Journal of Scientometric Research, 2024; 13(3):706-714.
https://www.jscires.org

Research Article

Bibliographic Coupling and Conceptual Similarity: Are the Bibliographically Coupled Papers also Conceptually Similar?

Abhirup Nandy[1,2,*], Aakash Singh[1,3], Vedika Gupta[4], Vivek Kumar Singh[1, 5]

[1]Department of Computer Science, Banaras Hindu University, Varanasi, Uttar Pradesh, INDIA.
[2]CPS Lab, Institute of Informatics and Communication, University of Delhi, Delhi, INDIA.
[3]School of Computer Science Engineering and Technology, Bennett University, Greater Noida, Uttar Pradesh, INDIA.
[4]Information Systems and Analytics, Jindal Global Business School, O.P. Jindal Global University, Sonipat, Haryana, INDIA.
[5]Department of Computer Science, University of Delhi, Delhi, INDIA.

ABSTRACT

Bibliographic coupling, over the years, has been referred to and used in different contexts related to scientific and technical literature. It is often believed that research papers that have bibliographic coupling deal with similar concepts and hence there may be high conceptual similarity between them. This study attempts to empirically asses this notion. To conduct this research, the study utilizes the data obtained from the Dimensions database and employs advanced machine learning algorithms to extract weighted keywords that better capture the conceptual content of documents. The Jaccard similarity measure is used to compute bibliographic and conceptual coupling matrices for different sets of research papers. The results show that even though bibliographic coupling is widely used to assess relationships between research papers, it often falls short of identifying actual conceptual similarities within documents. This study's findings carry important implications for areas such as information retrieval, interdisciplinary research and evaluation metrics, calling for a more refined understanding of how research documents relate to one another beyond their shared references.

Keywords: Bibliographic Coupling, Conceptual Coupling, Conceptual Similarity, Semantic Similarity.

Correspondence:
Abhirup Nandy[1,2]
[1]Department of Computer Science
Banaras Hindu University, Varanasi,
Uttar Pradesh, INDIA.
[2]Institute of Informatics and
Communication, University of Delhi,
New Delhi, Delhi, INDIA.
Email: abhirupnandy.online@gmail.com
ORCID: 0000-0001-8618-0847

Received: 11-01-2024 ;
Revised: 19-04-2024;
Accepted: 20-05-2024.

INTRODUCTION

In the large set of ever-growing academic literature, an analysis of the presence of relationships between articles is a difficult task. Despite the employment of various techniques to correlate publications from different perspectives, such as semantic and morphological ones, numerous unexplored connections remain between them. One such view includes the study of the bibliographic metadata of the publication, where the authors cite various other articles in their work. This was termed "Bibliographic Coupling" (or BC), first published by M. M. Kesslar in 1963.[1,2] While studying a method for grouping technical and scientific documents. The coupling method assigned a score to the relationship between two documents based on their common references. Initially, a network of publications and their references is created, which is quite similar to a co-citation network. Kessler used this network to build a hypothesis that strongly coupled documents might also be quite similar in their cognitive similarity. Hence, this network of coupled documents and their references helps to analyse two important aspects of document similarity: (i) How to determine if two documents are similar in nature? (ii) How to quantify the similarity between them? The first issue can be addressed by leveraging the inter-article linkages based on common references, while the second issue can be solved using the weights assigned to these linkages. Since these weights are assigned based on an intersection of the set of references, two similar documents are supposed to be coupled with higher relative weights. Co-citation analysis, document retrieval techniques and knowledge mapping techniques can utilize the coupling strength between the articles.

DOI: 10.5530/jscires.20041115

Publishing Partner : Manuscript Technomedia. [www.mstechnomedia.com]

In the initial study, Kesslar analysed the coupling strength between 265 articles of Physical Review from 1958 and concluded that the bibliographic coupling phenomenon can interpret the similarity between coupled papers.[1] Later, another study was performed using 1,712 publications to build an improved document indexing technique using citations.[3] These studies demonstrated the effective use of bibliographic coupling techniques in introducing a semantic comparison between documents. A similar study by Sen and Gan, based on the mathematical framework of bibliographic coupling,[4] inferred that the technique can be used to establish a “cognitive chain of information” and to group articles according to their relative similarities. Further studies of citation network analysis[5] used bibliographic coupling along with cluster analysis to group together scientific articles on specific topics. In later research, Jarneving used bibliographic coupling along with clustering techniques to successfully map a two-level cluster of scientific documents.[6,7] He used bibliographic coupling as a way to group semantically similar documents. Another study, based on a collaboration recommendation system, was conducted by Park *et al.*, where Bibliographic Coupling and Latent Semantic Analysis were used to generate recommendations for patent collaborations.[8] The authors showed that the use of bibliographic coupling revealed various relationships based on the semantic structure of articles, which traditional keyword analysis would not have found otherwise. Overall, these studies suggest that Bibliographic Coupling might be an efficient approach to establishing the semantic similarity between articles on various levels of similarity.

