OPEN ACCESS
International Journal of Innovations in Science & Technology

Experimental Evaluation of Related Papers Finding Techniques

Muhammad Ammad Idrees[1], M. Abdul Qadir[2], Maryam Nageen[1]
[1]Dept. of Artificial Intelligence Capital University of Science and Technology.
[2]Dept. of Computer Science Capital University of Science and Technology.
*Correspondence: sardarammad143@gmail.com

Citation | Idrees. M. A, Qadir. M. A, Nageen. M, "Experimental Evaluation of Related Papers Finding Techniques", IJIST, Special Issue pp 403-417, May 2026
Received | March 19, 2026 **Revised |** April 30, 2026 **Accepted |** May 06, 2026 **Published |** May 11, 2026.

Introduction/Importance of Study: Related-paper recommendation systems generally contain two categories: Content-Based (CB) approaches, which estimate relatedness using semantic similarity between research paper texts, and metadata-based approaches, which infer relatedness from bibliographic information such as citations, references, authorship, and publication venue. Although CB methods, such as Jensen–Shannon Divergence (JSD), computed over TF–IDF representations provide accurate relatedness scores, they are computationally expensive because they require processing of the full text of each paper. Metadata-based methods offer a more efficient substitute, but their relative effectiveness to strong CB measures remains unclear. This study investigates which bibliometric technique correlates most strongly with JSD-based semantic relatedness to identify a low-cost substitute for computationally expensive CB methods.

Novelty Statement: Since no existing dataset contained the required combination of full text, citations, references, and "related papers" lists, we constructed a new dataset of 1,225 papers, selected to statistically represent the population for a target keyword at 95% confidence with ±2.8% margin of error. No prior research has analyzed bibliometric methods using a unified dataset.

Material and Method: JSD-based relatedness scores were computed using full-text TF–IDF representations for all papers. We then calculated bibliographic relatedness using bibliographic coupling (BC), co-citation coupling (CC), and Katz similarity, and also extracted relatedness scores from Semantic Scholar (SS).

Result and Discussion: Correlation analysis revealed the following Pearson correlations with JSD: BC = 0.40, SS = 0.35, Katz = 0.01, CC = –0.11. These results indicate that BC-based relatedness aligns most closely with CB semantic similarity, followed by SS, while Katz and CC show negligible or negative correlation. Notably, the finding that Semantic Scholar's related-paper measure correlates less strongly with JSD than bibliographic coupling is both surprising and practically important.

Concluding Remarks: Overall, the results highlight the potential of BC-based methods as an efficient and reliable alternative to traditional full-text similarity computations for estimating relatedness.

Keywords: Paper Recommendation; Bibliometric Technique; Semantic Similarity; Jensen-Shannon Divergence; Citation-Based Technique.

Introduction:

Paper recommendation systems are platforms that share the same input papers and return semantically or contextually similar papers to each other. Existing methods for searching relevant research articles are content-based and metadata-based [1]. The search for relevant publications is integral to the evolution of knowledge and science. This problem has gained much attention in the last 20 years. Due to this scientific attention, many proposed methods and strategies have been suggested to address the challenge of finding relevant papers [2]. Existing methods for finding relevant research articles: are content-based and metadata-based approach. The three widely used strategies are co-citation, bibliographic coupling, and Katz similarity. Although bibliographic coupling is based on papers that are referenced by many articles, Co-citation analysis is based on papers that are cited together by the same citing documents [3], and Katz reveals direct and indirect connections [4]. Bibliographic coupling and co-citation are the basic procedures used for the development of citation analysis, the most widely used means to identify relationships between scholarly documents for many years [5]. The accelerating pace of scientific publications makes it much harder to identify suitable research papers. The most crucial research in a field can no longer be tracked effectively with traditional methods such as keyword searching and citation counts [6]. Traditional methods for recommending related papers have generally made use of individual bibliometric techniques.

Content-based recommendation systems usually conduct textual analysis of research papers, including title, abstract, keywords, and full texts, to elicit semantic representations. Such as TF–IDF [7], word embeddings, topic modeling, and semantic distances. Topic modeling has emerged as a powerful probabilistic approach for uncovering the underlying thematic structure within large and unstructured text, providing a robust foundation for enhancing research paper recommendation systems [7]. Feature-based methods form a foundational approach in research paper recommendation systems, operating on the principle that scholarly documents can be effectively represented by a set of discriminative attributes, or features, which can be quantitatively analyzed to determine similarity [8].

Semantic Scholar (SC) used a graph-based recommendation technique. SC was created to combat the ever-increasing number of scientific publications that exist by constructing a massive representation of research, or the literature graph. These nodes and edges, each of which has specialized properties, collectively support sophisticated algorithms that can automatically find relationships between research works [9]. The citation recommendation model that leverages deep language understanding and graph-based learning to enhance related research paper discovery. They utilized a model that incorporates Bidirectional Encoder Representations from Transformers (BERT), contributing to the semantic context of the citation blasts, and a Graph Convolutional Network operating on the citation network for learning structural relationships between papers [10]. MIReAD (Minimal Information for Representation Learning of Academic Documents) specifically addresses the creation of document-level representations of academic documents. Unlike other models that are reliant on citation graphs like SPECTER or CiteBERT, MIReAD simply learns embeddings based on title and abstract alone, making it especially appropriate for recently published or understudied works [11]. The growing number of scientific publications has led to a necessity for accurate methods for the classification and organization of research articles, especially if full-text access is unavailable. Metadata-based classification offers a good practical solution, as one can use the free information like: full document titles, abstracts, keywords, and general terms to categorize documents into specific domains of interest [12].

This study addresses the limitations of existing approaches and analyzes which bibliometric approach performs better to find semantically related papers. It investigates the results on a single dataset and which bibliometric method: bibliographic coupling, co-citation, and Katz similarity performs the best with JSD. Jensen-Shannon Divergence (JSD) as a content-based ground truth to provide a more objective evaluation of citation-graph techniques. The method that correlates most strongly with JSD-based semantic relatedness is also observed to identify a low-cost substitute for computationally expensive content-based methods. Based on bibliographic coupling, co-citation, and Katz's similarity, the correlation is found with JSD. It evaluates the Semantic Scholar-related papers with JSD, a method that performs well to find semantically related papers, such as co-citation, which is backward-looking and relies on future citations, whereas bibliographic coupling is forward-looking but sensitive to old references. The Katz similarity is included to be able to discover indirect but semantically grounded relations.

