Creado el 9/06/2014
Última versión: 9/06/2014
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

¡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.

El esquema que voy a seguir durante la presentación es el siguiente:

- Comenzaré con una breve introducción sobre cómo es interpretada la información geográfica por la mayoría de buscadores web hoy día
- Después repasaremos los conceptos básicos de la tarea de la recuperación de información geográfica
- A continuación explicaré en qué consiste la relevancia en esta tarea y analizaremos los principales componentes de la arquitectura de un sistema de este tipo.
- Después comentaré algunos de los sistemas y recursos disponibles más importantes y,
- para terminar, expondré algunos de los aspectos que todavía quedan por mejorar dentro de esta área

Empecemos pues sin más demora con la introducción.

Introducción

Sabemos que hoy en día la información geográfica, es decir, todo objeto o dato que hace referencia a alguna entidad espacial de nuestro mundo, forma parte de nuestra vida diaria innegablemente. Podemos encontrar información geográfica en cualquier tipo de medio, ya sean documentos de texto, vídeos, libros, bases de datos, imágenes, blogs, tweets, etc. De hecho, es tal la importancia de la información geográfica que diversos estudios científicos concluyen que aproximadamente el ochenta por ciento de las páginas web en Internet contienen alguna localización.

Siguiendo con algunos de esos estudios que analizan el comportamiento de los usuarios en Internet, las estadísticas dicen que alrededor del veinte por ciento de las consultas que se lanzan a los buscadores contienen al menos una referencia geográfica. Algunos ejemplos podrían ser consultas como “restaurantes cercanos al Museo del Prado”, “noticias sobre atentados en El Cairo”, “fotografías de las Torres Petronas”, etc. En estos ejemplos podemos ver cómo la necesidad de información del usuario viene restringida por un contexto geográfico específico.

Ahora bien, ¿cómo interpretan los buscadores web esa información geográfica? Un problema con el que se encuentran los sistemas de recuperación de información está relacionado con la gestión y manejo de las referencias geográficas presentes en los documentos y en las consultas del usuario. En ciertas ocasiones, el ser humano es capaz de interpretar frases como "cerca de Madrid" o "al sur de España", porque la mayoría de actividades ocurren en uno o más lugares de la superficie terrestre y disponemos de una gran cantidad de información que conocemos y que relaciona estos lugares. Sin embargo, esta interpretación no es del todo correcta en la mayoría de los motores de búsqueda actuales.

