CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Experimental Evaluation of Related Papers Finding Techniques

by

Muhammad Ammad Idrees

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing
Department of Software Engineering

2025

CERTIFICATE OF APPROVAL

Experimental Evaluation of Related Papers Finding Techniques

by

Muhammad Ammad Idrees

(MAI233005)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Amanullah Yasin	BU, Islamabad
(b)	Internal Examiner	Dr. Farah Haneef	CUST, Islamabad

Dr. M. Abdul Qadir
Thesis Supervisor
November, 2025

Dr. Nadeem Anjum
Head
Dept. of Software Engineering
November, 2025

Dr. M. Abdul Qadir
Dean
Faculty of Computing
November, 2025

Author's Declaration

I, **Muhammad Ammad Idrees** hereby state that my MS thesis titled "**Experimental Evaluation of Related Papers Finding Techniques**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Muhammad Ammad Idrees)

Registration No: MAI233005

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Experimental Evaluation of Related Papers Finding Techniques**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Muhammad Ammad Idrees)

Registration No: MAI233005

Acknowledgement

All praises to Almighty ALLAH who gave me the strength and ability to complete this work. His blessings made every difficult moment easier and gave me hope when things felt impossible.

All praises, respect, and love to the **Holy Prophet Hazrat Muhammad (P.B.U.H)**, whose life is a perfect example of guidance for all of us.

I am truly thankful to my supervisor, **Dr. M. Abdul Qadir (Capital University of Science and Technology, Islamabad)**, for his support and kind guidance throughout this research. His advice and encouragement helped me stay focused and motivated.

My heartfelt thanks to my **parents and siblings** for always believing in me. Their prayers, support, and love gave me the strength to keep going, even when it was hard.

The process has contributed significantly to intellectual development, analytical thinking, and knowledge acquisition. Thanks to everyone who walked even a step of it with me.

(Muhammad Ammad Idrees)

Abstract

Related-paper recommendation systems generally fall into two categories: content-based (CB) approaches, which estimate relatedness using semantic similarity between 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—are known to provide accurate relatedness scores, they are computationally expensive because they require processing the full text of each paper. Metadata-based methods offer a more efficient alternative, but their effectiveness relative to strong CB measures remains unclear. This study investigates which bibliometric technique correlates most strongly with JSD-based semantic relatedness, with the goal of identifying a low-cost substitute for computationally expensive CB methods. 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. 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). 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. Overall, the results highlight the potential of BC-based methods as an efficient and highly accurate alternative to traditional full-text similarity computations to find relatedness.

Contents

List of Figures

List of Tables

Abbreviations

AdaptiveUKE	Adaptive Unsupervised Key-phrase Extraction
AKE	Automatic Key-phrase Extraction
ANN	Artificial Neural Network
BC	Bibliographic Coupling
CC	Co-citation
CNN	Convolutional Neural Network
Doc2vec	Document to Vector
GNN	Graph Neural Network
JSD	Jenson Shannon Divergence
KT	Katz
NLP	Natural Language Processing
Node2vec	Node to vector
PLMs	Pre-Trained Language Model
SCI	Science Citation Index
Scival	Science Value Analytics

Chapter 1

Introduction

1.1 Overview

Paper recommendation systems are platforms that share the same input papers and return semantically or contextually similar ones. Reviews of available methods for searching relevant research articles, considering their range and characteristics, such as content-based methods and metadata-based methods [1].

The search for relevant publications is integral to the evolution of knowledge and science. It enables researchers to advance the state of the art and build on existing work. This problem has garnered 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 literature. Researchers are continually striving to improve the precision and recall of approaches to discover research publications using metadata and digital library contents [2]. In this thesis, we address the limitations of existing approaches and analyze which bibliometric approach performs better to find semantically related papers. There are two common methods for finding relevant research articles: a content-based approach and a metadata-based approach. Bibliographic information, especially citation, is indeed a valuable indicator of related research articles in metadata, but text is the building block of content-based relatedness. To address this challenge, three

widely used strategies are co-citation, bibliographic coupling, and Katz similarity. Bibliographic coupling is used to identify papers referenced by multiple papers, co-citation links are determined by papers mentioned by the citing article, and Katz centrality takes into account the number and length of paths between nodes when evaluating the significance between nodes. Although bibliographic coupling is based on papers that are referenced by many articles, citation analysis is based on papers that cite the same citing work, and Katz reveals direct and indirect connections [3].