Although many studies have shown Bibliographic Coupling to be effective in determining semantic similarity, it is still subject to limitations and criticisms. Martyn and Weinberg both criticized the Bibliographic Coupling technique, with Martyn noting its inability to pinpoint specific areas of similarity in subsections of paired publications, while Weinberg emphasized its failure to differentiate between varied contextual uses of the same cited document within coupled papers.[9,10] While considering publications that appear to be similar, it is still important to do a thorough study of their conceptual framework. Though the semantic structure summarises the main ideas of a document, the conceptual framework explores the underlying concepts that are present in each part. By highlighting previously disregarded details, such author-defined keywords, this method allows readers to gain a comprehensive and multifaceted knowledge of the text. As a result, this approach guarantees a comprehensive analysis of every concept that contributes to the thematic content and semantic meaning of a document, offering a more comprehensive and nuanced analytical viewpoint.

A related study utilizing 'author-provided keywords' and 'WoS indexed keywords' from the Web of Science database revealed that while author-provided keywords offer some conceptual structure, they tend to exhibit bias towards specific sections of the study, potentially skewing the overall representation of the research content.[11,12] The WoS indexed keywords are derived from the titles of cited publications^2, that focuses more on the titles than the full text of the publication. Hence, these two fields (author keywords and WoS keywords) do not fully represent the entirety of the article. However, the study's focus on a thematic dataset of economics articles from 2015 introduces a potential relative bias in the bibliographic coupling analysis, as noted by the thematic concentration of the selected literature.[10,11]

The current study attempts to overcome these issues by using the ‘concepts’ field from the Dimensions database.[3] The concepts are represented as tuples comprising weighted terms and relevance scores, derived through the parsing of publications by the database using various machine learning algorithms.[4] This ensures that the whole document is well represented within a set of weighted concepts and hence would be better suited to bring forth the semantics of the document. As a result, a more comprehensive approach to semantic similarity is taken into account when comparing two articles based on how similar their concept phrases are. Also, a large-scale dataset (~30,000 articles) is used in this study, which is adequately admissible to analyse the similarity between coupled articles.[13] The selected dataset contains articles from 2010 to 2019 and is divided into two subsets- (i) 10 journals from the field of ‘Scientometrics’ and (ii) articles mentioned with the theme of ‘Scientometrics’. Taking the two different datasets ensures a relatively thorough inclusion of articles on the theme. The motivation behind this study is to check for a possibility of conceptual similarity between bibliographically coupled articles. The conceptual similarity was chosen so that the whole theme of the article could be described using these weighted terms, without any bias, which might be the case with the author-provided keywords. Notably, the concept terms were selected with weight≥0.6 (out of 1), thereby focusing on the major concepts embedded in the article. Overall, this study attempts to answer the following research question:

RQ: Do the research papers with higher bibliographic coupling also show higher conceptual similarity and *vice-versa*?

The study is organised as follows: Section 2 represents a review of studies related to bibliographic coupling and its effectiveness in determining any semantic or conceptual relationship between articles. Section 3 describes the dataset in detail and the methodology used in this study to investigate the conceptual relations between articles. Section 4 displays the analysis's results, as well as discussions based on the findings. The paper concludes with a summary of the work in Section 6, along with the implications and applications of the analysis and the possible future scope of the study.

RELATED WORK

Bibliographic Coupling Approach and its Applications

As first described by Kessler in 1963,[1] the Bibliographic Coupling technique focuses on examining the connections between the cited references of two articles. Kessler established two separate

https://clarivate.com/webofsciencegroup/essays/concept-citation-indexing/
https://app.dimensions.ai/
https://docs.dimensions.ai/dsl/language.html#searching-using-concepts

standards for determining documents that are bibliographically coupled:

Criterion A

A number of papers constitute a related group if each member of the group has at least one coupling unit to a given test paper P_0. The coupling strength between P_0 and any member of is measured by the number of coupling units (n) between them. is that portion of that is linked to P_0 through n coupling units.

Criteria B

A number of papers constitute a related group if each member of the group has at least one coupling unit to every other member of the group.

Using a dataset of scientific publications, specifically 36 volumes of Physical Review, Kessler tested the Bibliographic Coupling method[2] and was able to demonstrate the accuracy in identifying subject-based similarity between papers that were bibliographically coupled. Kessler also examined various scenarios by relocating the test paper within the dataset based on its publication year. Consequently, Bibliographic Coupling demonstrated its applicability to papers spanning both historical and future contexts.[14] The significant influence on bibliometric studies based on this approach arose due to the existence of citation network structures. Various studies adopted similarity measures based on Bibliographic Coupling, such as comparing it with analytical subject indexing.[15] Notably, strong correlations and connections between document pairs formed by Bibliographic Coupling and subject indexing were found. Another study demonstrated the cognitive resemblance between word-profiles of documents and bibliographically coupled document pairs, further supporting the notion that Bibliographic Coupling works effectively within specific research fields.[16]

Moreover, Bibliographic Coupling has been employed to uncover "hot" research topics by identifying core documents through threshold values for common reference counts and coupling strength.[17] These studies showed that strong coupling links could trace the evolution of specialities within a subject area and classify core documents as significant community articles. Recent research extended this approach to assess the relationship between core research field strengths and research scale, contributing to diversifying research fields and enhancing productivity.[18]

Criticisms and Advancements in semantic similarity

While Bibliographic Coupling found acceptance within the bibliometric community, it faced criticism from various quarters. Weinberg criticized the approach for its bias towards Kessler's initial document selection, given that these documents had already formed meaningful clusters.[10] However, a study conducted on the entire 1981 SCI database countered this claim and demonstrated the feasibility of applying the Bibliographic Coupling approach to a vast dataset, yielding semantically related publications.[13]

Another critique contested the validity of Bibliographic Coupling as a unit of measurement for document relationships, suggesting it is a mere indicator of the likelihood of relatedness.[9] This criticism highlighted that even if two articles A and B both cite article C, the information cited might differ, challenging the assumption of relatedness through Bibliographic Coupling.

