Método para la extracción masiva de canales de sindicación

Manuel Blázquez Ochando

Scire, ISSN 1135-3716, Vol. 23, n. 1, 2017, p. 39-45

La redifusión de información, sinónimo de sindicación de contenidos, corresponde a una tecnología clave para soportar diversas actividades de la información y documentación. La sindicación se define como la transmisión de activos informativos y documentales, para su reutilización e integración en terceros recursos, a través de un archivo editado en lenguaje de marcado extensible XML, contenedor de la información, conforme a un formato de estructuración de datos RSS o Atom. Se utiliza para redifusión de registros documentales y autoridades, el intercambio de contenidos entre bibliotecas digitales, la alimentación de grandes proyectos europeos como Europeana y la difusión selectiva de la información. Existen 348 millones de instalaciones de programas CMS registradas, por lo que el principal problema es el descubrimiento de los canales de sindicación pertinentes. Este estudio propone combina programas web crawler y estrategias de búsqueda para dirigir el objetivo de recopilación. El método propuesto, consta de los siguientes pasos: 1) delimitación del área de conocimiento y desarrollo de un vocabulario representativo, 2) diseño de estrategias de consulta para la recuperación de sitios web pertinentes, 3) creación de una semilla de enlaces para su análisis con herramientas web crawler, 4) análisis de enlaces con programas web crawler, y 5) preparación previa de los canales de sindicación para su procesamiento en agregadores de contenidos. A continuación el autor pasa a detallar cada uno de estos procedimientos. El desarrollo de estudios

informétricos, de producción de información periodística, de opinión, tendencias o incluso de producción científica, según el objeto de estudio, depende de una mayor exhaustividad en las fuentes de información utilizadas. Esto significa no restringir las investigaciones a las fuentes conocidas y abrir el campo de estudio a nuevas fuentes que están por descubrir. Tomando como referencia la terminología organizada, se proponen estrategias de consulta que la combinan usando operadores de consulta avanzada en buscadores para obtener resultados más pertinentes posibles. De los resultados obtenidos se extraen los enlaces que serán procesados por herramientas web crawler para extraer los enlaces de los canales de sindicación. Finalmente se detectan errores en los canales de sindicación y se completa la información clasificatoria y descriptiva que caracteriza su contenido.

Resumen elaborado por Antonio Rodríguez Vela