Thus, a citation index such as the Science Citation Index (SCI) will tell us how many times two cited papers have been cited together. The Citation Index of the SCI is an online resource providing information about the citing papers for the two papers in these two lists. The degree of co-citation regarding the two cited papers is measured by the number of common citing items.

The relative frequency with which a pair of old literature is cited collectively by later literature is called co-citation. An identical citation item is a new paper that cites the same works as the original one [4]. We limited the bibliographic coupling-returned publication records to articles with a bibliographic coupling score of two or greater to increase efficiency.

Based on bibliographic coupling, co-citation, and Katz's similarity measure of the revisable relevance between papers, we find the correlation between JSD vs bibliographic coupling, co-citation, and Katz's similarity. The research tries to differentiate similar papers in a set of citations by considering the cited papers.

The focus of the study is to find a correlation between JSD vs bibliographic coupling, co-citation, and Katz similarity. Also, check the Semantic Scholar-related papers with JSD, which technique that performs well to find related papers.

Jensen-Shannon Divergence (JSD) as a content-based ground truth to provide a more objective evaluation of citation-graph techniques. Traditional methods for recommending related papers have generally made use of individual techniques such as bibliographic coupling, co-citation, or graph traversal-based methods [5].

Nevertheless, the performance of which bibliometric method is best to find semantically related papers, such as co-citation is backward-looking and relies on future citations, whereas bibliographic coupling is forward-looking but sensitive to old references. Recent studies recommend the technique for better performance for relevance detection in various research fields.

Specifically, the Katz similarity is included to be able to discover indirect but semantically grounded relations following multi-hop paths in the citation graph, thus expanding the contextual scope recommendations beyond direct citation links [6]. The use of the three citation-based indicators balances local citation behavior with global structural awareness and provides a more robust and scalable framework for relatedness detection in the scientific literature.

With the exponential growth of academic papers, it is becoming difficult to discover related research papers. The 1960s witnessed the emergence of preliminary approaches, such as co-citation analysis and bibliographic coupling. Introduced the concept of co-citation and demonstrated that it could be used to expose relationships between academic papers.

Established Bibliographic Coupling, a schema to connect documents through references shared between them. These basic procedures were also essential for the development of citation analysis, the most widely used means to identify relationships between scholarly documents for many years [7].

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. Over the course of the late 1990s and early 2000s, it became clear to researchers that citation counts are not a reliable source of measuring the importance of a paper, as not all citations are influential [8]. Researchers began to develop elaborate techniques to flag up relevant papers.

This study investigates the results on a single dataset and which bibliometric method: bibliographic coupling, co-citation, and Katz similarity performs the best

with JSD. Which method correlates most strongly with JSD-based semantic relatedness to identify a low-cost substitute for computationally expensive Content-based methods.

1.2 Techniques for Identifying Related Papers

1.2.1 Content-based Technique

Content-based recommendation systems have become an integral part of contemporary scholarly discovery platforms. These models usually conduct textual analysis of research papers including title, abstract, keywords and full texts to elicit semantic representations that indicate the latent topics or themes of a paper. Gradient-based methods calculate the similarity (affinity) between these representations to find papers that share common themes, use comparable techniques, or investigate similar research issues.

These systems can often identify complex relationships that go beyond a citation network or metadata-based bibliographic information: for example using means such as TF–IDF, word embeddings, topic modeling and semantic distances (e.g., cosine similarity, Jensen–Shannon Divergence). Content-based approaches, in contrast to citation-based techniques that require accessible citation linkage from articles, can recommend recently published or uncited work and are especially valuable in emerging scientific domains.

