STI 2018 Leiden

23rd International Conference on Science and Technology Indicators
"Science, Technology and Innovation Indicators in Transition"

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Do you cite what I mean? Assessing the semantic scope of bibliographic coupling in economics[1]

Maxime Sainte-Marie[*], Philippe Mongeon[**] and Vincent Larivière[***]

[*] *msaintemarie@gmail.com*
Chaire de recherche du Canada sur les transformations de la communication savante, École de Bibliothéconomie et des Sciences de l'Information, Université de Montréal, CP6128, Station Centre-Ville, Montreal, Quebec, H3C3J7 (Canada).

[**] *p.mongeon@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Leiden University, P.O. Box 905, 2300 AX Leiden, The Netherlands

[***] *vincent.lariviere@umontreal.ca*
Chaire de recherche du Canada sur les transformations de la communication savante; Observatoire des Sciences et les Technologies, Université du Québec à Montréal, Pavillon Paul-Gérin-Lajoie, 8e étage,1205 Saint-Denis, Montréal, H2X 3R9.

Introduction

Bibliographic Coupling (henceforth BC) is one of the earliest statistical methods used to analyze scientific production and structure at different granularity levels. Introduced by Kessler in the 1960s (1962, 1963a, 1963b, 1965), BC is a retrospective and static citation measurement technique based on the number of bibliographic references two articles share: two articles are bibliographically coupled if they cite at least one common document, and the strength of this coupling is proportional to the number of references shared.

As with citation analysis in general, BC is based on the idea that there is "an implied relation" (Kessler, 1965: 223) between papers that show similar bibliographic properties. But the precise nature of this "implied relation" is still open to debate. Following his analyses, Kessler himself concluded to the subject relatedness of bibliographically coupled documents (Kessler, 1963). Reviewing the latter's work on BC, Weinberg (1974) refused for its part "to subscribe to the notion of the citation being a unit of content, with its implication that papers with the same references are identical in content" (Weinberg, 1974: 195). On the other hand, many scholars went as far as conferring BC a semantic scope. In the first large-scale and multidisciplinary assessment of the validity, feasibility, and effectiveness of BC for subject analysis in large citation databases, Vladutz and Cook (1984) concluded in the affirmative, arguing that BC "may prove to be the easiest approximation to an algorithm for revealing the semantically closest neighbours of publications" (cited in Jarneving 2007: 290). The persistence of this semantic stance is even prominently featured in the Wikipedia page on co-citation analysis, which describes BC as "a semantic similarity measure for documents that makes use of citation relationships" (https://en.wikipedia.org/wiki/Co-citation).

[1] This work was supported by the Canada Research Chair program and the Canada foundation for innovation.

Beyond these speculations, no study has so far settled the matter, and the questions as to "when and to what extent [BC] can be considered [a measure] of semantic similarity" still "needs to be established theoretically" (Hjørland, 2013: 1315). Despite this theoretical gap, one thing can be taken for granted: if we assume that meaning is carried by words, no strong correlation between BC and semantic similarity can reasonably be expected. Two reasons can be given in support for that thesis: first, since the probabilities that any two articles have words or references in common are respectively very high and infinitesimal, the majority of scientific papers will be semantically related to some extent, yet totally unrelated bibliographically. Additionally, given that the scientific community as a whole promotes novelty in scientific contributions and takes a hardline approach to plagiarism, a lot of articles may have the same references, yet no two papers will ever have the same content. This limit to the semantic similarity of bibliographically coupled scientific documents is also furthered by the fact that landmark scientific contributions are often translated in different languages, a common practice which results in the production of hundreds of documents that have identical bibliographies but do not share a single word (Hjørland, 2013: 1315). Thus, from a strictly correlational perspective, the idea of using bibliographic similarity as a proxy for semantic similarity can be dismissed from the outset.