Research Objectives:

This study pursues the following specific objectives:

Build a cleaned, normalized corpus with structured references, citations, and related.

Compute BC, CC, and Katz scores; construct a JSD-based semantic proxy from TF-IDF text representations; normalize scores for comparability.

Use a representative sample to quantify correlations between each technique and JSD; select a production scorer based on evidence.

Novelty:

Existing studies utilize bibliometric methods such as BC, CC, and Katz similarity to identify related articles; no one has conclusively determined which technique best captures semantic relatedness. The dataset we used in this study is the combination of full text, citations, references, and "related papers" lists, which demonstrates its uniqueness. This dataset encompasses 1,225 papers, selected to statistically represent the population for a target keyword at 95% confidence with the error margin ±2.8%. Our research has been conducted to analyze bibliometric methods by using a unified dataset that efficiently covers the gap.

Literature:

The increasing scholarly literature has necessitated the development of automated systems to help researchers efficiently discover related work [13]. As a result, research paper recommendations have now become an important topic of interest in information retrieval and digital libraries. The whole literature can be grouped into three high-level paradigms: content-based, citation, and hybrid [2]. To aid retrieval, we can rely on content-based methods that analyze the text of papers, citation-based techniques that exploit bibliographic structures, and hybrid systems that attempt to integrate these signals for a better outcome [14].

The content-based systems work on the idea that papers with similar content would have comparable textual details. One common approach is to extract features (keywords, phrases, or embeddings) from the given document text (titles, abstracts, or full text) and compute similarity between these representations [15].

Key phrase extraction approach utilizing pre-trained language models (PLMs) that integrate an attention mechanism with semantic similarity. They demonstrate the limitations of conventional extraction models in modeling a sense of context at a deeper level and also show superior performance across complex documents. Furthermore, graph-based ranking and phrase-document similarity methods have been incorporated into PLMs for better performance in key phrase extraction, which has the advantages of not requiring labeled data to train the model and, therefore, can be applied to low-resource conditions and domain-specific tasks [16].

At the neural level, a study conducted a comprehensive survey of deep learning in AKE. Their research highlights the development of AKE from traditional ML methods towards state-of-the-art neural networks such as RNNs, CNNs, and autoencoders. The paper highlights both the extractive (from the corpus) and abstractive (to generate unseen phrases) perspectives. It also highlights problems such as the identification of semantically redundant key phrases and provides future directions in extraction improvements and hybrid solutions through a combination of supervised and unsupervised learning approaches [17]

An unsupervised key phrase extraction model: AdaptiveUKE, which improves upon the use of a gated topic model to capture semantic diversity. It infers topics related to the richness of the document independently and uses an independent scoring strategy for topic relevance and inter-relationship, ensured by relevance and diversity in extracting key phrases. Experiments on datasets, such as the Inspec and SemEval2010, etc., show that the model outperforms some state-of-the-art baselines. This work indicates the importance of changing behavior due to topic differences to extract information better from real-world documents [18].

A hybrid recommendation system, which is a combination of content-based filtering and collaborative filtering to reduce the downsides of every method; while it is conceptually logical to amalgamate these two methods, its implementation usually lacks the complexity and novelty as compared to more recent works, thus helping with cold-start or data sparseness issues. Similarly, the adoption of primitive feature engineering (e.g., traditional cosine similarity) in the content-based module is another potential but missed opportunity to replace with more advanced semantic representations using state-of-the-art NLP technologies like BERT or Doc2Vec to enhance contextual understanding and personalized results. Additionally, the collaborative filtering part shines at leveraging user-item interaction history in an efficient manner, but neglects changing preferences of users or the fact that human behaviors are not static across time, which are both crucial aspects when applied to large-scale real-world problems. The empirical validation claiming competitive results is not well scaled and barely diverse, which is conducted on only one dataset without a rigorous comparison with the deep learning-based state-of-the-art models. The mosaic style of the paper is praiseworthy for its clarity and practical relevance, although there is no mention of computational efficiency, scalability, and domain transferability, which diminishes its applicability to large-scale settings. In general, although the proposed hybrid model is a practical and rule-based approach toward better recommendation accuracy, it lacks the integration with the prevalent data-driven shape of contemporary state-of-the-art recommender systems [19].

Co-citation analysis is an important technique for the identification of intellectual structures and thematic developments in science. The fundamental statement that two papers cited together by films are conceptually related is rather powerful, and following recent citations by other films, probably even more so, and certainly appealing from an intuitive point of view. The work is strongest in methodological clarity and historical context, especially in the development of co-citation from raw frequency counts to network-based visualizations, clustering, etc. Nevertheless, despite its historical and theoretical strength, the paper exhibits severe limitations when viewed from the perspective of contemporary recommender systems. It is still reliant on the rate of accumulated citations over time, which is inherently not suitable for recently published or niche research with few citations. Also, the binary co-citation linkage between papers of the model cannot express the deep semantic similarity between papers, such as subtle differences in representative contents or citation context. The paper also fails to discuss algorithm scalability for large citation networks, as well as the incorporation of other graph-based or semantic techniques, e.g., co-citation with topic models, or graph neural networks. ca) Further disadvantage is that there is no personalized user modeling, and it cannot be directly used in the adaptive research recommendation environment. While co-citation continues to play an important role in bibliometric research, this study does not attempt to predict its future development to support current needs for real-time, user-centered scholarly discovery and interaction in a digital age of such dynamic content, rapid publishing cycles, and complex, interdisciplinary relationships [20].

Co-citation analysis, inscribing it into a robust material to reveal intellectual structures and links between themes in scientific literature, although offering a concise theoretical discussion and arguing co-citation's worth in creating maps of research fronts, this article fails to respond to major limitations preventing co-citation from becoming an important component of modern recommendation systems. The historical citation frequency-based approach has a natural limitation in that it omits the new research that is very little cited, and it does not take the citation context or polarity into consideration, e.g., whether the citation is positive, neutral, or negative.

