Improving Collection Understanding in Web Archives

Shawn M. Jones

Bulletin of the IEEE Technical Committee on Digital Libraries, ISSN 1937-7266, vol. 15, n. 1, 2019

Los archivos web sirven para que los periodistas documenten sus artículos, los historiadores almacenen información y los científicos sociales estudien periodos específicos de tiempo. Los usuarios no pueden conocer el contenido de las colecciones sin revisarlas manualmente. A lo largo del tiempo cambian los formatos de los mismos documentos, lo que sirva para saber cómo los recursos han ido evolucionando, pero dificulta su usabilidad. El objetivo del autor es ayudar a los creadores de colecciones y al público para hacer un mejor uso de las mismas a través de la más profunda comprensión de la colección. Archive-IT fue creada por Internet Archive en 2005, y permite a los conservadores elegir qué contenido preservar en el formato original. Para elegir el contenido más apropiado a lo que se busca, los metadatos utilizados no son suficientes, ya que se usan estándares inconsistentes y diversas interpretaciones. Para ayudar a los investigadores, se propone crear sumarios de colecciones automáticos, extrayendo los conceptos más relevantes. Los usuarios pueden apoyarse en pistas visuales y textuales para determinar si un recurso se ajusta a sus necesidades. Para realizar el estudio se seleccionaron mementos representativos (agregaciones de archivos web), y se visualizaron. Para ello se agrupó los mementos de la colección según sus concordancias, y después se seleccionaron los de mayor calidad. En cuanto a la visualización, se pueden utilizar miniaturas bajo encabezamientos que contengan las entidades dentro de las mismas. El estudio ha permitido identificar diferentes categorías semánticas de las colecciones de archivos web en Archive-It, que pueden ser predecibles. Con el desarrollo y evaluación de nuevos algoritmos y visualizaciones para la comprensión de las colecciones, será posible que se obtengan sumarios usando diferentes técnicas de reducción y conocidos paradigmas de visualización.

Resumen elaborado por Antonio Rodríguez Vela