However, that doesn't necessarily mean that bibliographic similarity is unconditionally uncorrelated with semantic similarity. Indeed, by limiting the scope of analysis to articles sharing at least one reference, a new and plausible hypothesis can be formulated: for any pair of articles that have at least one reference in common, coupling strength directly correlates with semantic similarity. In other words, the more references bibliographically-related articles share, the closer their meaning. If this were the case, coupling strength would constitute a good enough proxy for semantic similarity, thus warranting the use of bibliographical analysis for semantic purposes: the references section of articles could then provide a reliable overview of article content, thus reducing the reliance on text-rich data or methodologically and theoretically more complex text-based analyses for bibliometric purposes. But such use cases are only possible insofar as the real semantic scope of BC is known beforehand, hence the relevance of a proper assessment. The purpose of the present paper is to assess the correlation between BC strength and semantic similarity using word-based as well as 3- and 4-gram-based weighted vector space models.

Methods

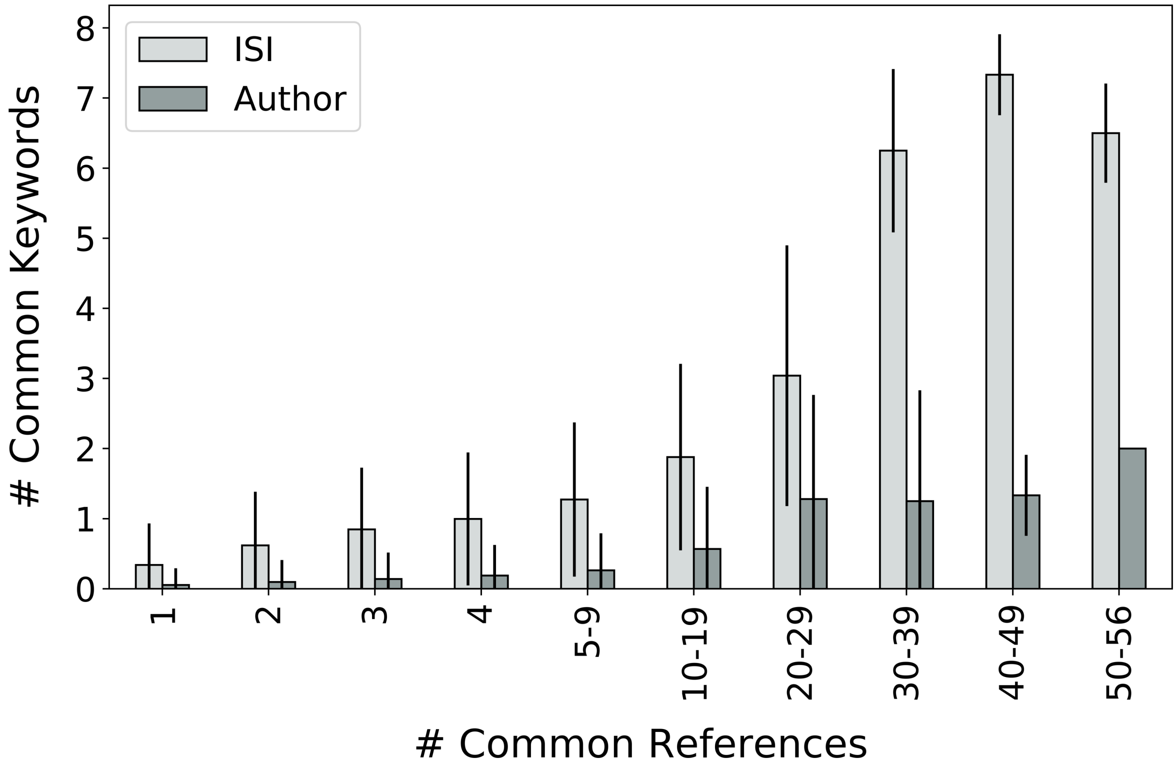
For testing purposes, relevant data (Title, Abstract, Author Keywords, ISI Keywords, References) from all 15,461 Web of Science articles published in 2015 and classified as Economics publications by to the NSF field classification of journals were extracted. Economics was chosen due to its strong and often-reported intradisciplinary character (Pieters, 2002; Jacobs & Frickel, 2009); we presumed this disciplinary feature could help reduce the impact cultural and extradisciplinary factors might have on both wording and citation behavior. In order to facilitate comparison between documents and data types, any article missing one or more of the analyzed fields was excluded, thus reducing the size of the dataset to 10,890 articles. Of the $(n(n-1))/2 = 59{,}290{,}605$ possible article combinations, only the bibliographically coupled ones (i.e., article pairs with at least one common reference to a source item) were considered for this study, amounting to a total of 493,590 combinations amongst 10,657 articles. Thus, while the selected article pairs only amount to 0.83% off all possible combinations, the following analysis will nonetheless involve 97.86% of all articles in the dataset. As shown in Table 1, the frequency distribution of coupling strength for bibliographically-related Economics article pairs is highly skewed, with a mode and a median of 1, as well as a mean and a maximum of 1.41 and 56 common references per article pair respectively.