Moreover, recent computational methods like graph embedding, dynamic citation networks, or hybrid models that account for content semantics with structural link semantics have not been discussed adequately. The paper thus fails to consider the demands of real-time recommendation, personalization, and user-interaction, which need to be treated as deployment requirements in digital scholarly environments. Although the work does not delineate a scalable or user-responsive manner in which to apply this providential potential of co-citation as an academic discovery mechanism at scale, it highlights the ongoing bibliometric utility [21].

TATKC: a temporal graph neural network that is based on learning approximate Katz scores in the dynamic networks. This model performed well in terms of prediction while adapting a time-heterogeneous graph structure. Embedding node proximity into Katz similarity is often tricky, and they instead seamlessly integrated GNNs with Katz, giving more contextualized embeddings. Nevertheless, the interpretability of the model is still limited, and it comes at a high price in training cost [22], nonetheless.

Katz score updates will occur quickly after we modify the graph. They provided formulas for analytical updates to the walk-count and communicability measures that you can increment or decrement without having to recompute from first principles. This made Katz even suitable for citation networks that are re-plotted at an often time period. Although this model showed more computational power, it focused on structural efficiency and did not improve semantic quality and content-based filtering [23].

A multi-stage recommender that uses clustering to group papers, then a graph-based model is built to represent structural relationships between them, which are fed to a deep learning algorithm for feature integration, reaching high recall and NDCG performance [24].

A temporal-GNN that allows recurrent updating of paper embedding when new citations appear, which will support proper recommendations on an evolving literature. Proposed a Temporal graph neural network for modeling co-author relationships with respect to evolving. These models provide timely recommendation which is contextually more relevant by capturing how co-authorship patterns shift, which addresses some limitations of the static network [25].

A bioinformatics-specific co-author recommender that weighted the recommendations according to how often people published together and how close they are in the domain. Representing deep-textual author influence and collaboration using their model, implemented in Spark and Scala, yielded 98% precision. This step shows the qualifying collaborative strength in the co-authorship networks [26].

Through meta-path frameworks, hybrid models started to combine co-authorship with citation, topic, and venue data. The combination of heterogeneous network analysis with machine learning improves recommendation accuracy and addresses the cold-start issue. In particular, multi-modal meta-path-based recommendation systems demonstrate the highest potential for personalized and context-aware paper discovery [27].

Synthesis and Identified Research Gaps:

The literature shows that the vigorous revolutions, as well as several gaps, are interconnected with this study. A common pattern across all three paradigms is the use of different, often non-overlapping datasets for evaluation, from small, domain-specific sets like the CiteSeer subsets to large, field-specific corpora like TREC Genomics or massive multi-domain graphs like MAG. This fragmentation makes a unified, fair comparison of core techniques (BC, CC, Katz, content models) practically impossible.

Secondly, while the efficiency of citation-based methods is frequently cited, and the accuracy of content-based methods is pursued, there is a striking absence of research that quantitatively correlates these efficient citation metrics with a robust, content-based semantic ground truth on a large scale. It remains an open empirical question which citation technique best approximates true topical relatedness as referred to by text.

Thirdly, academic research proposals are seldom benchmarked against the relatedness algorithms of major which deployed services like Semantic Scholar. Understanding how transparent bibliometric measures perform relative to these proprietary industrial systems is of significant practical value but is rarely investigated.

Finally, these gaps are exacerbated by the absence of a purpose-built, publicly available Dataset containing the triad necessary for such a study: full text (for content analysis), complete citation/reference graphs (for BC/CC/Katz), and reliable relatedness findings. This work aims to fill these gaps by constructing a new dataset and using Jensen-Shannon Divergence on text as a semantic benchmark to empirically evaluate and rank citation-based techniques.

Materials and Methods:

This study experimentally evaluates different related paper finding techniques by computing similarity scores using BC, CC, Katz, and Semantic Scholar, and comparing them against the semantic similarity measured using JSD. Citation-based recommendation systems using citation exploit the structure of scholarship citations to detect relations between research papers. Such systems use shared references (bibliographic coupling), co-cited references (co-citation analysis), and citation graph paths (e.g., Katz similarity) among other types of patterns to obtain conceptual or intellectual links between documents. Co-citation (CC) is a fundamental bibliometric method that computes the similarity of two documents according to the frequency of co-citation by other documents. In Figure 1.1, two or more documents (e.g., p3 and p4) are co-cited by another document, p1; they share some kind of relationship in terms of content or context. Bibliographic Coupling (BC) considers two documents similar if they are referenced by the same set of papers. In Figure 1.1 p13 cites (p11, p12) p14 cites (p11, p12) and p15 cites (p11, p12) paper p13, p14, p15 are bibliographically coupled with strength 2. Katz's similarity is a fundamental approach in network science and graph-based recommendation systems. Katz's similarity takes into account all paths between nodes (including the direct links), weighted by a damping factor which decreases the relevance of edges traversed in more steps [28]. In Figure 1.1, p13 cites p11 and p11 cites p10, so p13 can't cite p10 directly but is connected with a 2-hop distance, that means p13 and p10 are also related documents.

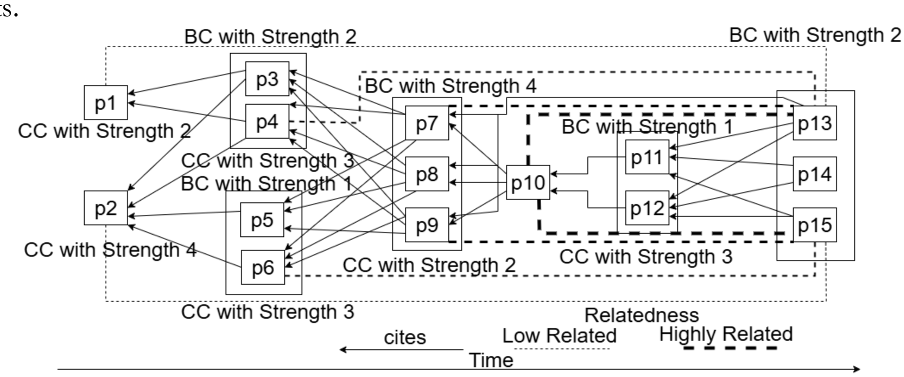

Figure 1. Citation Graph of CC, BC, and Katz Similarity Relations

Dataset Construction and Description:

A dataset was systematically constructed to enable the correlation analysis at the core of this research. The primary objective was to compile a statistically valid sample of the machine learning literature, from which both content-based (JSD) and citation-based (BC, CC, Katz) similarity scores are calculated between many pairs of papers. This database is the fundamental template for assessing which bibliometric measure is most similar to semantic relatedness. We chose Semantic Scholar for primary inspiration due to its wide coverage of computer science literature, extensive structured metadata, and the availability of a "related papers" feature, which allows us to measure it against the external standard.

