Creado el 20/06/2014
Última versión: 20/06/2014
José Manuel Perea Ortega,
Institute for the Protection and Security of the Citizen (IPSC)

Comunidad moocTLH
Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 3.0 Unported.

Aplicaciones de las TLH/RECUPERACIÓN DE INFORMACIÓN

Recuperación de información geográfica 2

¡Hola! Mi nombre es José Manuel Perea, trabajo como investigador para la Comisión Europea en el Joint Research Centre de Ispra en Italia, y voy a presentaros el tema dedicado a recuperación de información geográfica.

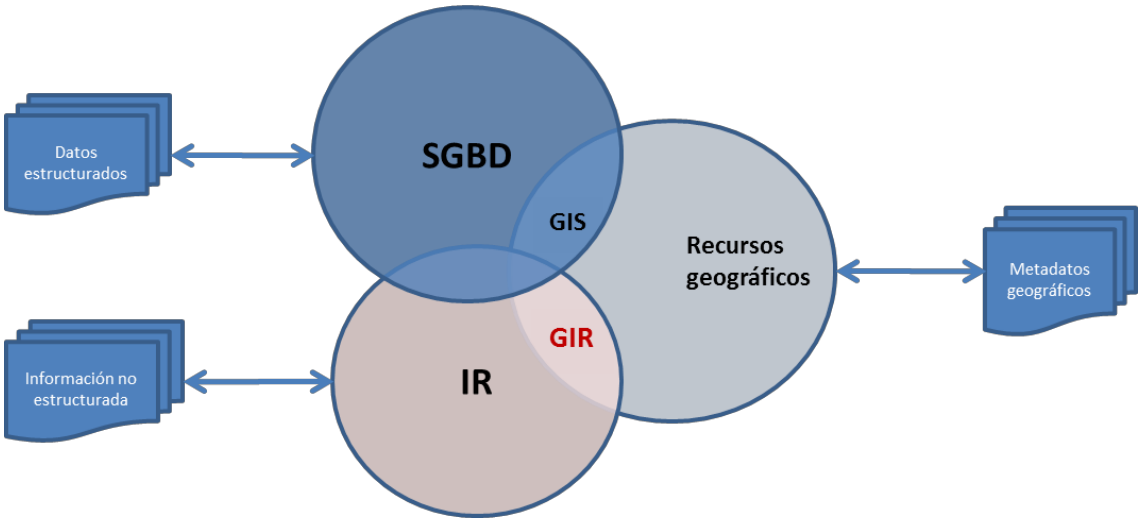
Conceptos básicos

Una vez introducidos brevemente algunos de los problemas a los que se enfrentan los motores de búsqueda actuales, podemos entonces definir la recuperación de información geográfica como el área dentro del procesamiento del lenguaje natural que se encarga de mejorar la eficacia y la eficiencia de los sistemas de recuperación de información que trabajan con entidades geográficas. Si centramos nuestro foco de aplicación en trabajar con colecciones de documentos exclusivamente, podemos ver la tarea de la recuperación de información geográfica como aquella que consiste en recuperar documentos relevantes de un corpus cuando la consulta del usuario viene restringida por un ámbito geográfico determinado.

Con respecto a las similitudes y diferencias que encontramos entre las áreas de recuperación de información geográfica y la de recuperación de información simple, podemos decir que el objetivo de ambas es recuperar información no estructurada, es decir, aquella que normalmente encontramos en cualquier documento de texto o página web. Además, la necesidad de información del usuario viene expresada en ambas mediante lenguaje natural. Por otro lado, la diferencia fundamental radica en que la consulta viene restringida por un ámbito geográfico en los sistemas GIR, mientras que esta restricción no existe en un sistema IR normal. En definitiva, podemos ver la recuperación de información geográfica como una combinación de recuperación de información y cierto razonamiento geográfico.

En cuanto a las similitudes y diferencias entre las áreas de la recuperación de información geográfica y la de sistemas de información geográfica, hay que destacar que en los sistemas GIS el objetivo es recuperar información que se encuentra ya estructurada y por tanto no presenta ambigüedades, es decir, un sistema GIS lo podemos ver como una combinación de bases de datos y mapas digitales. Otra diferencia importante es que las consultas en GIS se realizan mediante un lenguaje formal (normalmente SQL), mientras que las consultas en un sistema GIR se realizan en lenguaje natural. Resumiendo, podríamos decir que las diferencias entre un sistema GIR y un sistema GIS son parecidas a las que encontramos entre un sistema de recuperación de información y un sistema gestor de bases de datos.

En la siguiente diapositiva se resume de modo gráfico la idea explicada con anterioridad. La recuperación de información geográfica se podría considerar un área multidisciplinar ya que, por un lado, permite combinar técnicas procedentes de los sistemas GIS; y por otro, hace uso de métodos y heurísticas que provienen del área de la recuperación de información y del procesamiento del lenguaje natural.



