

Etiquetado social y blog-scraping como alternativa para la actualización de vocabularios controlados: aplicación práctica a un tesoro de Biblioteconomía y Documentación

Gonzalo Mochón Bezares, Eva Méndez Rodríguez y Ángela Sorli Rojo

Información, cultura y sociedad, ISSN 1851-1740, n. 37, 2017, p. 13-26

La actualización de los tesauros y otros sistemas de organización del conocimiento (SOC) es una tarea que presenta problemas en el mantenimiento que debe seguir a toda creación de vocabularios. Uno de estos problemas es que el alcance y uso de algunos descriptores o conceptos cambia con el paso del tiempo, a la vez que se incorporan términos nuevos y otros caen en desuso. Hay diferentes formas de actualizar los vocabularios controlados, principalmente sobre tesauros y ontologías. Se distinguen métodos automáticos o semiautomáticos. El objetivo de este trabajo es doble: se investiga la representación de conceptos en las etiquetas en lenguaje libre y en el lenguaje controlado no estructurado de una base de datos; y se busca conocer la terminología más novedosa para cada una de las categorías que integran el Tesoro de Biblioteconomía y Documentación (TByD). La metodología empleada combina la extracción terminológica de las listas de términos procedentes de la indización en la base de datos ISOC-BD, y la extracción de términos a partir de blogs, utilizando tecnologías de web scraping y text mining, lo que los autores denominan blog scraping. De la extracción de la base de datos ISOC-BD se recogieron 130 términos. Del análisis de 127 blogs se recogieron 186 etiquetas. Este estudio implica una nueva aportación a la tradicional dicotomía entre el lenguaje libre y los lenguajes controlados. Los resultados obtenidos en la comparación realizada demuestran que las etiquetas asignadas por autores al contenido de los blogs aportan una terminología más variada y actualizada que las palabras clave asignadas a artículos científicos por documentalistas. El mayor incremento terminológico se ha dado en las categorías relacionadas con la web y su desarrollo tecnológico.

Resumen elaborado por Antonio Rodríguez Vela