Keyword Distance Ratio: Evaluating Keyword Assignment with Word Embeddings

Brandon Sepulvado[1]

[1] *sepulvado-brandon@norc.org*
NORC at the University of Chicago, 4350 East West Highway, Bethesda, MD 20814 (USA)

Introduction

Despite the emergence of natural language processing (NLP) assisted search within certain new bibliometric databases (Wang et al., 2020), keywords remain a staple of information curation and retrieval within most major bibliometric sources, such as Web of Science, Scopus, and PubMed, and authors still choose keywords or phrases to summarize the information their documents contain. As such, keywords are an institutionalized component of the publication and literature search process, yet how does one empirically evaluate the keywords chosen for a document?

This text proposes the Keyword Distance Ratio (KDR) to respond to this question. Theoretically, keywords should, at the same time, be closely related to the content within a document while also remaining sufficiently distinct so as not to convey redundant information. The KDR is a document-level measure that relies upon the Relaxed Word Mover's Distance (RWMD; Kusner, Sun, Kolkin, & Weinberger, 2015) to quantify these intuitions.

In investigating the utility of the KDR, this text compares article keywords submitted by authors to those chosen by Scopus in its index terms and then illustrates a dramatic difference in document summary based upon the keyword source.

Keyword Distance Ratio

Relaxed Word Mover's Distance

The KDR relies upon Word Mover's Distance (WMD; Kusner et al., 2015) in order to obtain distances between keywords and the documents they describe. For an intuitive understanding of WMD, imagine that a document is a cluster of points in *n*-dimensional space. The points represent the embeddings for words (i.e., keywords and those in the article), and a word's location is based upon the word's embedding vector. The current analyses use 300-dimension FastText embeddings (Mikolov, Chen, Corrado, & Dean, 2013), which means that each word would be placed in a 300-dimension space. One plots the keywords and words in the document, and the distance then is a function of the effort it takes to move one set of points to the closest points in the other set. To reduce computational complexity, the RWMD relaxes a couple constraints imposed upon the WMD.

Keyword Distance Ratio

The KDR entails calculating two sets of RWMDs: (1) between each pair of keywords and (2) between each keyword and its corresponding document. The KDR then compares the sum of each type of distance, adjusting for the number of keywords listed. Let the KDR of document *d* be:

$$KDR_d = \frac{\sum_{j=1}^{n-1}\sum_{i>j} RWMD_{k_{i,d},k_{j,d}}}{\sum_{i=1}^{n} RWMD_{k_{i,d},a_d}} \cdot \frac{2}{(n-1)}$$

In this equation, the numerator sums all pairwise RWMDs between keywords k_i and k_j used to signify the content of scholarly document d; the denominator sums the pairwise RWMDs between keyword k_i and abstract a for the document d. This value however will increase automatically with each additional keyword because there are $\frac{n(n-1)}{2}$ possible comparisons in the numerator and only n comparisons in the denominator, where n is the number of keywords. As such, the second part of the KDR equation accounts for the number of keywords. To rephrase more intuitively, the KDR is the ratio of the sum of pairwise RWMDs for all keywords listed for a document to the sum of all the RWMDs between a document's keywords and its abstract; this value is then scaled to account for the number of keywords in a document.

Data

This text uses a small corpus of neuroethics articles to demonstrate the utility of the KDR for comparing author-provided keywords to Scopus-assigned keywords (i.e., index terms). The data used come from Scopus, which is an ideal bibliographic database because it has good journal coverage in both the humanities and health-related sciences and indexes more journals than other databases, such as the Web of Science (Falagas, Pitsouni, Malietzis, & Pappas, 2008). A keyword-based query was used to obtain publication records from Scopus because this approach has proven successful for neuroethics (Leefmann, Levallois, & Hildt, 2016). I searched Scopus for any articles published in English that contain "neuroethic*" in the title, abstract, and/or keyword fields. After excluding articles with missing abstracts and/or keywords, there are 727 publications. In calculating the distance from

keywords to documents, this paper uses abstracts—rather than full article text—to represent a document.

