Term-based topic extraction incorporating word embedding techniques: A comparative study

Yi Zhang[1], Jianjiu Ou[1], Zhinan Wang[2]

[1]Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia
[2]School of Management and Economics, Beijing Institute of Technology, China

Extended Abstract

Topic extraction, known as the way of modelling data, extracting features, detecting regularities, and identifying topics for representing the data (Velden et al., 2017), is a key concern of the bibliometric community for decades. As a part of feature extraction, how to effectively exploit bibliometric indicators (e.g., citation/co-citation counts, word co-occurrence statistics, and co-authorship information) has been well investigated in the literature (Boyack et al., 2011; Klavans & Boyack, 2017; Zhang et al., 2018). It is clear that either citation analysis or co-word analysis has its own strengths and weaknesses on topic extraction and further analysis such as science mapping (Zhang et al., 2017), e.g., distinguishing positive and negative citation linkages, and synthesizing technical synonyms.

The rise of word embedding techniques, as an application of shallow neural networks that map words from a vocabulary to vectors of real numbers and uncover latent semantics in large-scale texts (Mikolov et al., 2013), sheds light on the area of natural language processing (NLP), as well as text analytics. Given the circumstances, co-word statistics-based topic extraction, in which topics are usually represented by either a set of terms or a group of records, is attracting increasing interests from the bibliometric community. One pilot study that exploited word embedding techniques for representing records has been approved that such incorporation provides great benefits for further record-based topic extraction (Zhang et al., 2018). However, comparably, record-based topic extraction might emphasize the use of time line associated with records in the investigation of topic evolutions (Zhang et al., 2016), while term-based topic extraction provides a solution of profiling detailed topics (Carley et al., 2018). One critical comment for record-based topic extraction is that in a broad range of scientific disciplines (in particular business disciplines) one record would cover more than one topics, and thus, soft clustering (e.g., fuzzy c-means) sometimes is introduced to get rid of such negative influences (Zhu et al., 2018), even though such involvement might be case-sensitive (Zhang et al., 2018).

This study aims to develop a term-based method for topic extraction with the incorporation of word embedding techniques. In this paper, we mainly focus on the incorporation of word embedding for term representation, the design of validation measurements for term-based topic extraction on bibliometric data, and the comparison of popular clustering algorithms on given tasks of topic extraction. The research framework is designed in Figure 1.

Titles and abstracts of bibliometric data (e.g., academic papers) are the inputs of this study. The Word2Vec method (Mikolov et al., 2013) is applied as a representative approach of word embedding techniques and its different models (e.g., a continuous bag of words model and a skip-gram model) are compared, and, then, a strategy is to be proposed for term representation – i.e., how to generate term vectors using word vectors. A model of validation measurements is designed, combining with quantitative and qualitative approaches. On the one hand, a labelled dataset is constructed based on specific data sources, e.g., the subject category of the Web of Science (WoS), and the international patent classification code of the Derwent World Patent Index (DWPI), and indicators such as Herfindahl index (Klavans & Boyack, 2017) and counting pair-based clustering evaluation metrics (Xuan et al., 2018) are applied for evaluating

the performance of clustering solutions. On the other hand, expert knowledge is involved for empirical validation, which is based on the domain of case studies. Popular clustering algorithms, including K-means, topic models, principal component analysis, and hierarchical clustering, are then integrated with term vectors and their abilities in topic extraction are examined and compared. In particular, as an endeavor of topic extraction for profiling technical emergence, the Emerging Score indicator (Carley et al., 2018) and its associated emerging topics (Wang et al., 2018) have been recognized by the community, and thus, it becomes interesting to compare with them as well.

This study would provide experimental results for the development of novel clustering methods incorporating word embedding techniques, and create solutions of accurate topic extractions for the bibliometric community.

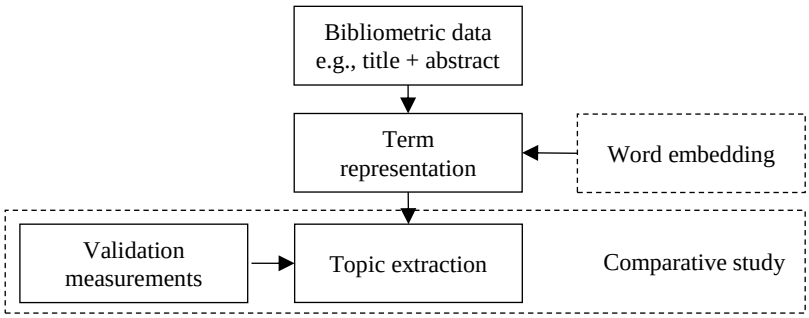

Figure 1. Research framework of the comparative study for term-based topic extraction incorporating word embedding techniques

References

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., . . . Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS One, 6*(3), e18029.

Carley, S. F., Newman, N. C., Porter, A. L., & Garner, J. G. J. S. (2018). An indicator of technical emergence. *115*(1), 35-49.

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology, 68*(4), 984-998.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111-3119.

Velden, T., Boyack, K. W., Gläser, J., Koopman, R., Scharnhorst, A., & Wang, S. (2017). Comparison of topic extraction approaches and their results. *Scientometrics, 111*(2), 1169-1221. doi:10.1007/s11192-017-2306-1

Wang, Z., Porter, A. L., Wang, X., & Carley, S. (2018). An approach to identify emergent topics of technological convergence: A case study for 3D printing. *Technological Forecasting and Social Change*. doi:https://doi.org/10.1016/j.techfore.2018.12.015

Xuan, J., Lu, J., Zhang, G., Da Xu, R. Y., & Luo, X. (2018). Doubly nonparametric sparse nonnegative matrix factorization based on dependent indian buffet processes. *IEEE transactions on neural networks and learning systems, 29*(5), 1835-1849. doi:10.1109/TNNLS.2017.2676817

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., & Zhang, G. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics, 12*(4), 1099-1117.

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology and a case study focusing on big data research. *Technological Forecasting and Social Change, 105*, 179-191.

Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017). Science evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science and Technology, 68*(8), 1925-1939.

Zhu, J., Han, L., Gou, Z., Yuan, X. J. J. o. t. A. f. I. S., & Technology. (2018). A fuzzy clustering - based denoising model for evaluating uncertainty in collaborative filtering recommender systems. *69*(9), 1109-1121.