Table 1. Distribution of extracted articles based on coupling strength

Number of common cited references	**Number of article pairs**
1	373,968
2	76,177
3	24,814
4	9,546
5-9	8,351
10-19	696
20-29	25
30-39	8
40-49	3
50+	2

Given this distribution and the often-presumed subject-relatedness of BC (Kessler, 1963a, 1965; Vladutz and Cook, 1984), the most straightforward way to analyze the relationship between bibliographic and semantic similarities for reference-related Economics articles would be to group article pairs by coupling strength (number of common references), and compute for each group thus formed the average number of shared keywords. The rationale is simple: if bibliographic and semantic dimensions of articles are intertwined, the more references bibliographically coupled articles share, the more keywords they ought to have in common. The aggregated distribution presented in Figure 1 seems to point to this interpretation, as the mean number of common keywords between two articles increases with the number of citations they share. This point of view is further supported by the fact that the correlation scores for ISI and Author Keywords are of 0.16 and 0.29 respectively.

Figure 1. Mean number of shared keywords by article pair, sorted by coupling strength

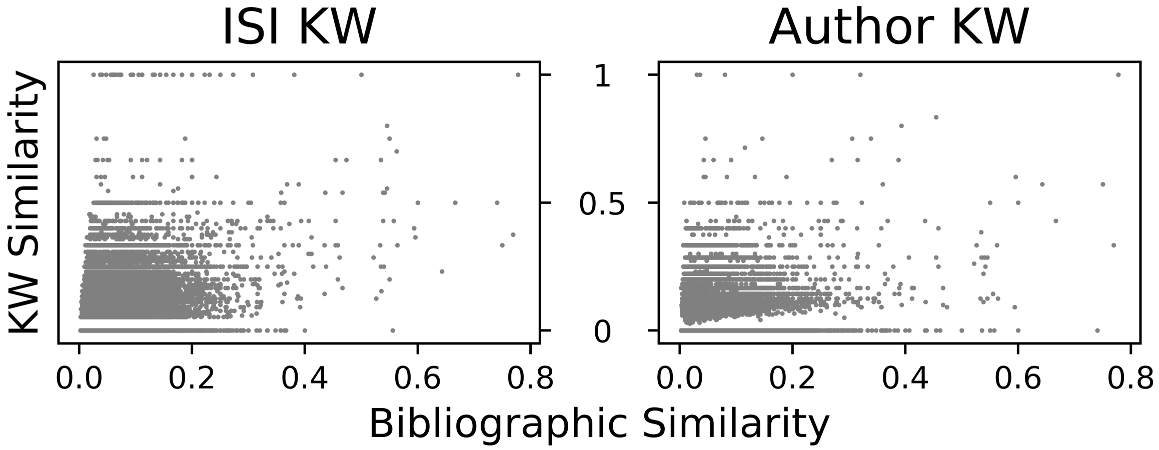


However, the relatively high standard deviation values in most bins suggest that the positive relationship between shared keywords and references may represent an aggregation artefact. Moreover, results may be biased by the fact that only common items are considered for article pairs. Indeed, considering only common keywords and references can be misleading, since two article pairs may have the same number of shared keywords and references, and yet have radically different total keywords and reference counts. Thus, in order to allow for meaningful comparisons between article pairs, shared references and keyword counts have to be weighted in proportion to the total number of entities involved. Mathematically, this can be done by calculating the Jaccard similarity coefficient for both references and keywords. Designed to measure the element overlap between two sets, the Jaccard similarity between two sets A and B is obtained by dividing the cardinality of their intersection (the number of common elements they share) by the cardinality of their union set (the total number of distinct elements they contain) :

$$Sim(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Thus, for each Economics article pair, correlation between semantic and bibliographical similarity is here obtained by calculating the Jaccard similarities of the paired articles' keyword and reference sets. Results are shown in Figure 2.