Sen and Gan proposed an alternative coupling strength to address these concerns and assess cognitive or semantic resemblance between document pairs.[4] This coupling strength was based on the vector norm instead of absolute values of intersection between the cited references. A value of $\theta = 60°$ was suggested as a threshold value for coupling strength, which corresponded to a value of approximately 0.5. Later, to quantify the similarity between sets of cited documents, researchers introduced cosine similarity as the coupling angle value for Boolean document vectors.[17]

Advancements involved incorporating additional elements into the Bibliographic Coupling score. Habib introduced a weighted approach based on the article's citation location,[19] boosting the significance of references within the same section. The inclusion of context passages or phrases, alongside the cited article, enabled the maximization of similarity between highly scoring document pairs.[20] Another method employed descriptions of articles with high coupling scores to enhance the approach.[21]

Notably, efforts emerged to leverage metadata and word profiles to overcome reliance on full-text data. Studies by Braam *et al.*,[16] showcased the utility of metadata fields like keywords in tandem with the Bibliographic Coupling value to enhance relatedness possibilities. However, token-based similarity studies questioned the effectiveness of the Coupling scores, revealing deviations in distributions.[22]

In the quest to explore the relationship between Bibliographic Coupling and semantic similarity, studies demonstrated inconsistencies. Although bibliographically coupled documents shared similar keyword terms, the coupling strength did not necessarily indicate semantic resemblance.[11] This discrepancy was evidenced in a study focusing on the Economics domain, which couldn't establish a clear pattern of covariation between Bibliographic Coupling Score and Semantic Similarity.

To further extend the study, we attempted to investigate a similar question of conceptual similarity between bibliographically coupled articles. We have used a set of concepts and their relevance scores which defines the document, rather than using author-provided keywords.

This use of a weighted conceptual strength was intended to ensure more focus on the main theme of the article rather than a list of unweighted keywords.

Data and Methodology

Data

The present study has used two sets of data, referred to as Dataset 1 (D_1) and Dataset 2 (D_2). The first dataset (D_1) comprises scientific articles extracted from ten selected journals (published between 2010 and 2019) in a single field of study, i.e., Scientometrics. We selected articles from journals in the same field, motivated by the possibility of several article pairs with high bibliographic coupling. The second dataset (D_2) comprises all the articles retrieved on a topical query "semantic similarity" with time period set to 2010-2019. The rationale behind employing a specific query to gather data was the expectation that obtaining articles on a particular topic would likely yield several pairs of articles with strong bibliographic correlation and semantic similarity. Dimensions database is used as the source of the data and was selected mainly due to the fact that it provides a list of 'concepts' discussed in an article. These concepts are weighted keywords extracted/generated from the full text of the article, using some machine learning approaches. The Dimensions database provides a set of such concepts, along with a relevance score for each concept term. Table 1 lists the metadata fields that were downloaded from the Dimensions database for both datasets.

From the extracted meta-data, we have mainly utilised the fields of ID, reference IDs and concept scores for our experimental work. The initial pre-processing involved removing missing data rows and other inconsistencies. Further, articles that had no concepts provided were filtered out. Dimensions provide a score (between 0-1) with the concepts, which denotes a kind of concept relevance to the article. Therefore, we have used a threshold of 0.5 for the concepts, i.e., only those concepts were considered that have a relevance score greater than or equal to 0.5. This operation was performed so that only the highly relevant concepts are used. The pre-processing left us with 4,475 publication records in D_1 and 11,960 publication records in D_2. Table 2 presents the details of the datasets.

Methodology

The cleaned dataset is utilized to compute bibliographic coupling and conceptual coupling strength. For this purpose, we calculate Jaccard similarity measures. Jaccard similarity is a trivial proximity measurement used to detect similarity between two objects or sets. The reason for using this metric is that it closely represents the normalized bibliographic coupling method.

$$BC_{ij} = \frac{\left|R_{p_i} \cap R_{p_j}\right|}{\left|R_{p_i} \cup R_{p_j}\right|}$$

where, R_{pi} is the reference set of a publication p_i.

This normalized value of Bibliographic Coupling was proposed to reduce the dependency on the number of references in the paper.[4]

We have employed a matrix approach to compute the Jaccard similarity between all possible publication pairs. Using the information described above, we have created two base matrices.

Let $M_{\mathrm{IR}}{}^{1}$ be the reference matrix of D_1 and $M_{\mathrm{IC}}{}^{1}$ be the concept matrix of D_1; and $M_{\mathrm{IR}}{}^{2}$ and $M_{\mathrm{IC}}{}^{2}$ be the respective matrices for D_2.

The objective is to compute paper-wise bibliographic coupling and conceptual similarity. These similarities can be defined as the following matrices: BC_1 and BC_2 are bibliographic coupling matrices computed from the datasets D_1 and D_2 respectively. Similarly, CC_1 and CC_2 are conceptual similarity matrices computed from the datasets D_1 and D_2 respectively. The construction of these matrices is described below.