Statistical Sampling Methodology:

The machine learning literature corpus is approximately N = 6.64M papers. For this reason, a statistically comprehensive sampling method was used to guarantee the representativeness of the database and generalizability of results. The sample size is calculated using the formula 'N' for a finite population for a given level of confidence (95%) and margin of error. The most conservative proportion (p=0.5) was utilized so that the sample size could be the largest possible to make firm analyses. An example of how the calculations were done to obtain these numbers, about a confidence level of 95% (Z = 1.96) and a margin of error of ± 2.8% (e = 0.028). This calculation indicated that a minimum of 1,225 papers was needed in order to attain the desired level of precision.

Sample Size Formula:

$$n_0 = \frac{Z^2 .p(1-p)}{e^2} = \frac{1.96^2 .0.5(1-0.5)}{0.028^2} \approx 1225$$

$$n = \frac{n_o}{1+\frac{n_o-1}{N}} = \frac{1225}{1+\frac{1225-1}{6600000}} \approx 1224.77$$

Proposed System:

The proposed study follows an organized data-processing workflow shown in Figure 1.2.

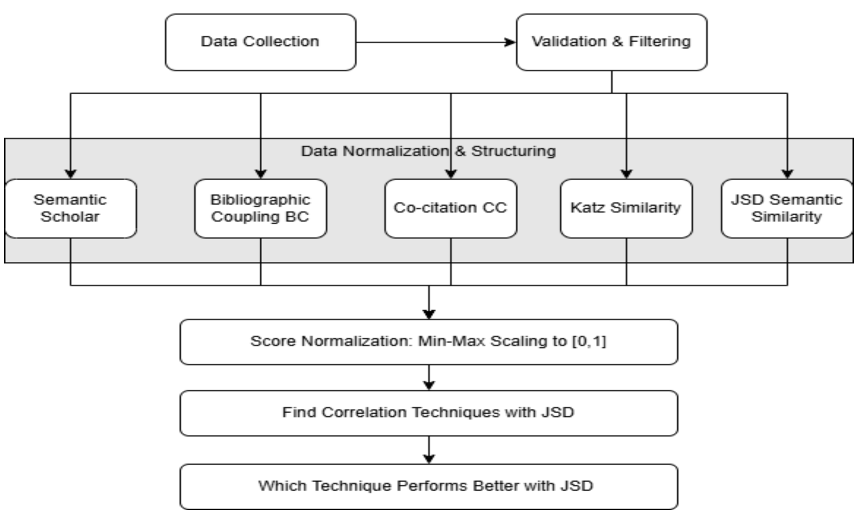

Figure 2. System Flow Diagram

In Figure 1.2, the first step is to collect data from the Semantic Scholar website, and the next step is to validate and apply filtering on the collected data. Then the data is converted into a normalized form, with all values in between 0 and 1, by using min-max scaling. The dataset is arranged in normalized form and finds correlation between all techniques, like CC, BC, Katz, and Semantic Scholar, with JSD. Finally, find the results for which techniques perform better with JSD.

This work takes a systematic, data-driven study to evaluate and compare the effectiveness of citation-based techniques for discovering semantically similar papers. The primary objective is to determine which bibliometric technique gives the Semantically similar papers. Here, we outline the design and implementation of such an analytic framework. All methods included the following common preprocessing steps:

For Content-Based (JSD) and Semantic Scholar (SS): The paper titles and abstracts are used to combine to create a document corpus. Text underwent standard NLP preprocessing: tokenization, lowercasing, stop-word removal, and lemmatization using the NLTK library in Python.

For Citation-Based Techniques (BC, CC, and Katz): A global directed graph was constructed as $G = (V, E)$, if the paper. V_i cites paper V_j then the set of all the unique papers, and also the directed edge is $e_{ij} \epsilon E$ exists are represented by V. This graph was stored as a sparse adjacency matrix A to facilitate the efficient computation on a large scale.

Content-Based Ground Truth: Jensen-Shannon Divergence (JSD): The JSD is employed as the semantic ground truth due to its theoretical robustness and established high correlation (90%) with human judgments of topical relatedness [29]. For each preprocessed document, a TF-IDF vector was fitted to create a term-document matrix. Then each document vector was normalized by the sum to 1, which acts as a probability distribution P over the vocabulary. The JSD of the distributions P and Q of two papers is given as follows:

$$JSD\ (P \| Q) = \frac{1}{2} DKL(P \| M) + \frac{1}{2} DKL(Q \| M)$$

Where $M = \frac{1}{2}(P + Q)$ *and* DKL is the Kullback-Leibler divergence, since JSD is a measure of dissimilarity, it is converted to a similarity score for consistent comparison: $JSD_{Sim}\ (P, Q) = 1 - JSD\ (P \| Q)$. Higher values of JSD_{sim} indicate greater semantic similarity.

P = Probability distribution of paper 1
Q = Probability distribution of paper 2
M = Average distribution
DKL = Kulback-Leibler divergence

Bibliographic Coupling (BC): BC measures the overlap in the reference lists of two papers, operating on the premise that shared citations indicate shared intellectual foundations [30]. For papers i and j with reference sets R_i and R_j, the raw BC strength is $BC_{raw}(i, j) = |R_i \cap R_j|$.

Co-citation (CC): CC quantifies the frequency with which two papers are cited together in the following research, characterizing their relationship from the perspective of researchers. For papers i and j with Citation paper sets C_i and C_j, the raw CC strength is $CC_{raw}(i, j) = |C_i \cap C_j|$.

Katz Similarity: Katz similarity is a global graph measure that considers all possible paths between two nodes in a network, weighting shorter paths more heavily. It captures indirect scholarly influence beyond direct citation or shared references. The Katz similarity between nodes i and j in the citation graph with adjacency matrix A is defined as:

$$Katz(i, j) = \sum_{i=1}^{\infty} \boldsymbol{\beta}^i (A^i)\text{ij}$$

Where β is a damping factor $(0 < \beta < 1/\lambda_{max})$.

