Nuevos retos de la tecnología Web Crawler para la recuperación de la información

Manuel Blázquez Ochando

Métodos de Información, ISSN 1134-2838, Vol. 4, n. 7, 2013, p. 115-128

Los programas web crawler constituyen una parte importante como elementos que ayudan en la recuperación de información de la cadena documental, de ahí su constante actualización y perfeccionamiento. Este artículo define los principales enfoques con los que se diseñan éstos programas, y aborda los futuros desafíos a los que deben dar respuesta. Se describe recuperación de información y su objetivo de proporcionar los documentos que mejor respondan a las necesidades informativas y documentales de los usuarios. Proceso que es llevado a cabo por sistemas conocidos como web crawler, cuya misión principal es rastrear páginas web a través de sus enlaces. Se exponen qué especificaciones les caracterizan, las cuales tienen como objetivo resolver los problemas más comunes a los que se enfrentan estos programas, entre las que se destacan: capacidad de indexar total o parcialmente páginas web; extracción de recursos y su archivado para preservación; capacidad de realización de estudios webmétricos. Se analiza el programa Mbot, web crawler capaz de realizar análisis a pequeña escala de la Web, y se compara con cuatro importantes rastreadores de código abierto: Apache Nutch, Heritrix, WIRE, y SocSciBot. Los resultados de esta comparativa quedan expuestos en tablas donde se pueden observar un conjunto amplio de las diversas características técnicas de estos

programas, así como los tipos de análisis e informes automáticos que pueden generar, destacándose el de análisis webmétrico. Se abordan aspectos esenciales como las dificultades que alguno de ellos presentan para su correcta instalación y configuración para realizar una misma tarea, así como recomendaciones para mejorar la configuración de futuros programas de recuperación, entre ellas, lograr la máxima sencillez de instalación, configuración y puesta en marcha, adaptación a las necesidades del documentalista, o la incorporación de aquellos elementos necesarios para funcionar de forma completa e integral. Se completa la exposición con la descripción de un conjunto de aspectos que el autor considera deben mejorarse y aplicarse a los futuros programas web crawler, entre otros, la simplificación de su uso, polivalencia, y capacidad de reconocimiento y de difusión. El futuro de estos sistemas pasa por convertirse en herramientas multipropósito, capaces de adaptarse a las necesidades de cada usuario (sencillez de uso), y especialmente orientados a la recuperación de contenidos semánticos – propios de la Web semántica – más que a los actuales contenidos hipertextuales y documentales.

Resumen realizado por la Sección de Documentación Bibliotecaria