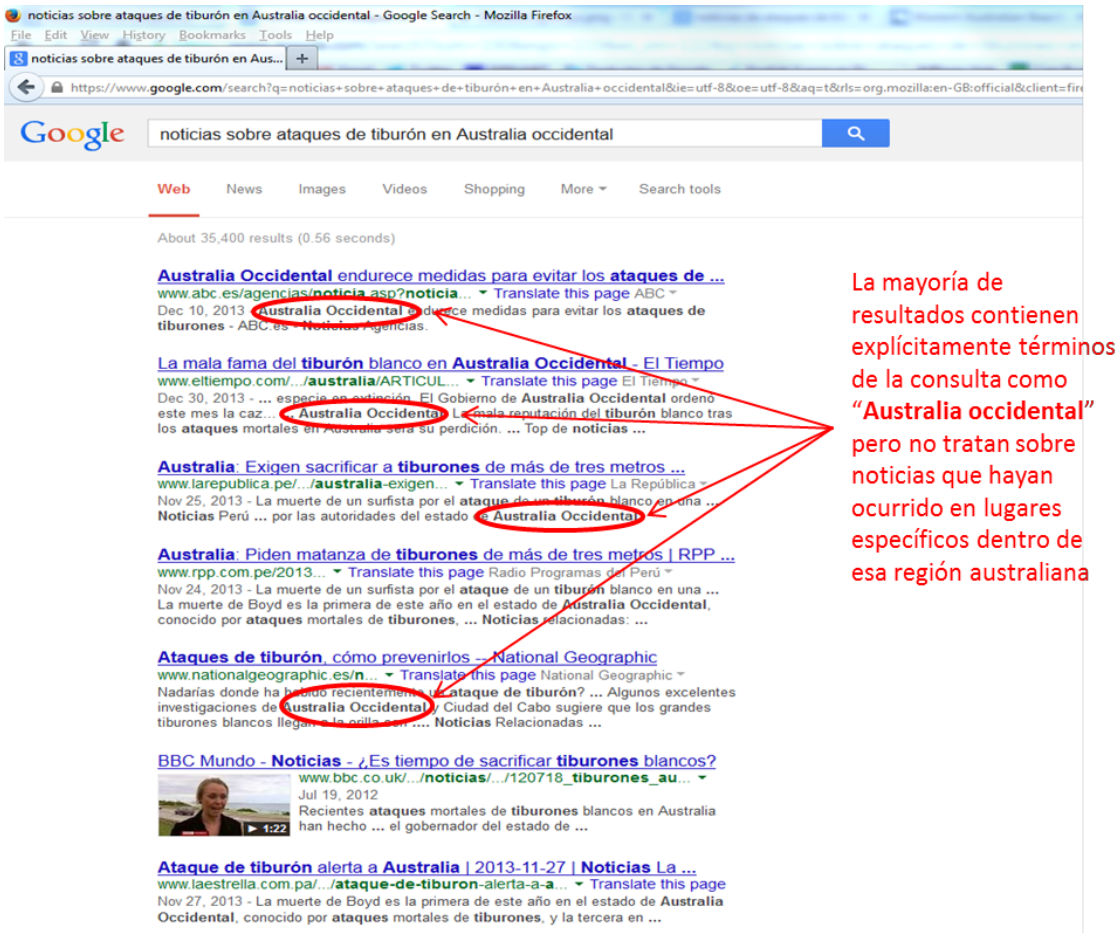
Esto se debe principalmente a varias razones:

- La primera es que la mayoría de buscadores aplican la teoría del emparejamiento simple (en inglés, keyword matching), que significa que cuando escribimos un nombre de lugar en la consulta, sólo los documentos que incluyen exactamente dicha localización son recuperados, ignorando la semántica de los términos geográficos utilizados. Por ello, es muy probable que muchos documentos relacionados con el ámbito geográfico de interés no sean recuperados.