With the exponential increase of academic literature, content-based recommendation systems are valuable in assisting researchers to find relevant studies efficiently, track the evolution of research topics and prevent information overload. We argue that methods for generating suggestions that can be used independently of citation network representations are crucial ingredients in hybrid recommendation approaches designed to provide more accurate and more comprehensive suggestions of related papers.

1.2.2 Topic-Model-Based Technique

Topic modeling has emerged as a powerful probabilistic approach for uncovering the underlying thematic structure within large and unstructured text corpora, providing a robust foundation for enhancing research paper recommendation systems. Unlike literal keyword-matching techniques, methods such as Latent Dirichlet Allocation (LDA) [9] automatically infer latent "topics" which are distributions over words from document collections, enabling a deeper, semantic understanding of content. By representing both papers and user interests as distributions over these discovered topics, these methods can compute similarity based on shared thematic proportions, effectively capturing conceptual relationships even when specific terminology differs. This capability makes topic models particularly valuable for facilitating interdisciplinary discovery and mitigating the vocabulary mismatch problem, thereby allowing for recommendations that are thematically coherent and aligned with a user's fundamental research interests rather than mere surface-level text overlap.

1.2.3 Feature-based Technique

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 [10]. These methods involve a two-stage process: first, a feature engineering phase where relevant characteristics are extracted from paper metadata and content, such as keywords, author affiliations, citation counts, journal impact factors, and term frequency vectors. Subsequently, these features are used to construct a feature vector for each document, and machine learning or similarity metrics are applied to compare these vectors and identify the most relevant recommendations. The primary strength of this approach lies in its transparency and controllability, as the contribution of specific attributes—such as favoring recent publications or papers from high-impact venues—can be explicitly weighted and tuned. However, a key challenge

is the effective selection and combination of these features to accurately capture semantic relevance beyond superficial correlations.

1.2.4 Citation-based Technique

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.

As they are rather an authors' admissive account of previous related work, citations-based indicators can be assumed to provide us a reliable estimate of both scholarly impact and thematic link. They are especially useful in well-established research fields that contain dense citation links which allow to recommend influential and similar papers.