Bibliographic Coupling matrix (BC)

For D_1, we have created a reference matrix. For this, we extracted the reference IDs from each publication and created the matrix $\boldsymbol{M_{IR}}$ of **[IDs× Reference IDs]** where the dimensions of the matrix were **[4,475×100,728]**. This was created as a matrix of Boolean vectors, where a publication with a set of references R would be represented in the form of a vector as -

$$V_i=[\, v_1,v_2,v_3,\ldots,v_n]$$

Where, i=1,2, 3…, 4475

Where, n is the total number of unique references in our dataset and,

$$v_{i_j} = \begin{cases} 1, Reference\ j\ present\ in\ R_{p_i} \\ 0, Reference\ j\ not\ present\ in\ R_{p_i} \end{cases}$$

Where, $j = 1,2,3,\ldots,n$, and $n = 100{,}728$.

Table 1: List of meta-data fields extracted from Dimensions.

Metadata Field	Description
Abstract	Abstract of the paper.
Authors	List of author names and their affiliations.
Category FOR	Dimension's own provided subject category.
Concept Scores	Dimensions' own provided concepts and their scores.
DOI	Unique DOI for the publication.
ID	Unique publication ID.
Journal ID	Unique journal ID.
Journal Title	Title of the journal as indexed in Dimensions database.
Reference IDs	Unique publication IDs of each reference.
Title	Title of the publication.
Year	Publication year

Table 2: Dataset Description.

Dataset	Description of the data	Number of articles	Number of total references	Number of total concepts
D_1	Metadata for published articles from the following 10 selected journals in the field of Scientometrics: Scientometrics Journal of Informetrics Journal of the Association of Information Science and Technology Research Policy Journal of Scientometric Research Aslib Journal of Information Management Online Information Review Science and Public Policy Quantitative Science Studies Research Evaluation.	12,748	380,121 (166,706 unique)	573,315 (114,026 unique)
D_2	Metadata of published articles retrieved on the search query “semantic similarity”.	17,655	657,265 (330,176 unique)	895,322 (181,757 unique)

Table 3: Description of data used after processing.

Dataset	Matrix type	Non-zero pairs	Intersecting Pairs
D_1	BC_1	243,904	66,341
	CC_1	514,976	
D_2	BC_2	2,164,063	649,156
	CC_2	6,597,022	

Using this matrix, we calculated the pairwise Jaccard Similarity for all possible paper pairs and represented it with the matrix BC_1. This is done using the formula for Jaccard Similarity computed between two vectors of binary data, which can be written as:

$$BC_{1ij} = Sim_{jacc_{ij}} = \frac{(\boldsymbol{M}_{IR}^{D_1})^T(\boldsymbol{M}_{IR}^{D_1})}{n - (\overline{\boldsymbol{M}_{IR}^{D_1}})^T(\overline{\boldsymbol{M}_{IR}^{D_1}})}$$

Where n = total number of unique references in the dataset.

Thus, we got a square matrix BC_1 of dimensions [4,475×4,475] containing pairwise Jaccard Similarity based on cited references of publications present in D_1, which computes to be equal to a normalized Bibliographic Coupling score between the two publications. Similar calculations were done to construct the matrix for dataset D_2.

Conceptual Similarity Matrix (CC)

The Conceptual Similarity matrix (can also be referred to as the Conceptual Coupling matrix) was constructed as follows. We have extracted the concepts from each article and created a matrix M_{IC} of [IDs × Concepts], where the dimensions of this matrix were [4,475×42,213]. It was created with similar steps as a matrix of Boolean vectors, where a publication with a set of concepts C would be represented in the form of a vector, as-

$$V_i=[\ v_1,v_2,v_3,\ldots,v_m]$$

where, $i = 1,2,3,\ldots,$ 4475 and m is the total number of unique references in our dataset and,

$$v_{ij} = \begin{cases} 1, Concept\ j\ present\ in\ C_{p_i} \\ 0, Concept\ j\ not\ present\ in\ C_{p_i} \end{cases}$$

where, $j = 1,2,3,\ldots,m$ and $m = 42,213$

Using this matrix, we calculated the pairwise Jaccard Similarity for all possible paper pairs and represented it with the matrix CC_1. This is done using the same mentioned formula for Jaccard Similarity computed between two vectors of binary data. The Jaccard coefficient is utilized for assessing the keyword similarity in the same manner as that of BC, which is -

$$CC_{1ij} = Sim_{jacc_{ij}} = \frac{(\boldsymbol{M}_{IC}^{D_1})^T(\boldsymbol{M}_{IC}^{D_1})}{n - (\overline{\boldsymbol{M}_{IC}^{D_1}})^T(\overline{\boldsymbol{M}_{IC}^{D_1}})}$$

Where n = total number of unique concepts in the dataset.

Thus, we got a square matrix $\boldsymbol{CC}_1$ of dimensions [11,960×11,960] containing pair-wise conceptual Similarity based on extracted concepts of publication present in D_1, which we take to be equal to a normalized Conceptual Coupling score between two articles. The conceptual similarity matrix for D_2 was also constructed in a similar fashion.

There were pairs with zero-value entries, meaning their respective mutual similarities were not significant. Since our main interest

Table 4: Summary of BC and CC values.