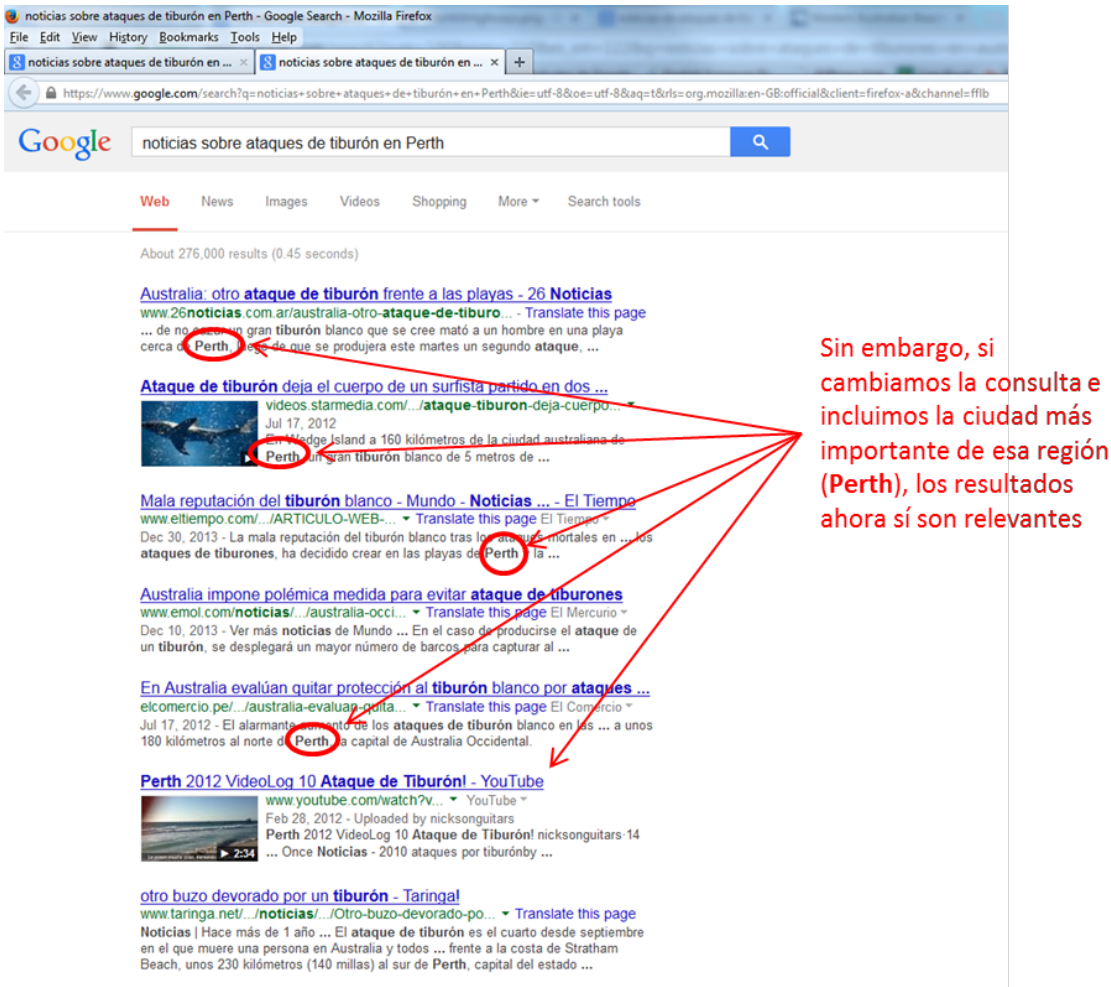
- Otro inconveniente que tienen los actuales motores de búsqueda tiene que ver con la utilización de relaciones espaciales relativas a la distancia o la inclusión por ejemplo, ya que a veces esta terminología puede llegar a ser compleja de interpretar hasta para el propio ser humano. Te pongo un ejemplo: si fueras un buscador y te hicieran la consulta "hoteles cerca de Madrid", ¿sería relevante devolver un documento que contenga un hotel situado en el pueblo de Coslada por ejemplo? ¿o en Getafe? ¿y en Toledo? Como se puede intuir, la interpretación geográfica y semántica de la relación espacial "cerca" es cuanto menos subjetiva, ya que estoy seguro de que dos personas se pondrían difícilmente de acuerdo en decidir qué se considera cerca de Madrid y qué no…

- Por último, otro de los grandes inconvenientes a los que se enfrentan los actuales motores de búsqueda es la ambigüedad presente en determinados contextos geográficos. Existen muchos nombres que son compartidos entre diferentes lugares, como por ejemplo Córdoba, Newport o Springfield. El término Córdoba puede hacer referencia a la capital andaluza, a la ciudad argentina de Córdoba, a un municipio de Colombia que también se llama así, a una ciudad del estado de Alabama en EEUU llamada también Córdoba, etc. Cuando leemos una referencia geográfica en un documento, tratamos de resolver la ambigüedad mediante el conocimiento extraído del contexto de dicho documento. Sin embargo, este comportamiento no es trivial para la mayoría de motores de búsqueda.