Semantic Scholar (SS) Relatedness: As an external baseline, we obtained the Semantic Scholar API provided relatedness score for each pair of papers in the sample. This is the output produced by Semantic Scholar's in a large-scale recommendation algorithm, which could contain a mix of content-based, as well as other features [31].

Normalization and Comparative Framework:

To contrast these five types of non-stationary features, all had different scales and distributions, we normalized the valuations of these scores following a normalization using the min-max technique on [0,1], 1 being the maximum related and zero indicating no relation for this method:

$$Score_{norm} = \frac{Score_{raw} - Min(Score)}{Max(score) - Min(score)}$$

$Score_{norm}$ = The normalized score is always between 0 and 1.
$Score_{raw}$ = The original value before normalization.
$\min(score)$ = The smallest value in the dataset.
$\max(score)$ = The largest value in the dataset.

Min-max normalization was selected because it preserves the relative relationships between similarity scores while ensuring all values fall within the range [0,1]. Without normalization, values are not comparable.

Correlation Analysis and Evaluation:

The main evaluation measure is the Pearson correlation coefficient P between each citation-based similarity score (BC, CC, Katz, SS) and the JSD-based semantic similarity. For n paper pairs, the correlation rXY between method X and $JSD\ (Y)$ is computed as:

$$rXY = \frac{\sum_{i=1}^{n}(xi - x^{-})(yi - y^{-})}{\sqrt{\sum_{i=1}^{n}(xi - x^{-})^2 \sum_{i=1}^{n}(yi - y^{-})^2}}$$

Where:

x_i = Normalized similarity score obtained from a bibliometric technique (e.g., BC, CC, Katz, and S_Scholar) for the ith paper pair.
y_i = Semantic similarity score (JSD) for the ith paper pair.
$\overline{x}$= Mean of all x_i values.
$\overline{y}$= Mean of all y_i values.
n = Total number of paper pairs in the dataset.

In this study, each observation corresponds to a pair of papers identified by Main Paper ID and Related Paper ID. Where similarity scores are computed using different techniques. The Pearson correlation coefficient is computed separately for each technique by taking x_i as the normalized score of the respective method (BC, CC, Katz, and S_Scholar) and y_i as the corresponding JSD-based similarity score.

Results and Discussion:

The main goal of this study was to find which relation measure based on reference counting is the most highly correlated with semantics or content relations and to validate their results. Based on a statistically sound sample of 1,225 papers, the results provide clear insights. The dataset is then used to find the correlation of all techniques with JSD. The Correlation results given in Table 1.1 BC, CC, Katz, and Semantic Scholar are correlated with JSD.

Table 1. Pearson Correlation (r) of Techniques with JSD

Technique	Correlation with JSD	Interpretation
Bibliographic Coupling	0.40	Moderate Positive Correlation
Semantic Scholar	0.35	Fair Positive Correlation
Co-citation	-0.11	Slight Negative Correlation
Katz similarity	-0.01	No Correlation

The data reveal a distinct hierarchy of effectiveness: As shown in Table 1.1, Bibliographic Coupling (BC) is the best-performing proxy with moderate positive correlation (r = 0.40) to semantic similarity. This implies a regular, systematic relationship in which papers with more references also typically have similar themes. Semantic Scholar relatedness score also exhibits a strong positive correlation (r = 0.35) and supports its applicability as a recommendation method. And CC and Katz similarity exhibit little semantic consistency. The weak feature for CC implies that, in terms of machine learning, the corpus is such that pairwise frequently cited papers are actually supply-side references rather than demand-side references. However, on this data set, it is empirically outperformed by the simpler transparent BC measure.