Dataset	Type of document pair	Highest Coupling Strength	Mean value of coupling strength	Standard deviation value of coupling strength
D_1	Non-zero pairs from the CC_1 matrix (514,976 pairs).	1	0.030057	0.030057
	Non-zero pairs from the BC_1 matrix (243,904 pairs).	0.645161	0.017028	0.016127
D_2	Non-zero pairs from the CC_2 matrix (6,597,022 pairs).	1	0.025489	0.013871
	Non-zero pairs from the BC_2 matrix (2,164,063 pairs).	1	0.105450	0.073757

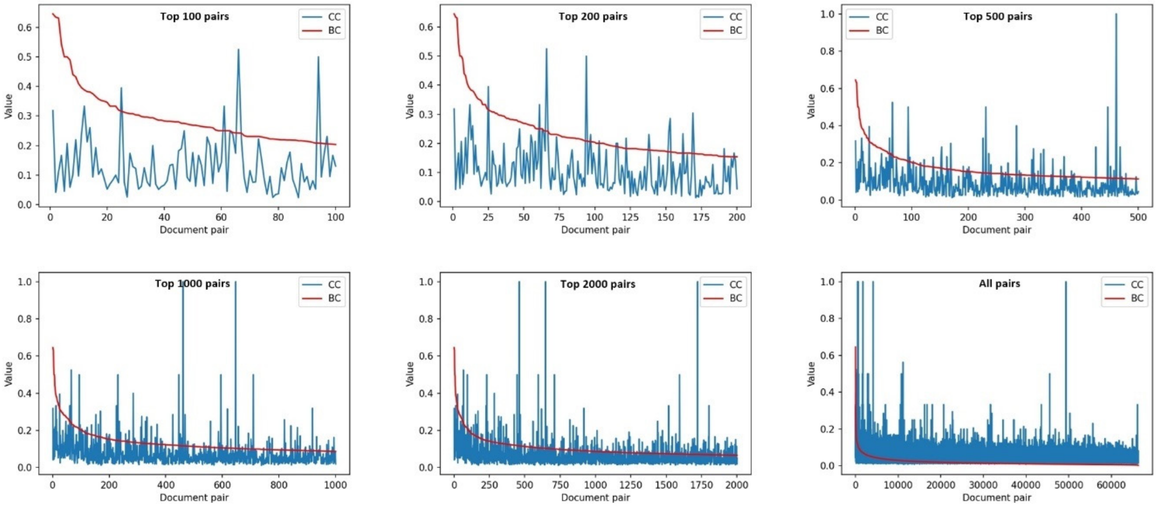

Figure 1: Dataset 1- BC (descending) vs CC -(a) top 100 pairs (b) top 200 pairs (c) top 500 pairs (d) top 1000 pairs (e) top 2000 pairs (f) all pairs.

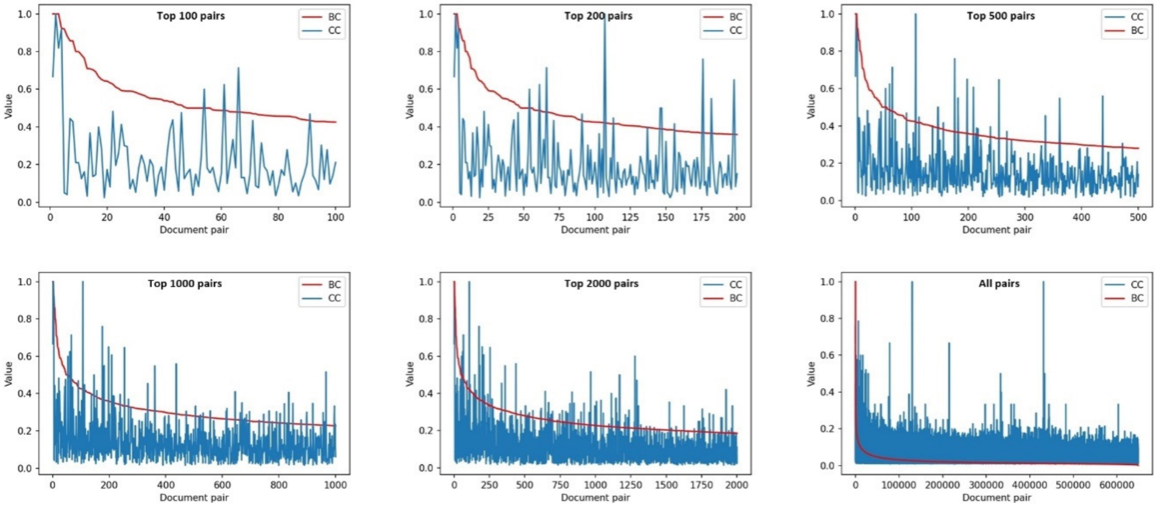

Figure 2: Dataset 2- BC (descending) vs CC-(a) top 100 pairs (b) top 200 pairs (c) top 500 pairs (d) top 1000 pairs (e) top 2000 pairs (f) all pairs.

