Do You Cite What I Mean? Assessing the Semantic Scope of Citations

Sainte-Marie, Maxime Mongeon, Philippe Larivière, Vincent

One of the most common and robust assumptions in bibliometrics is that citations are semantics-laden: regardless of motive, context or even content, when a document cites another one, meaning is both drawn from and conferred to the documents involved. In fact, it is this semantic kinship between citing and cited documents that would have made citation indexing both relevant and useful for information retrieval purposes in the first place [1].

Despite this critical role, very few studies have looked at the semantic dimension of citation relationships. In a most significant but already dated study [2], Harter *et al.* investigated the semantic relationship between keywords of citing and cited article pairs in various articles published between 1988 and 1989 in three information science journals (*College and Research Librairies, JASIS, Library Journal*), based on the descriptors assigned by three indexing and abstracting journals (*ERIC, LISA, Library Literature*). By calculating the Jaccard similarity coefficient ($Sim(A, B) = (A \cap B)/(A \cup B)$) between the keyword sets of citing and cited document pairs, the author obtained surprisingly low averages and high standard deviations, with minima and maxima of .08 and .39 in the first case and .08 and .350 in the second. In light of these highly variable but globally low results, Harter concluded that "the subject similarity among pairs of cited and citing documents is typically very small" [2, p.543].

However, the scope of that study is limited in several regards. First, the number of journals and citing/cited article pairs investigated is considerably small. But more importantly, reducing article semantics to the sole presence or absence of keywords in the database is rather oversimplifying. The purpose of the present research is to assess the semantic scope of citations using word-based as well as 3- and 4-gram-based weighted vector space models on a wider set of articles and article text fields.

To do this, relevant text data (Title, Abstract, Author Keywords, ISI Keywords, References) from all 15,461 Web of Science articles published in 2015 and classified as Economics publications by the NSF field classification of journals were extracted. Then, for each article, corresponding text data from each cited document was joined to the citing article information, and all citing/cited article pairs missing one or more of the analyzed fields were removed, thus reducing the size of the dataset to 126,202 citing/cited pairs, involving 10,801 citing Economics articles and 77,684 cited articles.

Following the method used in [2], Jaccard similarity scores between citing and cited article keyword sets were calculated for both ISI and Author Keywords. Different word space models were also built out of the above-mentioned text fields for each article pair. A special 'AllText' field was also created by joining together all text fields of each citing or cited article into one long string. In a first series of word space models, word tokenization was done on the text fields of all citing/cited article pairs, stop words were removed, and the remaining text data was vectorized based on TF-IDF-weighted values. Another series of words space models was generated by converting all text data into vectors of TF-IDF-weighted 3-grams and 4-grams. Then, for both series of matrices, cosine distances between the corresponding text fields of each article pair were calculated.

Results of these computations are shown in Table 1. In the case of Jaccard similarity scores, mean values for both keyword types are lower than those obtained in [2], while relative standard deviation values, expressed in percentage of mean values, are also very high. At first glance, it thus seems that the low degree of semantic similarity reported in [2] is even lower when Economics citing articles are considered. This trend is further supported by cosine similarity scores, as the aggregated average scores obtained for both word-based and NGram-based models, while being higher than the average Jaccard similarity scores, are still as a whole below those obtained in [2]. As for standard deviation scores, their values are also regu-

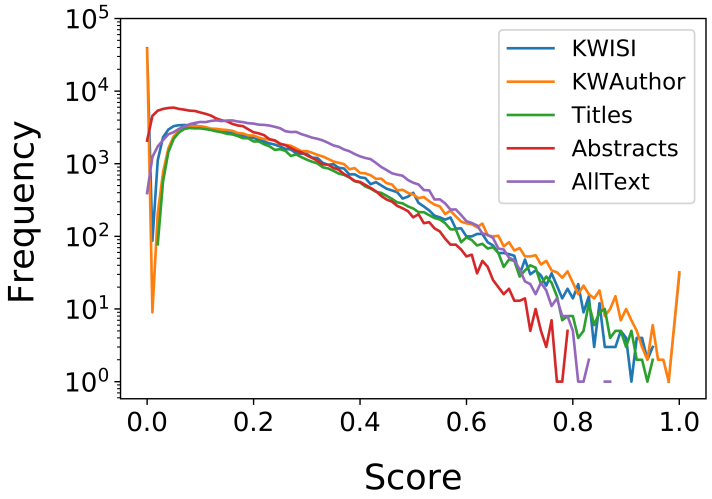
larly greater than their respective means, regardless of the similarity metric or the text field considered; such findings, consistent with those of [2], suggests that the semantic relationship between citing and cited documents is not only generally low, but also highly variable.

