El impacto de la calidad de los metadatos en los servicios de búsqueda de una IDE

Tolosana-Calasanz, Rafael[1]
Nogueras-Iso, Javier[2]
Zarazaga-Soria, F. Javier[3]
University of Zaragoza (Spain) , rafaelt@unizar.es [1], jnog@unizar.es [2], javy@unizar.es [3]

ABSTRACT
A medida que se avanza en la investigación y en el desarrollo en el contexto de Información Geográfica y crece el volumen de metadatos creados, surgen nuevos requisitos en el rendimiento de los sistemas de recuperación de información geográfica. En este sentido sorprende detectar que los aspectos concernientes a la creación de metadatos con calidad razonable han recibido relativamente poca atención. En el mundo de la informática se suele utilizar el acrónimo GIGO (Garbage In, Gargage Out) para describir el problema derivado de utilizar como entrada de un sistema datos erróneos: la salida será inevitablemente inexacta o incorrecta; es decir, el uso de información de poca calidad limita el rendimiento de un sistema. Consecuentemente, es imprescindible que los sistemas de recuperación de información geográfica cuenten con metadatos de calidad para que se puedan obtener buenos resultados en general y, en particular, en los servicios de búsqueda.

Sin embargo, antes de estudiar concretamente el impacto de la calidad de los metadatos geográficos en los servicios de búsqueda, hay que analizar el concepto de calidad. La calidad es una cuestión subjetiva, influenciada por la complejidad de muchos factores humanos. Estos factores dependen de cada individuo y, además, los juicios personales suelen tener variabilidad a medida que pasa el tiempo y evolucionan las circunstancias de esas personas. La noción de calidad es un concepto que parece percibirse de forma inmediata y directa pero que a menudo es difícil de argumentar utilizando razonamientos lógicos. Debido a estas razones, la comunidad científica reconoce que la definición de la calidad de los metadatos no está exenta de dificultades. Un metadato de buena calidad se define a veces como un registro que es útil en un número de contextos diferentes, siendo útil también respecto a las estrategias de búsqueda y términos que se pueden emplear para localizarlo. Otras definiciones son menos ambiciosas y simplemente hablan de adecuación al propósito perseguido (“fitness for purpose”). Siguiendo estos razonamientos, parece que los metadatos geográficos serán adecuados para su propósito si describen bien los datos y esas descripciones son útiles para sus usuarios.

Uno de los problemas que los sistemas de recuperación de información geográfica tienen que resolver es proporcionar al usuario la información que satisface sus requisitos o sus intereses. Un usuario especifica una consulta y el sistema le devuelve una lista de resultados ordenados de manera que los que se han considerado más relevantes aparecen en las primeras posiciones de la lista. Los métodos de clasificación o “ranking” que utilizan los sistemas para hacer eso están basados en cuantificar la similitud entre la pregunta del usuario y el recurso en la colección. Esa cuantificación o valoración puede interpretarse como una estimación de la relevancia o utilidad de un recurso candidato a satisfacer las necesidades de un usuario. No obstante, la mayoría de las veces, para el usuario el hecho de especificar sin ambigüedad sus intereses es difícil o, lo que es peor, hay usuarios que no saben exactamente lo que están buscando. Dos grandes aspectos pueden considerarse para resolver el problema adecuadamente: en primer lugar, mejoras en las interfaces gráficas de usuario que faciliten tanto la exploración de la información como la especificación de las consultas y, en segundo lugar, mejoras en los algoritmos, métodos, estructuras de datos y modelos en el área de la recuperación de la información que se adapten a las particularidades especiales que presentan los metadatos geográficos.

Lo propuesto en este resumen está dentro del ámbito de este segundo aspecto. A grandes rasgos, las peculiaridades de los metadatos geográficos son dos principalmente: el componente conocido como espacial, atributo geométrico o extension espacial que describe la localización, la forma o la orientación; y el componente que describe los datos geográficos por medio de atributos no espaciales. La recuperación de información teniendo en cuenta el componente espacial se basa en asignar puntuaciones y elaborar “rankings” o clasificaciones a partir de las características geospaciales tales como el tamaño, la forma y la distancia. La recuperación de información atendiendo al componente no espacial puede basarse en técnicas tradicionales de recuperación de información que se apoya en propiedades estadísticas que se aparecen en los términos de los metadatos. Existe una enorme y extraordinaria variedad de técnicas al respecto que puede encontrarse en la literatura para resolver el problema, aunque la naturaleza de la información geográfica debe tenerse en cuenta.

Cuando se combinan ambas estrategias de “ranking” se pueden obtener buenas listas de resultados; aunque, en determinadas circunstancias, podrían no ser suficientemente satisfactorias. Por ejemplo, los catálogos de geodatos están progresivamente recibiendo metadatos de diferentes organismos, metadatos que fueron creados de forma heterogénea por equipos de trabajo distintos y con criterios diferentes, metadatos que describirán con mayor o menor fortuna datos geográficos. En estos casos, los usuarios esperan recibir una lista ordenada de manera que las descripciones que mejor le convengan aparezcan al principio. Está claro, que aquellas descripciones más ricas y mejor realizadas, esto es, aquellos metadatos de mejor calidad deberían aparecer antes que los de calidad más baja. En estudios anteriores, hemos tratado de estimar la calidad de los metadatos geográficos. Pretendemos, a partir de ahora, definir nuevos algoritmos de “ranking” que incorporen esa estimación y podamos, de esa manera, recuperar información geográfica con mayor grado de satisfacción.