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 3

¡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.

Relevancia en GIR

Una vez introducidos los conceptos básicos de la tarea, pasaremos a continuación a explicar en qué consiste la relevancia en recuperación de información geográfica.

La relevancia en los sistemas GIR va a consistir en dos aspectos principales. Por un lado, debemos tener en cuenta la relevancia temática, es decir, los documentos recuperados por el sistema tienen que responder al qué de la consulta del usuario. Por otro lado, y esto es lo que los diferencia realmente de cualquier sistema de recuperación de información simple, deben tener en cuenta también la relevancia geográfica del documento, es decir, deben responder a la parte dónde detectada en la consulta. Si tomamos como ejemplo la consulta "ataques de tiburones en las costas australianas" podemos distinguir en color verde cuál sería el foco o topic (ataques de tiburones) y en color rojo la respuesta al dónde, es decir, la restricción geográfica, que en ese caso sería "las costas australianas". En color naranja tendríamos la relación espacial que establece la relación entre el foco y el ámbito geográfico de la consulta.

Por consiguiente, en un sistema GIR, la ordenación final de los documentos recuperados deberá tener en cuenta ambos criterios de relevancia. Por un lado, se debe calcular la relevancia temática entre la consulta q y el documento d, y por otro, se debe tener en cuenta la relevancia geográfica entre el documento y la consulta. La relevancia temática viene determinada por las técnicas habituales utilizadas en recuperación de información, como el modelo de espacio vectorial o técnicas de emparejamiento simple. Para el cálculo de la relevancia geográfica existen varios métodos entre los que podemos distinguir los de huella espacial y los de distancia. A continuación pasaré a explicar brevemente las principales características de cada uno.

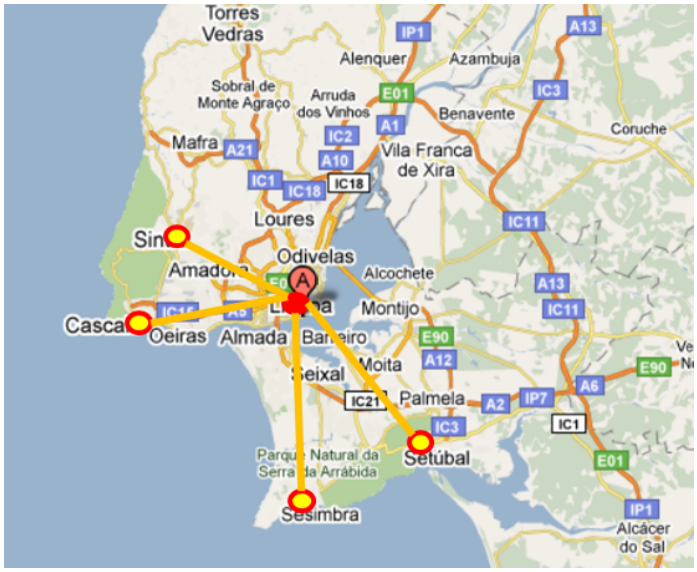
En los métodos de huella espacial o footprint, tanto la consulta como el documento se representan mediante algún tipo de estructura espacial o geométrica (puntos, polígonos, bitmaps, etc.). Veámoslo más claro con un ejemplo. Si tenemos la consulta "hoteles cerca de Lisboa", podríamos representar la huella espacial o footprint de la misma mediante un rectángulo de color rojo como se ve en la imagen, que tendría en cuenta la capital Lisboa y un poco de su área metropolitana. Este cuadrado rojo sería la representación en un mapa de dicha consulta. Por otro lado, si tenemos un documento recuperado con diferentes referencias geográficas como por ejemplo las localizaciones de Sintra, Cascais, Sesimbra y Setúbal, podríamos representar la huella espacial de dicho documento mediante el polígono azul mostrado sobre el mapa.

¿Cómo podríamos entonces calcular la similitud geográfica entre la consulta y el documento mediante sus huellas espaciales? Existen varios métodos para ello. Unos son los métodos de overlapping, que miden la superficie de solapamiento o superposición entre ambas huellas espaciales.

Otros son los métodos de distancia mínima o distancia entre centroides, que calculan la distancia entre determinados puntos de ambas huellas espaciales.

Otros tipos de métodos para calcular la similitud geográfica entre consulta y documento son los métodos de distancia. Estos métodos consideran cada documento como una colección de puntos, limitándose a medir la distancia entre cada punto del documento y el representado

para la consulta. Una aproximación simple consistiría en tener en cuenta el valor más alto de relevancia de todos los puntos del documento. Otra estrategia consiste en calcular la media de todas las distancias de los puntos del documento con la consulta. Entre las técnicas más utilizadas para calcular la distancia entre dos puntos destacan las del método del gran círculo, que considera los puntos sobre una esfera, y el método de la distancia euclídea, que proyecta cualquier punto de la superficie terrestre sobre un plano bidimensional.