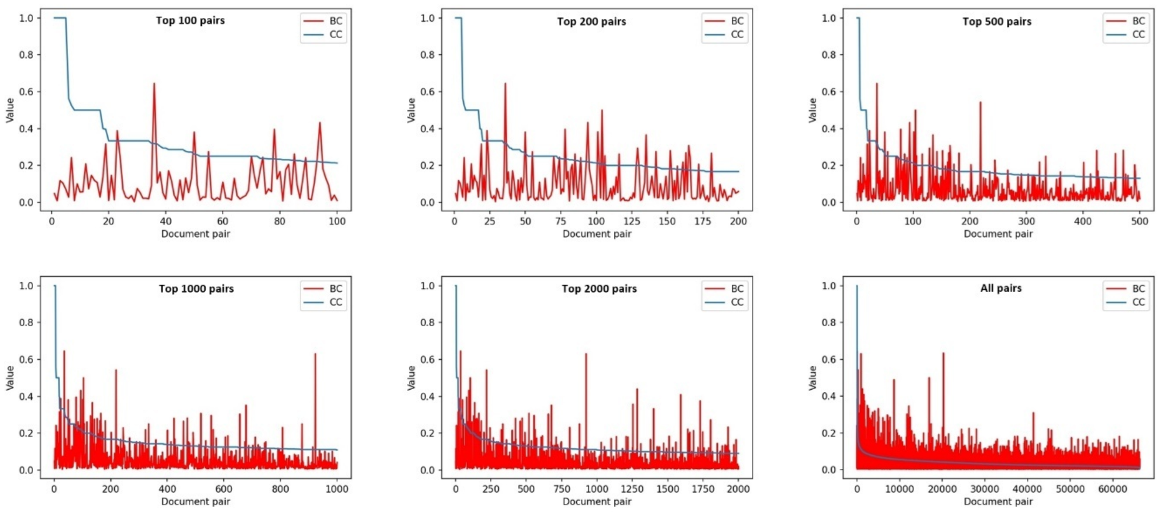

Figure 3: Dataset 1- CC (descending) vs BC-(a) top 100 pairs (b) top 200 pairs (c) top 500 pairs (d) top 1000 pairs (e) top 2000 pairs (f) all pairs.

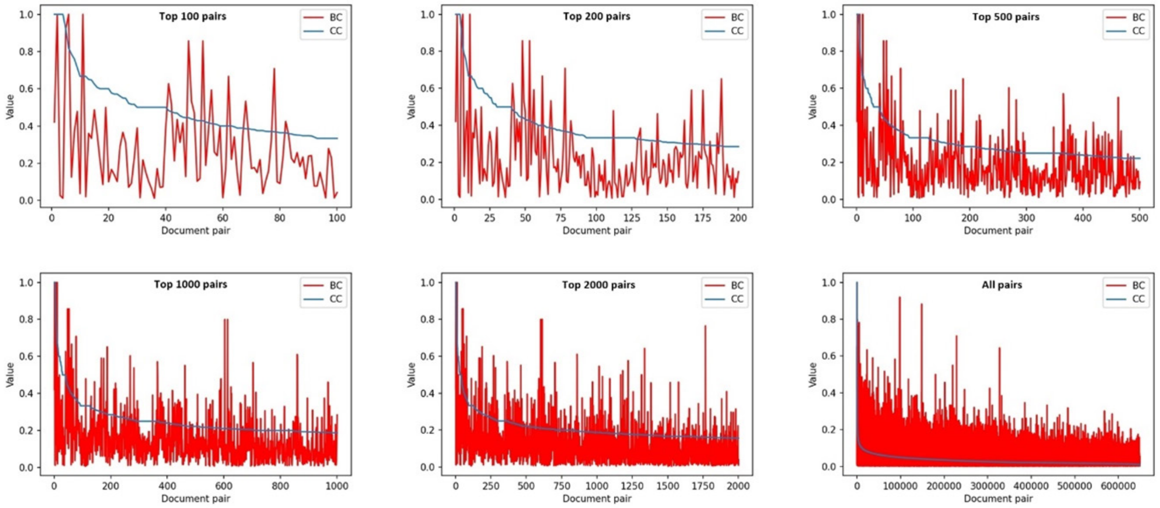

Figure 4: Dataset 2- CC (descending) vs BC-(a) top 100 pairs (b) top 200 pairs (c) top 500 pairs (d) top 1000 pairs (e) top 2000 pairs (f) all pairs.

Table 5: Correlation between matrices M_1 and M_2 for D_1 and D_2.

Dataset	No. of Intersecting Pairs	Rank Correlation between Bibliographic coupling and Conceptual coupling			
		Spearman		Kendall	
		Rho (ρ)	$-log_{10}(p)$	Tau (τ)	$-log_{10}(p)$
D_1	66,341	0.02001	6.5941	0.01445	7.6245
D_2	649,156	0.02220	70.8604	0.01656	88.4245

was to find whether papers with high bibliographic coupling have higher conceptual similarity, the zero-value pairs were dropped from further analysis.

This process also contributed to reducing the dataset to approximately one-third of its original size, resulting in significant memory savings. Subsequently, the next step involved identifying intersecting pairs of non-zero entries in the BC and CC matrices. A comprehensive description of the obtained matrices is provided in Table 3. The BC and CC values were visualized by arranging one of them in descending order. The top 1000 values of the BC and CC matrices were also identified and visualized to understand the relationship between them.

The rank correlation coefficients of the intersecting pairs in the BC and CC matrices were also computed using Spearman's and Kendall's methods.

$$Spearman's\ \rho = 1 - \frac{6\sum_i^n d_i^2}{n(n^2-1)}$$

$$Kendall's\ \tau = \frac{no.of\ concordant\ pairs - no.of\ discordant\ pairs}{\frac{n(n-1)}{2}}$$

where n = total number of intersecting article pairs in the dataset, and d = differnce of ranks.

