Categorización jerárquica de gran contenido usando topología de concepto

Andrew Yates, y otros

Journal of library metadata, ISSN 1937-5034, Vol. 18, n. 3-4, 2018, p. 113-134

Se necesitan métodos que sean computacionalmente factibles y prácticamente efectivos para dar sentido a los grandes corpuses de contenido, o «gran contenido». Por ejemplo, las técnicas de categorización supervisadas para publicaciones académicas de acceso abierto no son adecuadas para la categorización automatizada porque se basan en un esquema ya existente, pero ningún esquema supervisado puede mantenerse al tanto del panorama en la rápida evolución del trabajo académico. Este problema también se aplica a cualquier dominio con grandes cantidades de documentos donde no exista un buen esquema de categorización. Para enfrentar este desafío, presentamos un método no supervisado para ajustar un esquema de categorización jerárquico a un corpus basado en agrupar la red de conceptos compartidos en el corpus, o su «topología de conceptos». Nuestro método se aplica potencialmente a cualquier tipo de contenido, y Se escala a grandes redes de millones de vértices. Hemos demostrado la aplicación de nuestro método a un corpus de 1,5 millones de textos académicos que representan a la mayoría de las publicaciones académicas de acceso abierto (Open Access) en la web, validando nuestros resultados utilizando anotaciones de bibliotecarios expertos. Hemos hecho nuestros conjuntos de datos de acceso abierto para la investigación de otros. Creemos que nuestro esquema de categorización resultante

representa mejor la publicación académica OA tal como existe en la actualidad.

Traducido del resumen de la propia publicación