

Uso de R en bibliometría. Exploración de técnicas para la detección de temas locales de investigación

Claudia M. González^{1,3}, Sebastián Varela^{2,3}, Sandra E. Miguel^{1,3}

¹ Universidad Nacional de La Plata. Facultad de Humanidades y Cs de la Educación. Dpto de Bibliotecología, Argentina
cgonzalez@fahce.unlp.edu.ar

² Universidad Nacional de La Plata. Facultad de Humanidades y Cs de la Educación. Dpto de Sociología, Argentina
svarela@fahce.unlp.edu.ar

³ IdIHCS - Instituto de Investigaciones en Humanidades y Cs Sociales (CONICET/UNLP), Argentina
smiguel@fahce.unlp.edu.ar

Palabras Claves: Cienciometría - Clustering - Modelado de tópicos - Latent Dirichlet Allocation (LDA)

1 Introducción

La investigación de aspectos de la ciencia a partir de las publicaciones puede hacerse desde una perspectiva bibliométrica, que implica el análisis estadístico sobre la información descriptiva, de contribuciones y de citación de las producciones científicas de carácter bibliográfico. O también desde una perspectiva temática, que permite detectar los frentes de investigación a partir de elicitar los tópicos allí tratados de manera que se pueda obtener información resumida útil para establecer relaciones analíticas. Esta identificación de temas puede hacerse utilizando el método de análisis de contenido tradicional, generalmente usado en estudios cualitativos sobre poca cantidad de documentos, o utilizando modelado estadístico, lo que permite extraer y analizar información de grandes colecciones de texto, aplicando algoritmos no supervisados. Diversas técnicas se han utilizado para la realización de representaciones temáticas de corpus textuales. El análisis de co-palabras es la técnica básica, la cual ha sido trabajada complementariamente con técnicas estadísticas de reducción del volumen de datos sin mayores pérdidas de información, tales como el análisis de *clusters*, el escalamiento multidimensional (MDS), el análisis factorial (FA), el *blockmodeling*, el análisis de redes sociales (ARS) y dentro de esta, técnicas específicas como las redes *Pathfinder*. En el último tiempo, la técnica *Latent Dirichlet Allocation* (LDA) ha sido una de las más usadas (Blei et al, 2003), y puede utilizarse además para mostrar la evolución de tópicos y asociaciones entre temas y autores. Es una forma de minería textual que funciona identificando patrones de co-ocurrencia de palabras y produciendo agrupamientos por similitud que denominamos tópicos. Cuando las técnicas basadas en co-ocurrencias de palabras se aplican sobre las bases de datos bibliográficas comprensivas, explotando los títulos y los resúmenes de los documentos como corpus textual, se las toma como técnicas bibliométricas.

2 Materiales y Métodos

En este trabajo se muestran resultados preliminares obtenidos al aplicar la técnica de *clustering* basado en *k-means* y un modelado de tópicos usando *Latent Dirichlet*