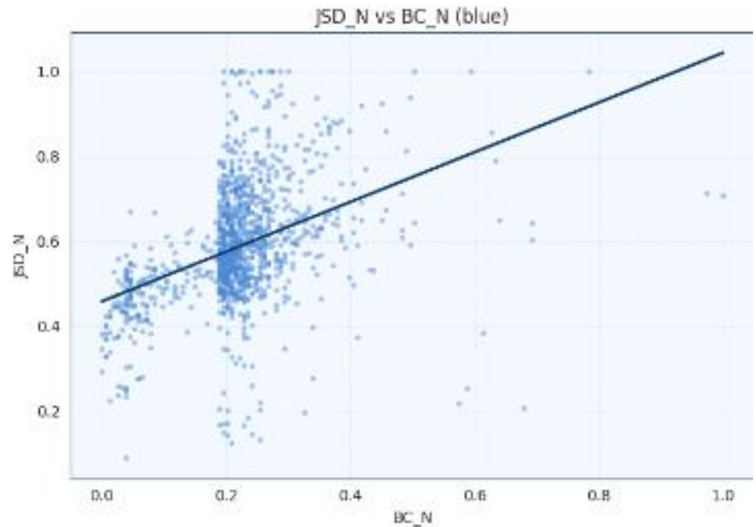

Figure 3. Bc vs JSD

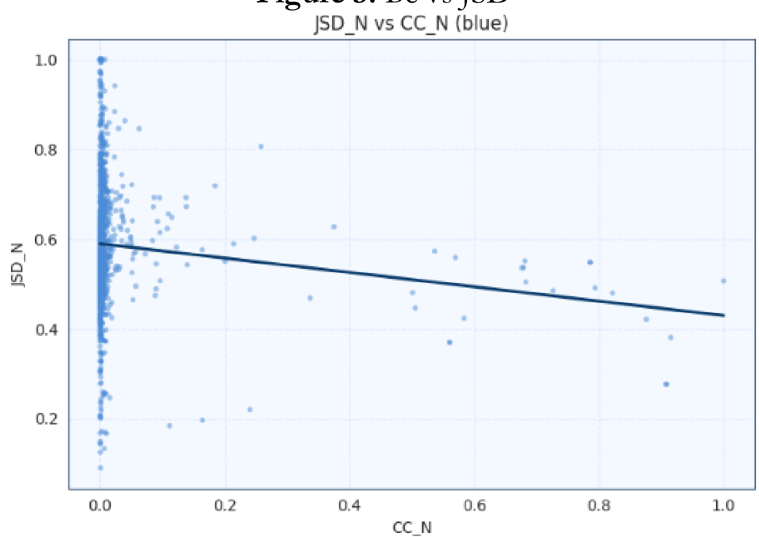

Figure 4. CC vs JSD

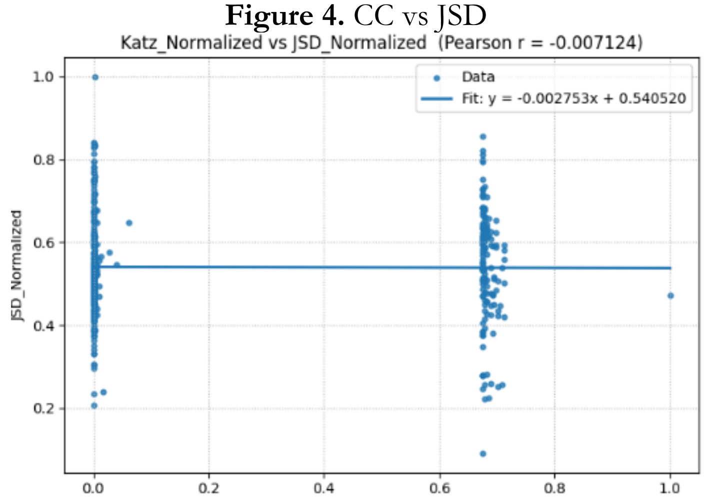

Figure 5. Katz vs JSD

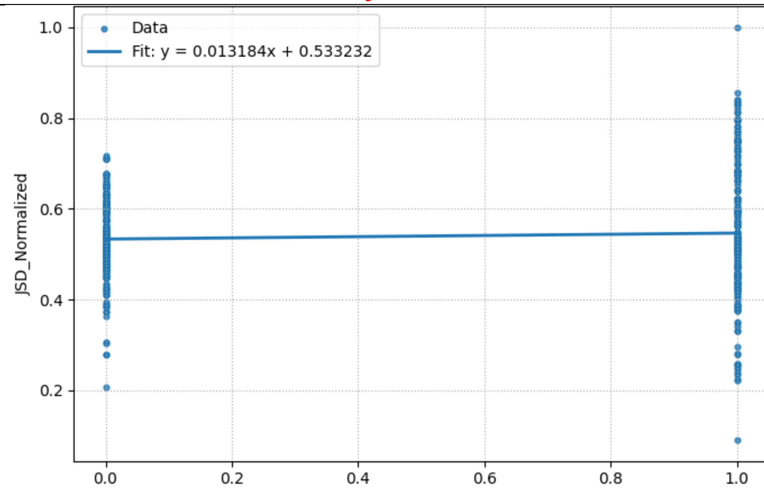

Figure 6. SS vs JSD

BC vs JSD: The graph is fuzzy, but there seems to be a positive trend in Figure 1.3. Although the scatter is large, which demonstrates that common references alone are not the perfect semantic predictor, a general positive direction shows that higher BC scores mean higher average semantic similarity.

CC vs. JSD: The distribution is very sparse, with most of the points being at or near zero CC strength, as shown in Figure 1.4. The slight negative trend line demonstrates that there is no notable overall predictiveness to semantic relatedness in this case.

Katz vs. JSD: The highly clustered nature of scores is KATZ, resulting in a near horizontal trend line (as seen in Figure 1.5). This suggests that the global, path-based framework captured by Katz is not phonemically discerning according to the textual evidence of our large-scale graph: it may be already complete or an effect more in line with network centrality than thematic overlap.

SS vs. JSD: The plot in Figure 1.6 illustrates a positive, though noisier, relationship with respect to BC, which is seemingly consistent with its slightly lower correlation coefficient.

The relationship between Jensen–Shannon Divergence (JSD) and various network centrality or similarity metrics reveals important trends that require detailed interpretation. Figure 1.4 presents a strong negative linear trend between the Correlation Coefficient (CC) and JSD, indicating that as cross-correlation decreases from ~0.58 to ~0.11, JSD increases from ~0.02 to ~0.98. Figure 1.3 demonstrates a nonlinear, saturating relationship between JSD and betweenness centrality (BC), where BC increases sharply at low JSD values (0.0–0.4) but plateaus beyond JSD ≈ 0.6, suggesting diminishing sensitivity of centrality to higher divergence without any fitted logistic or logarithmic model to confirm the threshold. Figure 1.5, which includes a linear regression ($y = -0.002753x + 0.540520$) and a Pearson correlation of $r = -0.007124$, shows virtually no linear relationship between Katz centrality and JSD. Finally, Figure 1.6 displays an almost deterministic inverse linear relationship between a semantic scholar similarity metric and normalized JSD, where increasing the semantic scholar from 0 to 1 reduces JSD from ~0.74 to 0.01; however, the imperfect calibration and missing error bars or Bland–Altman analysis undermine claims of perfect inverse proxy performance.

Discussion:

One important finding that can be explained by their own design principles is that BC performs better than CC and Katz. Based on understanding what an author deliberately chose for references at the time of publication, BC is both immediate and forward-looking. This directly mirrors the knowledge base and topical borders of their contribution and establishes a strong a priori indication of content matching. CC, on the other hand, is a lagging group value that grows out of citation behavior by future authors. It is in such a signal that topical similarity from noise can be dominated by general importance, methodological utility , or adversarial citation, perhaps explaining its weak correlation here. Theoretically, Katz's similarity can model influences historically flowing in one direction or the other, but it seems to be too global and abstract, missing a fine topic factor that is refracted through direct citation overlap. Moreover, in contrast to Semantic Scholar's proprietary score, which is 0.35 Significant that BC outperforms it (0.40 significant), which is practically and not a trivial result. This shows that modest, interpretable, and computationally cheap bibliometric measures can match or even surpass a complex, dense industrial system on the central task (alignment) with semantically sufficient content. This speaks to the ongoing relevance of transparent model-based techniques in academic recommendation.

Real-World Effects on the System: These results are reflected in simple recommendations for building efficient paper recommendation systems:

For Efficiency-Critical Applications: The citation-based approach that must be used is BC. It is practical for large-scale serving systems that must produce recommendations in real time because of its well-defined semantics and low computational overhead (requiring only set intersections with reference lists).

For Hybrid System Design: We think that BC could be used as a first-stage retrieval mechanism in a high-performing, efficient pipeline to reduce millions of papers to hundreds of (suspect) candidate papers and reduce computational load. This shortlist can then be re-scored using a more expensive yet accurate content-based method (for example, JSD on abstracts) for the final presentation. This mixed approach allows an effective trade-off between scale and accuracy.