Results

Table 1 presents the descriptive statistics on both sets of keywords. It is notable that, although the numerator and denominator are higher in the index terms KDR than in the author keywords KDR, the numerator is proportionally much higher in the index terms KDR. The mean and median KDRs are much lower for author keywords. Note that values below 1 indicate that a document's cumulative keyword-to-abstract distance is greater than the document's keyword-to-keyword cumulative RWMD; values greater than 1 indicate that a document's cumulative keyword-to-abstract distance is less than the document's cumulative keyword-to-keyword RWMD. In other words, KDR < 1 indicates that keywords are more similar to each other than to the abstract, and KDR > 1 indicates that keywords are less similar to each other than to the abstract. Table 1 suggests that authors tend to choose keywords that are more similar to each other while Scopus assigns index terms that maximize their differences.

Table 1. Descriptive statistics about the KDR and its main components for both author keywords and Scopus index terms

Value	Mean	Med.	SD	Min.	Max.
Author Keywords					
Numerator	20.3	13.5	21.4	1.06	184
Denominator	22.6	23.1	23.1	0.00	200
KDR	.881	.896	.0997	.381	1.16
Index Terms					
Numerator	406	243	562	.456	8107
Denominator	71.7	58.0	57.4	0.00	549
KDR	4.82	3.62	5.74	.0822	106

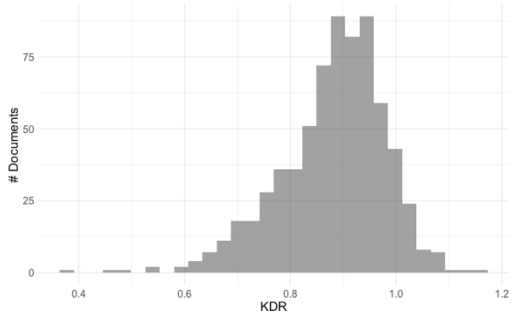

Figure 1. KDR distribution based upon author keywords

Aside from these descriptive statistics, it is instructive to visualize the distribution of these KDRs calculated from different types of keyword. Figure 1 presents the KDR distribution when calculated based upon author keywords, and Figure 2 presents the KDR distribution for index terms. The difference in the *x*-axes was to be expected from Table 1. However, the difference in distribution shape is striking. The author keyword KDR distribution is somewhat normally distributed around 1, while the index term KDR distribution is dramatically skewed to the right.

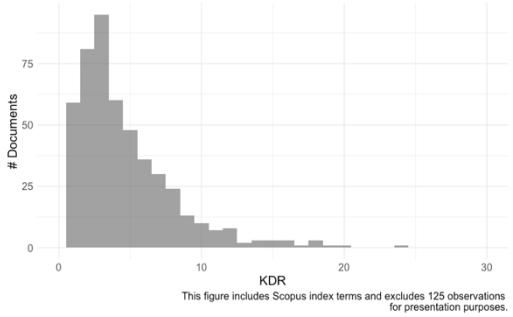

Figure 2. KDR distribution based upon Scopus index terms

Ongoing analyses (not included) examine the impact of potential limitations by increasing sample size and looking at less interdisciplinary fields/topics as well as other data sources (e.g., Web of Science). Future research should investigate specific mechanisms linking keyword selection practices to KDR distributions, how KDR differences are associated with perception of keyword accuracy and utility, alternative embeddings for calculating distances, and the use of full article text rather than abstracts.

Acknowledgments

This material is based upon work supported by the U.S. National Science Foundation under Grant Number 1939965.

References

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB Journal*, 22(2), 338–342.

Kusner, M. J., Sun, Y., Kolkin, N. I., & Weinberger, K. Q. (2015). From word embeddings to document distances. *32nd International Conference on Machine Learning*, 37, 957–966.

Leefmann, J., Levallois, C., & Hildt, E. (2016). Neuroethics 1995–2012. A Bibliometric Analysis of the Guiding Themes of an Emerging Research Field. *Frontiers in Human Neuroscience*, 10(336), 1–19.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1-12.

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies*, 1(1), 396–413.