledge and Data Engineering 24(3), 426–439 (March 2012)
4. Nebert, D., Whiteside, A., Vretanos, P.: OpenGIS Catalogue Services Specification. Open Geospatial Consortium (OGC) (2007), http://www.opengeospatial.org/standards/cat
5. Vretanos, P.: OpenGIS Web Feature Service 2.0 Interface Standard. Open Geospatial Consortium (OGC) (2010), http://www.opengeospatial.org/standards/wfs