The computations for the study were performed using Python (ver. 3.10) and various Python libraries NumPy (ver. 1.24.3) and SciPy (ver. 1.11.1). The similarity calculation functions were written as custom scripts and the large matrices for the Jaccard similarity computation was stored using the NumPy library. (The scripts can be shared upon request).

RESULTS AND DISCUSSION

The BC and CC values of all paper pairs were computed using the BC and CC matrices, based on the equations by Sen and Gan.[4] The highest value, mean value and standard deviation value of the bibliographic coupling strength amongst the non-zero pairs for the two datasets are shown in Table 4. It can be observed that most of the bibliographic coupling scores lie near the mean, with a standard deviation of 0.016127 in the case of D_1 and 0.073757 in the case of D_2. A similar occurrence is found in the case of conceptual coupling, where the data has only 0.030057 and 0.013871 standard deviations from the mean for D_1 and D_2 respectively. Thus, there are not very large numbers of pairs in the datasets with high bibliographic coupling or conceptual coupling value. Hence, it may be a better idea to arrange the pairs in descending order of BC and CC values each and look at some top valued pairs (say 1000).

The relationship between bibliographic coupling and conceptual coupling is visualised and interpreted as follows: First, the top 1000 BC pairs from D_1 are selected and their BC values are plotted in descending order (Figure 1). For each pair in this Figure 1, the CC value for the pair is also shown. The same process is then repeated for dataset D_2 (Figure 2). It may be observed that there is a lack of clear congruence between the BC and CC values of the different article pairs.

Next, we try to look at the relationship from the other side, i.e., to select the top 1000 conceptually coupled papers and see the bibliographic coupling between them. Figure 3 plots the top 1000 pairs arranged in descending order of CC values for the dataset D_1. The corresponding BC values for the article pairs are also shown. Figure 4 presents a similar plot for the dataset D_2. In both these cases too, no congruence is observed suggesting that pairs with high conceptual coupling need not have high bibliographic coupling. The patterns observed indicate fluctuating difference values, indicating no congruence between the BC and CC values of the given article pairs. This result aligns with the conclusion drawn by Martyn, who asserted that two bibliographically linked papers may not necessarily share similar semantics or concepts,[9] as is evident in our case.

Finally, we calculate the Spearman Rho (ρ) and Kendall Tau (τ) rank correlation coefficients between bibliographic coupling rank and conceptual coupling rank of the intersecting pairs.[23,24] These 2 measures are useful to compare the consistency and both provide quantitative measures of the monotonic relationship between two sets of similarity matrices.[25] This way, we can measure how linearly the two coupling values relate, based on the ranked coupling scores. Table 5 presents the value of the 2 rank correlation coefficients for both datasets. In summary, when Spearman's correlation coefficient and Kendall's tau values are low, around 0.01~0.02. This indicates a lack of agreement in the ranking between the two sets of similarity matrices. These findings indicate that the BC and CC values for the article pairings exhibit significant disparities. The result is reinforced by very small p-values, which suggest very strong evidence that the sample correlation coefficients represent the true correlations (since the p-values are extremely small, we have reported $-\mathrm{lo}\,g_{10}($ $p -$ value)'s for better interpretability).

CONCLUSION

The study revealed a weak correlation between the linking of bibliographies and the linking of concepts, as evidenced by Jaccard similarity and the concept fields from the Dimensions database.[3] This result aligns with the conclusion of Martyn, who asserted that two bibliographically linked papers may not necessarily share similar semantics or concepts,[9] as observed in our case. The notably low correlation values in these coupling approaches highlight the need for a more detailed examination of the underlying dynamics (Refer to Section 4). Our findings immediately call for a re-evaluation of the conceptual coupling of the academic literature network. The study delves into the concepts, derived algorithmically from the full text of the articles, using Dimensions data. Our analysis challenges conventional wisdom by proposing that two documents may possess substantial conceptual similarity despite a lack of extensive citation connections, even when they share a similarity in references. By employing this analysis, one can uncover concealed innovations in publications that are conceptually related but may not share similar references. We need to conduct further research to understand the factors contributing to this variation in coupling values and to pinpoint potential variations within specific disciplines, subjects, or publication years.

The study lays the foundation for subsequent investigations in this area. Conducting an in-depth analysis of document pairs that demonstrate substantial disparities between bibliographic and conceptual coupling may provide invaluable insights. This could facilitate a more comprehensive comprehension of the interplay between the content of scholarly articles and their cited references. Essentially, it can facilitate interdisciplinary research by identifying similar research areas across multiple subject disciplines through the comparison of publication pairs with high conceptual similarity and low citation similarity (which may not be cited together). It may also find possible applications in refining search in recommendation systems, enhancing literature review strategies, facilitating cross-disciplinary learning and innovating new strategies.

The work reported here, however, has certain limitations as well. It relies only on the concepts obtained from the Dimensions database as a representative of the thematic structure of research papers. Subsequent research endeavours may be enhanced by exploring more sophisticated techniques for concept extraction and contextualization, which may uncover latent parallels among bibliographically linked articles. Further, to fully capture the spectrum of relationships between documents, future investigations might be required to examine alternative similarity metrics in addition to Jaccard similarity.