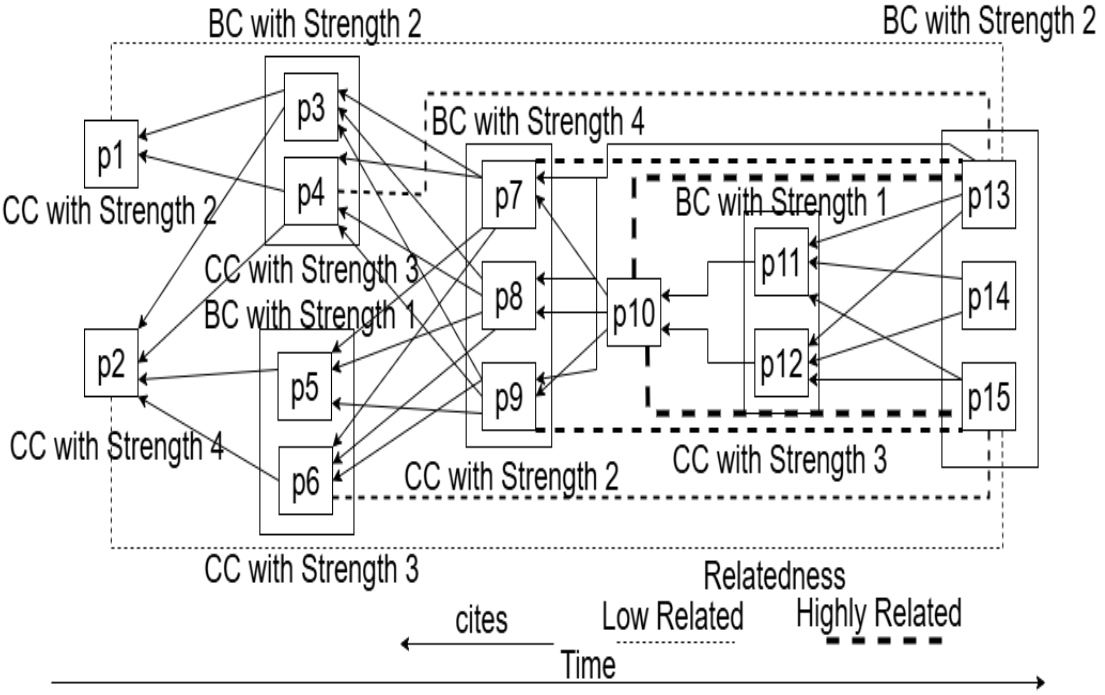

FIGURE 1.1: Citation Graph Representing Co-citation, Bibliographic Coupling, and Katz Similarity Relations.

1.2.4.1 Co-citation Analysis

Co-citation (CC) is a fundamental bibliometric method that computes the similarity of the top 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.

Continuously recited by the Canonical documents, repeated co-citations between more than one paper are considered as an indicator of scholarly proximity. Such frequencies are recorded in a co-citation matrix that can be used to cluster papers or visualize the intellectual structure of one or several domains.

Two papers are co-cited (CC) if one paper cites both of them. In the figure 1.1 p3 and p5 are co-cited by p6, p7, and p9. Therefore, they are co-cited (CC). The greater the number of times two papers are co-cited, the greater their co-citation strength. The bold edges, which consist of the co-citation triangle (for instance, p6 $\rightarrow$ p3, p6 $\rightarrow$ p5), belong to the graph base as well, but have a similarity value due to the co-cited-in-the-same-article property [4].

By the time different models were proposed where the co-citations network was dynamically modified and as a result the representation of scientific growth is more flexible. Riding on the shoulders of this work, recent progress in statistical modeling has empowered us to monitor the evolution of citation behavior through time in more detail. These dynamic models not only trace the development of academic research but also uncover growth and change patterns of research hotspots as well as the intellectual structure of scientific domains.

There are still some limitations remaining in co-citation analysis, however despite its development. In particular one of the major challenges is the computationally expensive nature of large-scale bibliographic networks. The analysis of large citation datasets is computationally expensive and may limit the application on the fly or at scale. This problem of scalability is a fundamental obstacle to the more general use of co-citation techniques in the current contexts of research [11].

In addition, co-citation analysis is fundamentally based on the existing citation data, and this method has the disadvantage of the 'cold start' problem. New or less-cited papers might be underrepresented or omitted due to a lack of citation counts.

This dependence precludes the system from attaching importance to new or innovative research that is outside the influential power of academia and may undervalue important contributions in new fields [12].

1.2.4.2 Bibliographic Coupling

Bibliographic Coupling (BC) considers two documents similar if they are referenced by the same set of papers. Unlike co-citation, which directs its attention forward toward who cites the paper, BC by contrast is a rearward-looking metric it evaluates the evidence trail left in reference lists. For example, in Figure 1.1 Paper p13 cites p11, p12, and Paper p14 cites p11, p12, they are bibliographically coupled with strength 2 (i.e., with p11, and p12 in common). This enables BC to work as soon as the paper is published, making it particularly beneficial while newer papers do not yet have the opportunity to be cited.

We say that two papers are BC-linked when they have common references. In the Figure 1.1 p1 and p2 are bibliographically coupled (BC) since they both cite p3, p5, and p6. These common outgoing links (p 1 and p 2 go to the same nodes) establish the BC relation. The citation links directly connecting a paper to its references are not used to find similarity, but exploited to compute the BC relationship [13].

Bibliographic coupling states that papers sharing the same set of references are likely to be topically related. This approach was an innovative approach to citation analysis, as a measure of the overlap in reference lists between papers, to provide a perspective of relationships among academic works. The original bibliographic coupling was a static approach that only considered the strength of the link between documents as a function of the number of references to common documents, regardless of their context or sequential appearance.