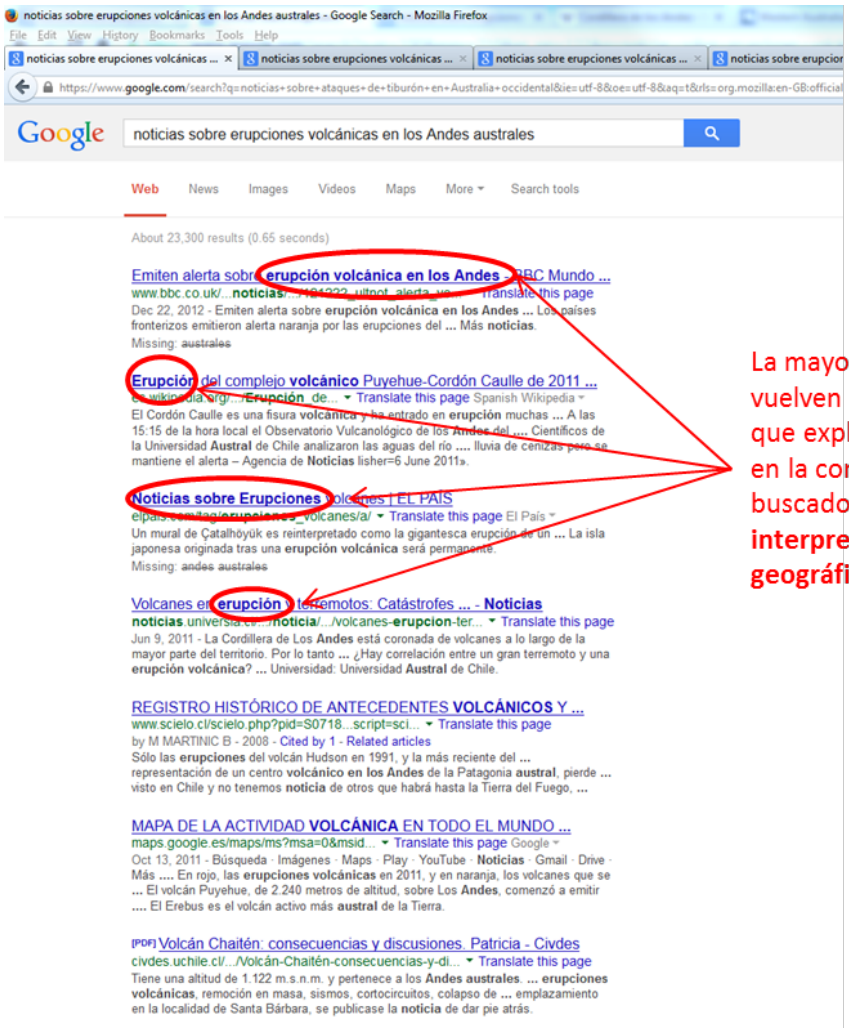
Para demostrar esta problemática a la que se enfrentan la mayoría de buscadores web, veamos a continuación un ejemplo del comportamiento de Google ante una consulta restringida por un contexto geográfico determinado. Por ejemplo, supongamos que estamos interesados en obtener noticias sobre ataques de tiburones en Australia occidental. Si lanzamos esa consulta en Google, podemos observar que la mayoría de resultados contienen explícitamente términos de la consulta como "Australia occidental", debido a la aplicación de la teoría del keyword matching comentada anteriormente. Como vemos en el ejemplo, ningún resultado muestra alguna noticia de ataque de tiburón que haya ocurrido en Perth, por ejemplo, una de las ciudades más importantes de esa región australiana.



Teniendo en cuenta el resultado anterior alguien podría pensar: "bueno, vale, eso es porque no ha habido ataques de tiburones en esa parte de Australia o en la ciudad de Perth". Sin embargo, nada más lejos de la realidad. Si modificamos la consulta anterior y simplemente sustituimos los términos "Australia occidental" por "Perth" y lanzamos la nueva consulta en Google, veremos que ahora los resultados sí son relevantes. Por tanto, podemos comprobar finalmente que sí existían noticias sobre ataques de tiburones en las costas de Australia occidental…