Figure 2. Keyword (KW) and bibliographic Similarity Ratios of 2015 economics article pairs



At first glance, similarity scores are rather low, as most article pairs score below 0.4 for both citation and keyword similarity. As for the relationship between both types of similarity, Figure 2 tells a radically different story than Figure 1. While it is true that the lower keyword similarity bounds in both subfigures slightly increases with bibliographic similarity, more so in the case of author keywords, their correlation seems tenuous at best. In fact, coefficient of determination (R^2) scores between keyword similarity and citation similarity are pretty low, with 0.03 and 0.07 for ISI keywords and Author keywords respectively. These scores mean that variation in keyword similarity can by itself only account for less than 5% of the variance in citation similarity, which isn't much. However, the significance of these results may be hampered by the fact that reducing article content to the sole presence or absence of keywords in the database oversimplifies the semantic dimension of articles. To give but one example, while keywords such as 'information science' and 'information retrieval' are clearly related in content, an "all-or-nothing", character-by-character match does not take this semantic kinship into account. This questionable keyword-based approach is all the more limited given the fact that other, richer text data from all extracted economics articles is available for analysis.

In order to conduct a more thorough analysis, different word space (Schütze, 1993) or bag-of-words models were built out of the relevant field information for each article pair, namely article titles, abstracts, ISI Keywords, and Author Keywords. A special 'All Text' field was also created by joining together, for each article of each article pair, all the previously mentioned field information units into one long string.

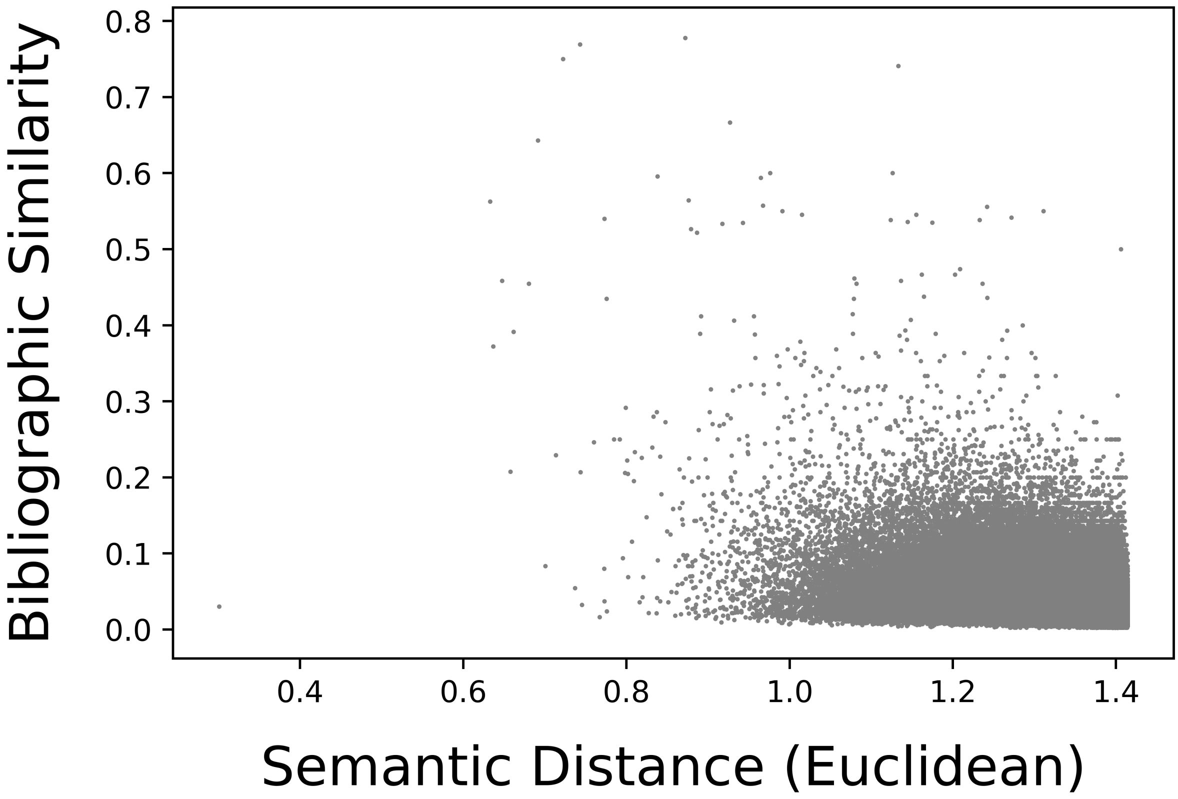
In a first series of word space models, word tokenization was done on all the above-mentioned text fields for each article pair, stop words were removed from each text field, all remaining text data was converted into word vectors based on TF-IDF-weighted values. Another series of word space models was generated by converting all text data for each article pair into vectors of TF-IDF-weighted 3-grams and 4-grams. As a justification for this second approach, research has shown that text analysis tasks based on character sequence tokenization offer comparable results to natural language, word-based, approaches (Cavnar & Trenkle, 1994; Damashek, 1995; McNamee & Mayfield, 2004). For both series of matrices, Euclidean and cosine distances between the corresponding text fields of each article pair were calculated, and R^2 scores were calculated between these computed distances and the Jaccard similarity scores obtained above for article references. Since the P-values for the different R^2 scores were all below 10^{-293} (the default value below which the Scipy module returns 0.0), all results were deemed statistically significant and thus kept for further analysis. Final results are presented in Table 2 as percentages of variation explained.

