Evaluación de la calidad de los Sistemas de Recuperación de Información

Paúl Rodríguez Leyva[1] y Juan Pedro Febles Rodríguez[2]

[1,2]Profesor; Universidad de las Ciencias Informáticas; La Habana, Cuba.

[1]pleyva@uci.cu

[2]jfebles808@gmail.com

Recibido: 20 ene. 2018 Aceptado: 17 abr. 2018

La recuperación de información se enfoca en el procesamiento de colecciones de documentos con el objetivo de brindar resultados que satisfagan interrogantes de búsqueda. Los documentos pueden ser de varios tipos (html, imágenes, videos, texto plano, entre otros) y las herramientas más utilizadas internacionalmente para ejecutar este proceso son los Sistemas de Recuperación de Información o buscadores, como comúnmente se les conoce. Estos sistemas basan su funcionamiento en tres procesos fundamentales:

- El rastreo: proceso mediante el cual se recorre la web con el objetivo de extraer información que posteriormente es almacenada. Los buscadores ejecutan esta actividad basados en arañas o rastreadores, que son los encargados del parseo de la estructura HTML de los sitios.

- La indexación: proceso encargado del procesamiento y almacenamiento de la información rastreada. Esta actividad contempla la transformación de la información en estructuras que sirven de base para ejecutar los modelos matemáticos encargados de establecer que documentos son relevantes en relación a determinada consulta.

- La visualización de la información: proceso encargado de brindar las interfaces donde se le muestra a los usuarios las opciones de inserción de consultas, ya sean simples o avanzadas, y se visualiza los resultados que el buscador extrae como relevantes de la colección de todos los documentos almacenados.

La calidad de un Sistema de Recuperación de Información está determinada por factores como la capacidad de almacenamiento, velocidad de las respuestas, grado de actualización y el nivel en que son satisfechas las interrogantes de búsqueda de los usuarios. Esta última característica hace que el proceso de evaluación de los buscadores sea muy complejo debido a que la relevancia de un resultado es definida de forma distinta por los usuarios. Las métricas más abordadas en la literatura para establecer criterios de calidad sobre los resultados de un buscador, son la precisión y la exhaustividad.

El valor de la primera se calcula dividiendo la cantidad de documentos relevantes que son recuperados entre los documentos que son relevantes y la segunda, dividiendo la cantidad de documentos relevantes recuperados entre los documentos recuperados. En ambos casos existe una dependencia total del factor relevancia, lo que implica que estos valores dependan de criterios subjetivos que establece el usuario sobre lo que podría ser o no un resultado relevante. Para una misma consulta, dos usuarios distintos pueden seleccionar de los resultados devueltos por el buscador, grupos distintos de resultados relevantes; ocasionando que la conformación de grupos de usuarios homogéneos para ejecutar el proceso de evaluación sea determinante para establecer criterios acertados sobre la calidad del Sistema de Recuperación de Información objetivo de la evaluación.