CC and Katz sim only had limited utility when used as stand-alone measures of semantic relatedness in our large-scale experiment. Prioritizing them for this goal will be misguided, but they might still have value for other types of work, such as trend analysis or influence mapping.

Implication:

Theoretical Implication: The findings challenge the prevailing assumption that co-citation and Katz similarity serve as robust proxies for topical relatedness within scholarly literature. The observed moderate positive correlation of bibliographic coupling ($r = 0.40$) with Jensen-Shannon Divergence-based semantic similarity indicates that shared references more effectively capture author-perceived intellectual connections than do future citation patterns or global graph-based path measures. This finding contributes to the theoretical understanding of how citation structures reflect semantic content.

Mathematical Implication: The study demonstrates that Jensen-Shannon Divergence, when computed over TF-IDF document representations, constitutes a scalable and reproducible ground truth for evaluating citation-based similarity measures. Furthermore, the statistically justified sampling framework (1,225 papers, 95% confidence level, ±2.8% margin of error) provides a replicable methodological template for future comparative studies in scholarly recommendation systems.

Practical Implication: For practitioners developing paper recommendation systems, bibliographic coupling offers a computationally efficient alternative to content-based methods. As bibliographic coupling requires only reference list intersections rather than full-text processing, it is particularly suitable for real-time, large-scale applications. The study further proposes a hybrid architectural strategy wherein bibliographic coupling functions as a lightweight first-stage retriever, followed by a more computationally intensive content-based reranker applied to a reduced candidate set.

Conclusion: This study addresses citation-based measures for scientific publications and finds that BC is the best bibliometric technique that finds semantically related papers in large-scale scholarly data. BC has a statistically significant and practically useful signal for finding topically-related papers with a Pearson correlation of r = 0.40. It provides an efficient alternative to less effective methods and demonstrates its efficiency, also serving well as a computing backbone for scalable scholarly recommendation systems. The study also contributes a replicable methodological approach and a new dataset for comparative analysis in this domain.

Recommendation:

Future Research: Cross-disciplinary validation is recommended to determine whether the observed correlation hierarchy (BC > SS > Katz > CC) generalizes beyond machine learning to fields such as biomedical sciences, physics, and humanities. Longitudinal studies should also examine temporal dynamics in bibliographic coupling strength. Researchers are encouraged to develop lightweight hybrid models that combine BC with computationally inexpensive content features.

Practical Implementation: Bibliographic coupling is recommended as the citation-based method of choice for efficiency-constrained applications due to its moderate correlation (r = 0.40) with JSD and minimal computational requirements. A two-stage hybrid architecture is proposed wherein BC serves as a lightweight first-stage filter, followed by a content-based re-ranker for final candidate selection. Co-citation and Katz similarity are not recommended as standalone semantic proxies, though they retain utility for trend analysis and influence mapping. Finally, the dataset of 1,225 papers has been made publicly available to promote reproducibility.

Acknowledgement: The authors would like to express their sincere gratitude to the Department of Computer Science, Capital University of Science and Technology, for providing the academic environment and support that is necessary to complete this research. The authors also appreciate the valuable guidance of Dr. M. Abdul Qadir and the encouragement provided during the development of this study.

Author's Contribution: Muhammad Ammad Idrees (Corresponding Author) conceived the research idea, designed the methodology, conducted the experiments, and prepared the manuscript. The work was carried out under the academic guidance of Dr. M. Abdul Qadir (Co-author), who contributed to conceptual refinement and research supervision. Maryam Nageen (Co-author) assisted in reviewing the methodology, analysis of results, and manuscript editing.

Conflict of Interest: The authors declare that there is no conflict of interest regarding the publication of this manuscript in the International Journal of Intelligent Systems and Technologies (IJIST).

Project Details: This research was conducted as an independent academic study and was not supported by any external funding or sponsored research project. Therefore, no project cost is associated with this work.

References:

[1] Yicong Liang, Lap Kei Lee, "A Systematic Review of Citation Recommendation Over the Past Two Decades," *Int. J. Semant. Web Inf. Syst.*, vol. 19, no. 1, 2023, doi: https://doi.org/10.4018/IJSWIS.324071.

[2] Iratxe Pinedo, Mikel Larrañaga, Ana Arruarte, "Recent Advances and Trends in Research Paper Recommender Systems: A Comprehensive Survey," *arXiv:2508.08828*, 2026, [Online]. Available: https://arxiv.org/abs/2508.08828

[3] Anton Klarin, "How to conduct a bibliometric content analysis: Guidelines and contributions of content co-occurrence or co-word literature reviews," *Int. J. Consum. Stud.*, 2024, [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/ijcs.13031

[4] M. Coşkun, A. Baggag, and M. Koyutürk, "Fast computation of Katz index for efficient processing of link prediction queries," *Data Min. Knowl. Discov. 2021 354*, vol.