Table 2. Percentage of Bibliographic Similarity variation explained (based on R2 scores)

	Words		NGrams	
	Cosine	Euclidean	Cosine	Euclidean
Titles	4.14	4.28	3.64	3.80
Abstracts	7.35	7.46	4.68	4.91
KW - ISI	10.52	11.31	6.69	7.52
KW - Author	6.41	6.63	5.48	5.79
All Text	11.14	11.47	7.65	8.13

Interesting patterns can be discerned from Table 2. For instance, percentages for Euclidean distances are systematically higher than those for cosine distances scores. Also, results for word tokens are higher than those for n-gram tokens, while Title and All Text scores are respectively the lowest and highest registered. But more importantly, regardless of the article field type, the distance used or of the tokenization method applied, variations in coupling strength usually account for less than 10% of semantic variations as measured using word space models. These scores, in conjunction with those presented in Figure 2, shed serious doubt on the content- or subject-relatedness of BC. This interpretation is further supported by Figure 3, which shows the similarity distribution of the highest-scoring setting, based on euclidean distances of Words from All Text Fields (11,47%).

Figure 3. Similarity Distribution for 2015 WoS Economics Articles



While the distribution slope is inverted from the one in Figure 2, similar trends can however be observed. First, both similarity scores fill the lower half of their respective axis, most article pairs being less similar than different from either aspect. An angular narrowing similar to those noticed in Figure 2 can also be found here for both axes, meaning that both similarity types slightly constrain the lower bound of their counterpart's score. Beyond these interesting structural patterns, however, one general finding stands out: in light of the different results obtained by the current analysis process, no clear patterns in covariation can be found between BC and semantic similarity, thus excluding the possibility of finding any strong linear or non-linear correlation in the data.

Discussion

Overall, our results are clear: as far as titles, abstracts, and keywords of bibliographically coupled Economics articles published in 2015 are concerned, the semantic scope of BC is very small, especially given its often-intended use as a "semantic similarity measure for documents" (https://en.wikipedia.org/wiki/Co-citation). Bibliographically coupled articles strength might at times be semantically similar, but a high coupling strength between two articles can by no means stand as an indicator of semantic similarity.

In our opinion, these results can be seen as a reminder of the unavoidably social nature of scientific communication. Just as academic disciplines represent both bodies of knowledge and social units, scientific papers are not simply knowledge repositories, but knowledge claims made in and aimed at specific communities. And while scientific communication is at the same time an expression of "the social and the intellectual organization of knowledge" (Hjørland, 2013: 1324), results of the present research suggests that BC pertains more to the former than the latter.