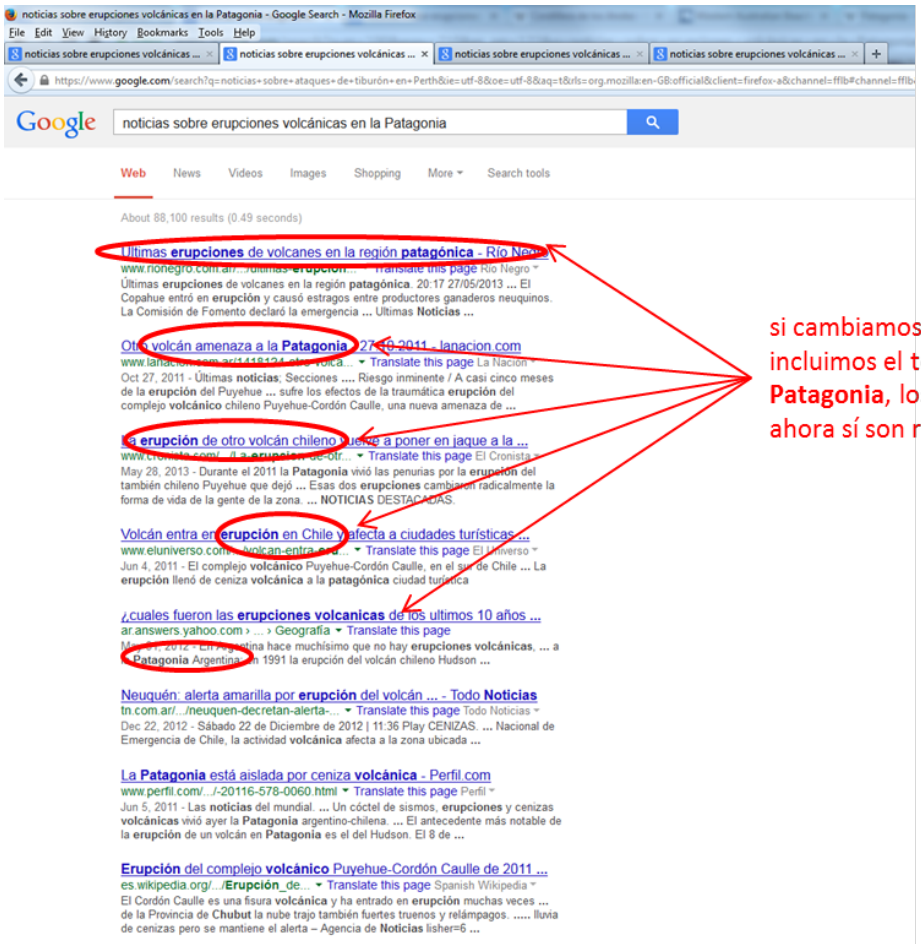


Veamos otro ejemplo más de este tipo de comportamiento. Si ahora estamos interesados en noticias sobre erupciones volcánicas en la región de los Andes australes que, según Wikipedia, comprende la zona sur de los Andes, abarcando parte de la Patagonia, Chile y Argentina, podemos comprobar que los resultados obtenidos no son los deseados. Esto es porque Google no es capaz de interpretar el contexto geográfico de los términos “Andes australes”



La mayoría de resultados vuelven a contener términos que explícitamente aparecen en la consulta, pero el buscador **no es capaz de interpretar el contexto geográfico "Andes australes"**

En cambio, si modificamos la consulta y sustituimos los términos "Andes australes" por "Patagonia" por ejemplo, los resultados ahora sí se pueden considerar relevantes.



Lo mismo ocurre si sustituimos el contexto geográfico de la consulta original por otros términos geográficos que pertenecen a esa región, como por ejemplo Chile o Argentina. Con esta serie de ejemplos que hemos visto, he pretendido demostrar que los motores de búsqueda todavía tienen mucho trabajo por hacer para conseguir una interpretación correcta cuando un usuario realiza determinadas consultas geográficas. Y es aquí donde entra en juego el área de la recuperación de información geográfica…