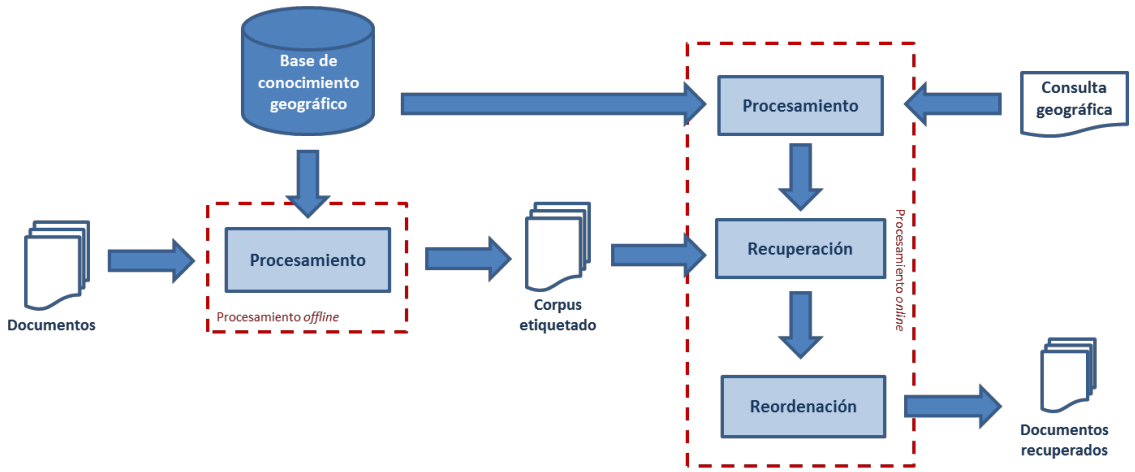
Para terminar este apartado, una vez que hemos calculado la similitud temática y geográfica entre la consulta q y un documento d, hay que combinar ambos criterios para obtener un único valor que nos dé, en una determinada escala, cómo de relevante es dicho documento con respecto a la consulta. Para ello, la estrategia más utilizada a día de hoy es aplicar una combinación lineal entre ambos valores. Entre las heurísticas más comunes destacamos las siguientes:

- Multiplicar directamente ambas similitudes entre sí
  - $sim(q,d) = sim_{text} \times sim_{geo}$
- Tener en cuenta solamente el máximo valor, o
  - $sim(q,d) = max(sim_{text}, sim_{geo})$
- Multiplicar ambas por un valor alfa que nos permita darle mayor importancia a un criterio o a otro.
  - $sim(q,d) = sim_{text} \times \alpha + sim_{geo} \times (1 - \alpha)$

Arquitectura de un sistema GIR

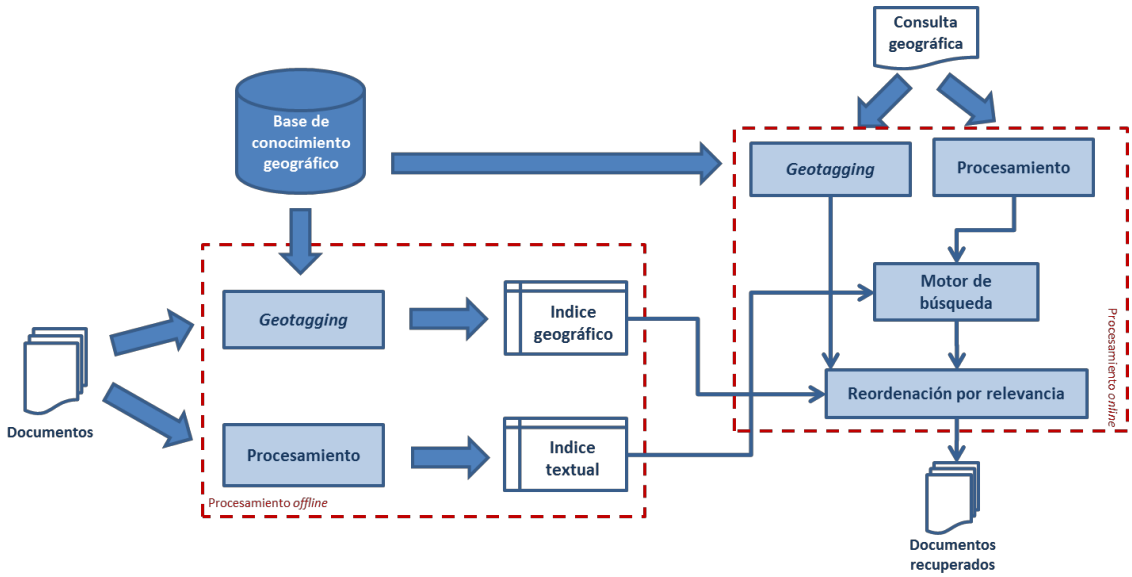
A continuación explicaré las principales características de la arquitectura de un sistema de recuperación de información geográfica.