However, the validity of these conclusions may be challenged on the basis that various unaccounted factors might have affected the relationship between semantic and bibliographical similarities. On the one hand, since there are both spatial and cognitive limits to referencing, the latter is thus necessarily partial, subjective, and thus prone to individual variations which have little to do with semantics. In addition, other referencing factors such as methodological borrowings, obliteration by incorporation, namedropping, and domain- or paradigm-specific practices could have affected the relationship between semantic and bibliographic similarity. The same relationship might also have been compromised by wording behavior: after all, semantically similar text units can be written using very different words, whereas opposite statements such as is\is-not claims can be made using near-identical sets of words. Similarly, it could also be argued that reducing article semantics to title, abstract, and descriptor information might be too simplistic. Indeed, given the fact that the best R^2 scores were obtained when all the available text was considered, research based on the full text of articles might deliver better and more representative results.

Beyond the arbitrariness of both referencing and wording behavior, the quality of the database used in this research might also be questioned on various grounds. For once, indexing of Economics publications by the Web of Science is far from complete: monographs, proceedings, non-English publications and other types of scientific communications of relevance to the field are left out; also, only references to articles already indexed are included in the database. In parallel, field partitioning based on the NSF classification of journals is certainly not perfect: surely, not all articles included in Economics-classified journals pertain to that field, and not all

Economics articles are included in Economics-classified journals. In sum, journal, article, and reference coverage might all be troublesome to some extent.

In light of these various limitations, the present study does not rule out the possibility of any significant or robust relationship between BC and article content. However, given the amount of data considered and the unequivocalness of the results, this claim certainly seems less plausible and reasonable now.

References

Boyack, K.W. & Klavens, R. (2010). Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately? Journal of the American Society for Information Science and Technology, 61, 12: 2389-2404.

Cavnar, W.B. & Trenkle, J.M. (1994). N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval (Las Vegas, NV, 1994), 161–175.

Damashek, M. (1995). Gauging Similarity with *n*-Grams: Language-Independent Categorization of Text. Science, 267: 843-846.

Harter, S.P, Nisonger, T.E. & Weng, A. (1994). Semantic Relationships between Cited and Citing Articles in Library and Information Science Journals. Journal oft he American Society for Information Science 44, 9: 1993.

Hjørland, B. (2013). Citation analysis: A social and dynamic approach to knowledge organization. Information Processing and Management 49: 1313-1325.

Jacobs, J. A. & Frickel, S. (2009). Interdisciplinarity: A Critical Assessment. Annual Review of Sociology, 35: 43-65.

Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. Journal of Informetrics 1, 4: 287-307.

Kessler, M. M. (1962). *An experimental study of bibliographic coupling between technical papers* (No. 62 673TN1). Cambridge, MA: Massachusetts Institute of Technology, Lexington Lincoln Lan.

Kessler. M. M. (1963a). Bibliographic Coupling Between Scientific Papers. Journal of the Association of for Information Science and Technology, 14, 1: 10-25.

Kessler, M. M. (1963b). Bibliographic coupling extended in time: Ten case histories. *Information storage and retrieval*, *1*(4), 169-187.

Kessler, M. M. (1965). Comparison of the results of bibliographic coupling and analytic subject indexing. *Journal of the Association for Information Science and Technology*, *16*(3), 223-233.

Lu, K. & Wolfram, D. (2012). Measuring Author Research Relatedness: A Comparison of Word-Based, Topic-Based, and Author Cocitation Approaches. Journal of the American Society for Information Science and Technology, 63, 100: 1973-1986.

McNamee, P. & Mayfield, J. (2004). Character *N*-Gram Tokenization for European Language Text Retrieval. Information Retrieval 7: 73-97.

Pieters, R. and Baumgartner, H. (2002). Who talks to whom? Intra- and Interdisciplinary Communication of Economic Journals. Journal of Economic Literature XL: 483-509.

Schütze, H. (1993). Word Space. In: Advances in Neural Information Processing Systems,
Ed. S. Hanson, J. Cowan , C. Giles, Morgan Kaufmann Publishers Inc., pp. 895-902.

Vladutz, G., & Cook, J. (1984). Bibliographic coupling and subject relatedness. *Proceedings of the American Society for Information Science*, *21*, 204-207.

Weinberg, B.H. (1974). Bibliographic coupling: A review. Information Storage and Retrieval, 10, 5-6: 189-196.