Topic Modelling in Social Networks with Formal Concept Analysis

P. Cordero, M. Enciso, D. López-Rodríguez, A. Mora
Universidad de Málaga, Andalucía Tech, Málaga, Spain
pcordero@uma.es, enciso@uma.es,dominlopez@uma.es, amora@uma.es

Abstract

In the age of social networks, the amount of the written material published every day exceeds our processing capacity. Topic models can help to organise and to understand extensive collections of unstructured text documents. In machine learning and natural language processing, a topic model is a statistical model for discovering the abstract "topics" in a collection of documents, uncovering hidden semantic structures and clusters of similar words.

Topic modelling formalises this idea mathematically in a framework that allows discovering the topics and each document's balance of topics. One of the first topic models [1] was *latent semantic indexing*. Later, the probabilistic latent semantic analysis (PLSA) was presented [2], serving as a basis for many others. Notably, its extension *latent Dirichlet allocation* (LDA) [3] is one of the most common topic model currently in use.

To approach topic modelling in social networks, we use Formal Concept Analysis [4], a mathematical tool firmly based on lattice theory and logic. Our approach uses the knowledge contained in the concept lattice to extract the topics. Thus, this approach to topic modelling is not statistical. For example, we do not need to assume a prior distribution of terms. Instead, the actual data structure is used to infer the semantic relationships between attributes.

The procedure is as follows: a formal context is built from the document-term matrix of the set of documents. Then, we use FCA tools to construct the concept lattice that contains, in each concept, knowledge about the topics in the documents. Once this lattice is built, the concepts are clustered. Concept clusters are then used to induce topic models on the original documents.

An experiment with a dataset with tweets about some hashtags is conducted with our approach to show how Formal Concept Analysis can be used in Social Network Analysis. In addition, a comparison with classical techniques is being addressed.

Acknowledgment

This work has been partially supported by the projects TIN2017-89023-P, PGC2018-095869-B-I00 of the Science and Innovation Ministry of Spain, and UMA2018-FEDERJA-001, funded by the European Regional Development Fund (ERDF).

References

[1] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[2] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] Bernhard Ganter and Rudolf Wille. Formal concept analysis. *Wissenschaftliche Zeitschrift-Technischen Universitat Dresden*, 45:8–13, 1996.