Con respecto a cómo se caracterizan las consultas geográficas en los sistemas GIR, el enfoque que comúnmente se sigue en la actualidad es el que representa la consulta mediante una tupla de tres componentes:

< qué, relación espacial, dónde >

- La parte "qué", que se corresponde con el aspecto no geográfico de la consulta, es decir, la parte temática o foco
- La parte "dónde", que se corresponde con el ámbito geográfico que restringe la necesidad de información del usuario, y
- La relación espacial, que establece la conexión entre el foco y el ámbito geográfico

En los siguientes ejemplos de consultas se puede observar esta representación, resaltando en color verde, rojo y azul, respectivamente, los tres componentes de la tupla mencionados con anterioridad.

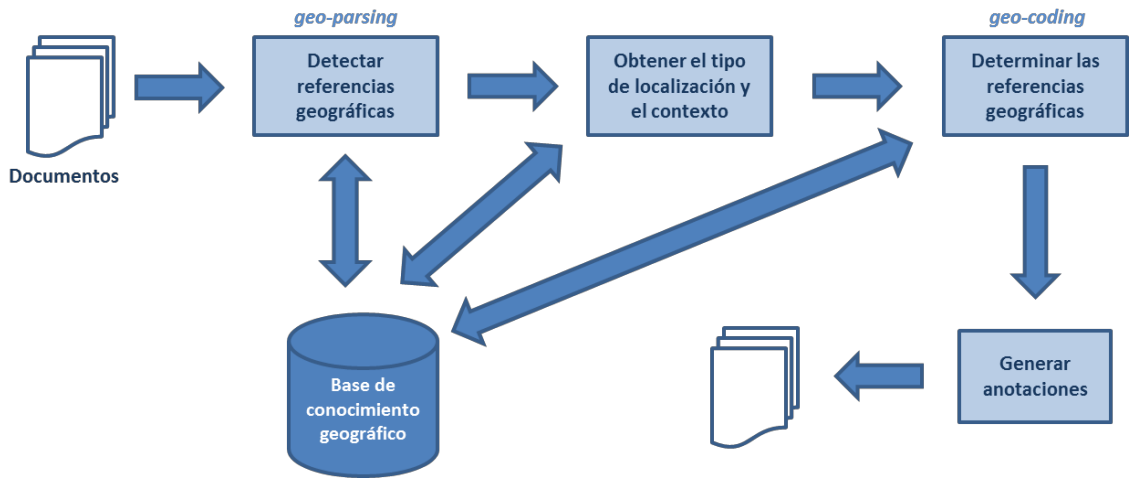
- *"Tormentas de nieve en Norteamérica"*
- *"Atentados con coche bomba cerca de Madrid"*
- *"Incendios forestales en el norte de Portugal"*
- *"Reuniones de la OSCE en el este de Europa"*

Otro concepto importante en recuperación de información geográfica tiene que ver con las relaciones espaciales, es decir, aquellas partículas que establecen una relación entre la temática de la consulta y su ámbito geográfico. Se distinguen principalmente dos tipos de relaciones espaciales: las geométricas, que se refieren a las dimensiones lineales reales, es decir, las que tienen que ver con la dirección y la distancia entre objetos, y las topológicas, que se preocupan más por los aspectos cognitivos de la información geográfica e incluyen conceptos como la adyacencia o la inclusión. Uno de los métodos más utilizados para resolver

relaciones geométricas es el método del gran círculo (*great circle*), que permite averiguar la distancia más corta entre dos puntos sobre una esfera.

Como último concepto básico dentro de esta tarea veremos la definición de geotagging o etiquetado geográfico. El geotagging es el proceso que consiste en generar anotaciones de la información geográfica detectada en un texto. Se distinguen dos fases principales durante el geotagging: el geoparsing y el geocoding. Durante el geoparsing o parseado geográfico se detectan las referencias geográficas en el texto, ignorando aquellas que no tengan un significado geográfico. Por su parte, el geocoding es el proceso mediante el cual se determina la localización o el área geográfica concreta de la entidad detectada durante el geoparsing, es decir, se puede ver como un proceso de desambiguación de topónimos.

En la siguiente diapositiva se muestran las fases principales que componen el etiquetado geográfico de un texto.



En primer lugar, se filtran las referencias geográficas encontradas, obteniendo su tipo y el contexto donde fueron localizadas en el documento. A continuación, se determina a qué localizaciones en concreto hacen referencia dichas entidades y, por último, se generan las anotaciones correspondientes sobre el documento original. Como se puede observar, todo el proceso de etiquetado se apoya en una base de conocimiento geográfico que bien podría ser una base de datos geográfica, una ontología, un gazetteer, etc.