Table 1: Similarity of Citing/Cited Text Fields

	Jaccard		Cosine			
	Keywords		Words		NGrams	
Field	μ	RSD%	μ	RSD%	μ	RSD%
ISI	.05	64.8	.15	126.9	.20	126.6
Author	.04	49.2	.28	255.7	.22	163
Title	–	–	.12	83.6	.18	121.5
Abstract	–	–	.17	117.6	.15	97.7
AllText	–	–	.14	93	.34	276.7

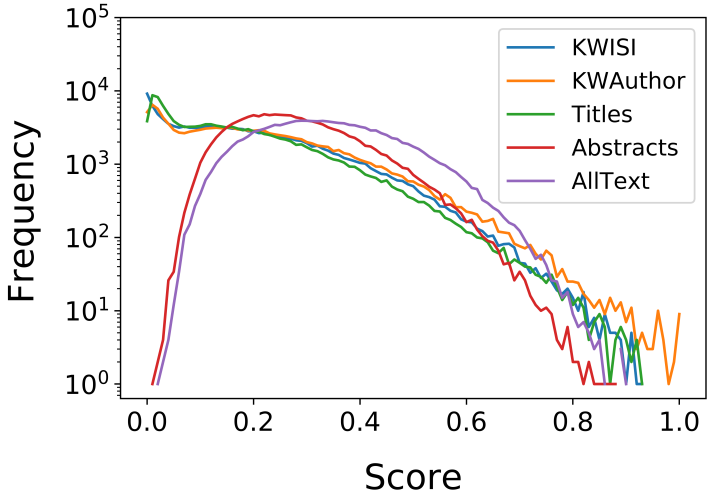
Also interesting to analyze are the distributions of the similarity scores obtained by word-based and n-gram-based models, shown in Figures 1 and 2 respectively.

Figure 1: Cosine scores for word-based models



Overall, all model types are right-tailed, the overwhelming majority of citing/cited pairs being scarcely, if at all, similar. Except in the case of author keywords, no citing/cited pairs are completely similar. At the other spectrum, no citing/cited abstract and AllText pairs are completely dissimilar in n-gram context; after all, the probability that two titles or abstracts from the same field do not have any 3-character substrings in common must be extremely low. Most surprising however is the less asymmetric and more curve-like shape of the two corresponding distributions. The fact that this structural peculiarity is

Figure 2: Cosine scores for n-gram-based models



more pronounced in the case of the allText n-gram model, which has more text content than its abstract n-gram counterpart, seems to indicate that similarity scores positively correlate with the quantity of text analyzed. These changes can be explained mathematically: longer texts have more n-gram occurrences, which densifies text matrices and thus results in dimensionally richer vectors and finer-grained similarity computations. While this explanation raises the necessity for a better control of text sample size, the unique shapes of both n-gram models also raises the possibility that analyzes based on the full-text of citing and cited articles might provide results that are more consistent with the the prevailing assumptions on the semantic scope of citations. In this sense, the present study does not rule out the possibility of any significant or robust semantic relationship between citing and cited articles. However, given the amount of data considered and the unequivocalness of the results, this claim certainly seems less plausible and reasonable now.

References

[1] E. Garfield and R. K. Merton. *Citation indexing: Its theory and application in science, technology, and humanities*, volume 8. Wiley New York, 1979.

[2] S. P. Harter, T. E. Nisonger, and A. Weng. Semantic relationships between cited and citing articles in library and information science journals. *Journal of the American Society for Information Science*, 44(9):543–552, 1993.