La siguiente transparencia muestra la arquitectura genérica de un sistema GIR.



Como se puede observar, se distinguen dos fases principales. Una, en la que el sistema no estaría en producción, llamada fase de procesamiento offline, en la que se llevaría a cabo el procesamiento de la colección de los documentos a indexar por el buscador.

En este procesamiento, como veremos en la siguiente transparencia con más detalle, se realizaría todo el proceso de etiquetado y extracción de información de los documentos con los que el sistema va a trabajar. La otra fase, llamada fase de procesamiento online o fase de producción, en la que el sistema, una vez indexados todos los documentos, analizaría la consulta geográfica del usuario y daría respuesta devolviendo los documentos relevantes encontrados para dicha consulta.

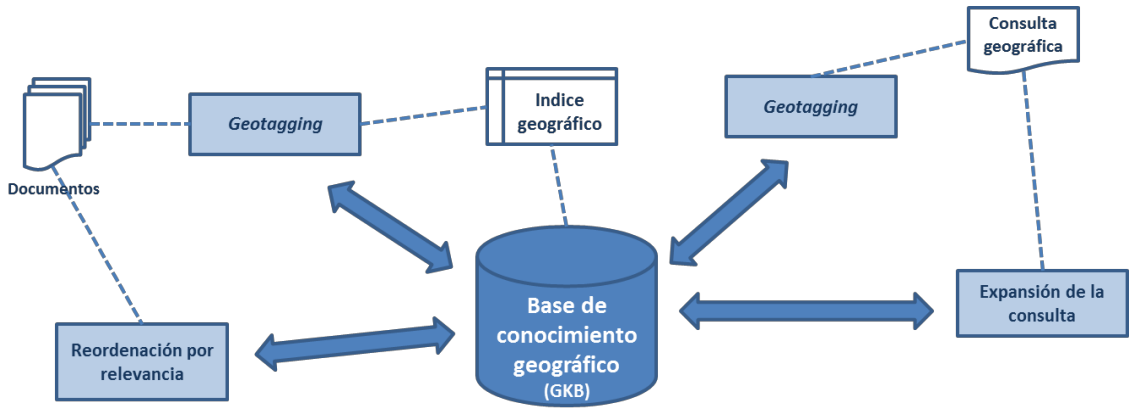


Analicemos con un poco más de detalle las dos principales fases de esta arquitectura. Durante la fase de procesamiento offline se generan dos índices: uno textual, que contendrá los términos relevantes para cada documento, y otro geográfico, en el que se almacenan las

referencias geográficas detectadas durante el proceso de geotagging. Como ya hemos explicado en la sección dedicada a conceptos básicos, este proceso de geotagging consistiría en detectar y desambiguar las referencias geográficas del texto. Por otro lado, durante la generación del índice textual, se aplica un procesamiento que suele estar basado en la utilización de lematizadores, eliminación de las palabras vacías, e incluso en ocasiones se utilizan analizadores léxicos y sintácticos.

Durante la fase de producción se analiza la consulta del usuario aplicando un procesamiento similar al realizado con cualquier documento de la colección. Una vez tenemos la consulta procesada, ésta se envía al motor de recuperación. La lista de documentos devueltos por el motor de búsqueda será reordenada posteriormente teniendo en cuenta la función de relevancia definida en el sistema, que como se ha visto en la sección anterior, deberá contemplar los criterios de similitud temática y geográfica entre consulta y documento.

Como podemos intuir una vez analizada la arquitectura de un sistema de este tipo, el papel que juega la base de conocimiento geográfico que utilicemos en él es crucial. En la siguiente diapositiva pretendo destacar esa cuestión.



Como queda reflejado, la base de conocimiento geográfico interviene prácticamente en todas las fases de estos sistemas: en la interfaz de usuario, para determinar el ámbito geográfico de interés o incluso para reformular la consulta expandiéndola con entidades geográficas relacionadas, durante el proceso de etiquetado geográfico de la colección a la hora de generar el índice de referencias y, por último, durante el proceso de reordenación por relevancia de los documentos devueltos por el motor de recuperación. En la mayoría de ocasiones, esta base de conocimiento geográfico está compuesta por uno o varios recursos externos, que bien podrían ser tesauros, bases de datos geográficas o gazetteers, ontologías geográficas, etc.