ACKNOWLEDGEMENT

This work is partly supported by the extramural research grant provided by the HPE Aruba Centre for Research in Information Systems at BHU (No.: M-22-69 of BHU).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ABBREVIATIONS

BC: Bibliographic Coupling; **CC:** Conceptual Coupling; **WoS:** Web of Science; Other abbreviations are mentioned within the text wherever necessary.

REFERENCES

1. Kessler MM. An experimental study of bibliographic coupling between technical papers. IEEE Trans Inf Theory. 1963;9(1):49–51. doi:10.1109/TIT.1963.1057800.
2. Kessler MM. Bibliographic coupling between scientific papers. Am Doc. 1963;14(1):10–25. doi:10.1002/asi.5090140103.
3. Bichteler J, Eaton EA. The combined use of bibliographic coupling and cocitation for document retrieval. J Am Soc Inf Sci. 1980;31(4):278–82. doi:10.1002/ASI.4630310408.
4. Sen SK, Gan SK. A mathematical extension of the idea of bibliographic coupling and its applications. Ann Libr Sci Doc. 1983;30(2):78–82.
5. Sharabchiev JT. Cluster analysis of bibliographic references as a scientometric method. Scientometrics. 1989;15(1–2):127–37. doi:10.1007/BF02021804.
6. Jarneving B. A comparison of two bibliometric methods for mapping of the research front. Scientometrics. 2005;65(2):245–63. doi:10.1007/s11192-005-0270-7.
7. Jarneving B. Bibliographic coupling and its application to research-front and other core documents. J Informetr. 2007;1(4):287–307. doi:10.1016/j.joi.2007.07.004.
8. Park I, Jeong Y, Yoon B, Mortara L. Exploring potential R&D collaboration partners through patent analysis based on bibliographic coupling and latent semantic analysis. J Technol Transf. 2014;27(7):759–81. doi:10.1080/09537325.2014.971004.
9. Martyn J. Bibliographic coupling. J Doc. 1964;20(4):236. doi:10.1108/eb026352.
10. Weinberg BH. Bibliographic coupling: A review. Inf Storage Retriev. 1974;10(5–6):189–96. doi:10.1016/0020-0271(74)90058-8.
11. Sainte-Marie M, Mongeon P, Larivière V. Do you cite what I mean? Assessing the semantic scope of bibliographic coupling. In: Proceedings of the 23rd International Conference on Science and Technology Indicators (STI 2018). 2018. p. 649–57.
12. Uddin S, Khan A. The impact of author-selected keywords on citation counts. J Informetr. 2016;10(4):1166–77. doi:10.1016/j.joi.2016.10.004.
13. Vladutz G, Cook J. Bibliographic coupling and subject relatedness. In: Proceedings of the ASIS Annual Meeting, 21. 1984. p. 204–7.
14. Kessler MM. Bibliographic coupling extended in time: Ten case histories. Inf Storage Retriev. 1963;1(4):169–87. doi:10.1016/0020-0271(63)90016-0.
15. Kessler MM. Comparison of the results of bibliographic coupling and analytic subject indexing. Am Doc. 1965;16(3):223–33. doi:10.1002/asi.5090160309.
16. Braam RR, Wed HF, Van Raan AFJ. Mapping of science by combined co-citation and word analysis. II: Dynamical aspects. J Am Soc Inf Sci. 1991;42(4).
17. Glänzel W, Czerwon HJ. A new methodological approach to bibliographic coupling and its application to research-front and other core documents. In: Proceedings of the Fifth Biennial International Conference of the International Society for Scientometrics and Infometrics. 1995. p. 167–76.
18. Li H, Wu M, Wang Y, Zeng A. Bibliographic coupling networks reveal the advantage of diversification in scientific projects. J Informetr. 2022;16(3):101321. doi:10.1016/j.joi.2022.101321.
19. Habib R, Afzal MT. Sections-based bibliographic coupling for research paper recommendation. Scientometrics. 2019;119(2):643–56. doi:10.1007/s11192-019-03053-8.
20. Liu RL. Passage-based bibliographic coupling: An inter-article similarity measure for biomedical articles. PLoS ONE. 2015;10(10). doi:10.1371/JOURNAL.PONE.0139245.
21. Liu RL. A new bibliographic coupling measure with descriptive capability. Scientometrics. 2017;110(2):915–35. doi:10.1007/s11192-016-2196-7.
22. Ahlgren P, Colliander C. Document–document similarity approaches and science mapping: Experimental comparison of five approaches. J Informetr. 2009;3(1):49–63. doi:10.1016/J.JOI.2008.11.003.
23. Spearman C. The proof and measurement of association between two things. Am J Psychol. 1904;15(1):72. doi:10.2307/1412159.
24. Kendall MG. A new measure of rank correlation. Biometrika. 1938;30(1/2):81. doi:10.2307/2332226.
25. Schober P, Schwarte LA. Correlation coefficients: Appropriate use and interpretation. Anesth Analg. 2018;126(5):1763–8. doi:10.1213/ANE.0000000000002864.

Cite this article: Nandy A, Singh A, Gupta V, Singh VK. Bibliographic Coupling and Conceptual Similarity: Are the Bibliographically Coupled Papers also Conceptually Similar?. J Scientometric Res. 2024;13(3):706-14.