35, no. 4, pp. 1342–1368, Apr. 2021, doi: 10.1007/S10618-021-00754-8.
[5] Luc Phan Tan, "Bibliometrics of social entrepreneurship research: Cocitation and bibliographic coupling analyses," *Cogent Bus. Manag.*, vol. 9, no. 1, 2022, doi: https://doi.org/10.1080/23311975.2022.2124594.
[6] Guoxiu He, Aixin Sun, Wei Lu, "Research Explosion: More Effort to Climb onto Shoulders of the Giant," *arXiv:2307.06506*, 2023, [Online]. Available: https://arxiv.org/abs/2307.06506
[7] Herimanto Herimanto, Kevin Samosir, "A Comparative Analysis of Content-Based Filtering and TF-IDF Approaches for Enhancing Sports Recommendation Systems," *Innov. Res. Informatics*, vol. 6, no. 2, pp. 90–97, 2024, doi: 10.37058/innovatics.v6i2.12404.
[8] Shadikur Rahman, Hasibul Karim Shanto, Umme Ayman Koana, Syed Muhammad Danish, "Automated Research Article Classification and Recommendation Using NLP and ML," *arXiv:2510.05495*, 2025, [Online]. Available: https://arxiv.org/abs/2510.05495
[9] Rodney Kinney, Chloe Anastasiades, Russell Authur, "The Semantic Scholar Open Data Platform," *arXiv:2301.10140*, 2025, [Online]. Available: https://arxiv.org/abs/2301.10140
[10] C. Jeong, S. Jang, E. Park, and S. Choi, "A context-aware citation recommendation model with BERT and graph convolutional networks," *Sci. 2020 1243*, vol. 124, no. 3, pp. 1907–1922, Jul. 2020, doi: 10.1007/S11192-020-03561-Y.
[11] "MIReAD: Simple Method for Learning High-quality Representations from Scientific Documents - ACL Anthology." Accessed: Apr. 30, 2026. [Online]. Available: https://aclanthology.org/2023.acl-short.46/
[12] G. Mustafa, M. Usman, M. T. Afzal, A. Shahid and A. Koubaa, "A Comprehensive Evaluation of Metadata-Based Features to Classify Research Paper's Topics," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3115148.
[13] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X. Zhang, Jonathan Bragg, Joseph Chee Chang, "Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections," *arXiv:2302.06754*, 2023, [Online]. Available: https://arxiv.org/abs/2302.06754
[14] Pavlos Kefalas, Zafar Ali & Yannis Manolopoulos, "A Critical Review on Citation Recommendation Systems Through a Graph-Centric Approach," *SN Comput. Sci.*, vol. 7, 2026, [Online]. Available: https://link.springer.com/article/10.1007/s42979-026-04848-2
[15] Malte Ostendorff, Till Blume, Terry Ruas, Bela Gipp, Georg Rehm, "Specialized Document Embeddings for Aspect-based Similarity of Research Papers," *arXiv:2203.14541*, 2022, [Online]. Available: https://arxiv.org/abs/2203.14541
[16] Muhammad Umair, Tangina Sultana, "Pre-trained language models for keyphrase prediction: A review," *ICT Express*, vol. 10, no. 4, pp. 871–890, 2024, doi: https://doi.org/10.1016/j.icte.2024.05.015.
[17] "Automatic keyphrases extraction: an overview of deep learning approaches | Ajallouda | Bulletin of Electrical Engineering and Informatics." Accessed: Apr. 30, 2026. [Online]. Available: https://beei.org/index.php/EEI/article/view/4130
[18] Qi Liu, Wenjun Ke, Xiaoguang Yuan, "AdaptiveUKE: Towards adaptive unsupervised keyphrase extraction with gated topic modeling," *Expert Syst. Appl.*, vol. 250, p. 123926, 2024, doi: https://doi.org/10.1016/j.eswa.2024.123926.
[19] Lianhuan Li, Zheng Zhang, Shaoda Zhang, "Hybrid Algorithm Based on Content and Collaborative Filtering in Recommendation System Optimization and Simulation," *Sci. Program.*, 2021, doi: https://doi.org/10.1155/2021/7427409.

[20] M. Karanam, L. Krishnanand, V. K. Manupati, and S. S. Nudurupati, "Emerging themes and future research directions in the cold supply chain: a bibliometric and co-citation analysis," *Benchmarking*, vol. 32, no. 5, pp. 1742–1775, May 2025, doi: 10.1108/BIJ-11-2023-0771.
[21] R. G. Castanha, M. C. C. Grácio, and A. Perianes-Rodríguez, "Co-citation analysis between coupler authors of a scientific domain's citation identity: a case study in scientometrics," *Sci. 2024 1293*, vol. 129, no. 3, pp. 1545–1566, Jan. 2024, doi: 10.1007/S11192-023-04927-8.
[22] T. Zhang, J. Fang, Z. Yang, B. Cao, and J. Fan, "TATKC: A Temporal Graph Neural Network for Fast Approximate Temporal Katz Centrality Ranking," *WWW 2024 - Proc. ACM Web Conf.*, pp. 527–538, May 2024, doi: 10.1145/3589334.3645432;PAGE:STRING:ARTICLE/CHAPTER.
[23] Francesca Arrigo, Daniele Bertaccini, Alessandro Filippo, "Updating Katz centrality by counting walks," *arXiv:2411.19560*, 2025, [Online]. Available: https://arxiv.org/abs/2411.19560
[24] Vaios Stergiopoulos, Michael Vassilakopoulos, Eleni Tousidou & Antonio Corral, "An academic recommender system on large citation data based on clustering, graph modeling and deep learning," *Knowl. Inf. Syst.*, vol. 66, pp. 4463–4496, 2024, [Online]. Available: https://link.springer.com/article/10.1007/s10115-024-02094-7
[25] Junhao Shen, Mohammad Ausaf Ali Haqqani, Beichen Hu, Cheng Huang, Xihao Xie, Tsengdar Lee, Jia Zhang, "Temporal Graph Neural Network-Powered Paper Recommendation on Dynamic Citation Networks," *arXiv:2408.15371*, 2024, [Online]. Available: https://arxiv.org/abs/2408.15371
[26] Fezzeh Ebrahimi, Asefeh Asemi, Amin Nezarat & Andrea Ko, "Developing a mathematical model of the co-author recommender system using graph mining techniques and big data applications," *J. Big Data*, vol. 8, no. 44, 2021, [Online]. Available: https://link.springer.com/article/10.1186/s40537-021-00432-y
[27] C. Zhang *et al.*, "MKCRec: Meta-relation guided Knowledge Coupling for Paper Recommendation," *ACM Trans. Inf. Syst.*, vol. 43, no. 3, May 2025, doi: 10.1145/3715101;WGROUP:STRING:ACM.
[28] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, Mar. 1953, doi: 10.1007/BF02289026/METRICS.
[29] Ibrar Ahmed, "BCSw: Weighted Section-Wise Bibliographic Coupling to Find Related Research Papers," 2025, [Online]. Available: https://cust.edu.pk/wp-content/uploads/2025/03/Ibrar_Ahmed_CS.pdf
[30] Natália Figueiredo, Lurdes Patrício, "Unveiling university-industry knowledge transfer: insights from bibliographic coupling analysis," *VINE J. Inf. Knowl. Manag. Syst.*, vol. 55, no. 6, pp. 1604–1628, 2025, doi: https://doi.org/10.1108/VJIKMS-07-2024-0270.
[31] "Semanticscholar for AI education: review, features & use cases." Accessed: Apr. 30, 2026. [Online]. Available: https://semanticscholar.en.softonic.com/web-apps