The strong and weak ties of bibliographic coupling can led us to the introduction of a more refine set of analytical tools. A change in emphasis toward documents' references and in particular on the recency of their co-citations, which became an order to detect the time-to-time relevance and strength of relations between documents. As Bibliographic coupling is a measure for collaboration in scientific research and shown its value for identifying topics of research that are new in centrosymmetric studies. Based on this family of techniques, more advanced methods have been proposed for increasing the precision of the bibliographic coupling approach by resorting to a set of weighting strategies. For example, weighted BC adds different weights to references independent of their contribution, as not all citations have the same effect on the thematic similarity. Some section-based models have also been developed, which puts more weight on the references in the core sections (e.g., Methodology or Results), which are likely to be a more substantive context of study reference. Research in the field has also developed to consider time through dynamic bibliographic coupling [14], which incorporates time-sensitive factors to better reflect the development of research topics. More recently, a hybrid deep learning recommendation system was proposed that provides context-aware mapping of scientific literature and addresses several inherent limitations of traditional coupling methods especially their reliance on static historical citation data.

1.2.4.3 Co-citation vs Bibliographic Coupling

Co-citation and bibliographic coupling are two significant approaches to discovering relations between academic works. Co-citation reveals two papers that are cited together more than expected by later papers, indicating an established or historical relationship in the academic conversation. Bibliographic coupling highlights the forward-looking perspective by associating papers having common references and hence a direct node overlap of themes or ideas. The major advantage of bibliographic coupling is, however, its dynamic nature, changing with each new document, each added reference to the already resident references that precede it. In contrast, co-citation is inherently backward-looking, in that it records rather

static mechanisms of literary connections, based on the citation behavior of older literature. This difference in time dimension highlights the temporal superiority of the bibliographic coupling for characterizing emerging patterns and the status of recent cooperation [15].

1.2.5 Graph-based Technique

1.2.5.1 Semantic Scholar

Semantic Scholar was created to combat the ever-increasing number of scientific publications that exist by constructing a massive machine-readable representation of research, or the literature graph. These nodes and edges, each of which has specialized properties [16], collectively support sophisticated algorithms that can automatically find relationships between research works. To present related papers, Semantic Scholar employs a variety of methods:

Citation Network: A directed link is generated from one paper to another if a citation was made, allowing relatedness based on direct citations, co-citation, and bibliographic coupling.

Entity Extraction and Linking: Neural models to extract scientific concepts from titles/abstracts, which are linked to canonical knowledge bases (UMLS, DBpedia). So papers that have common entities are connected as related.

Citation Quality and Recommendation Models: Machine learning models are used to differentiate between essential and nonessential citations, and to decide which papers should be cited in a draft. This enhances the ranking and relevance of recommended related papers.

1.2.5.2 Katz Similarity

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. Its ability to identify the latent or indirect relation in the large citation graph, social network, and recommendation system makes it especially valuable [17].

Modified versions of Katz similarity have been tailored to new computational contexts in decade-old research on recommending research papers. With the exponential growth of academic databases, the discovery of more profound semantic and structural relationships among documents has ushered in a new era where dynamic, personalized, and deep learning-based extensions of Katz similarity have emerged.

This work demonstrates progression from the theory first presented by Katz to practical systems that can deal with large, real-time citation networks.

Katz Similarity is a graph-based similarity significance measure used in citation networks to measure how similar two nodes (research papers) are, based on the total number of paths available, which are weighted by the path length, between them. It takes into account direct relationships (e.g., traditional citation) and indirect paths of different lengths and thus reveals deeper, hidden connections in a given network.

Katz's similarity is especially suitable for a dense citation graph with many indirect connections. It's also used in recommended systems, in particular when it's integrated into neural architectures or put as part of hybrid models. But it is also computationally costly as it needs to compute all the paths between every pair of nodes of citation networks at a large-scale level. The redundancy can easily be reduced by truncation techniques, such as matrix approximation methods. Katz remains a powerful method for modeling scholarly proximity over and above mere co-occurrence or shared references.

In Figure 1.1, the methods used to recommend the Research Paper are via citation graph-based methods, mainly concentrating on Co-Citation(CC), Bibliographic Coupling(BC), and Katz Similarity-hope distance(D). In Figure 1.1, each box is a

paper (like p13, p11, and so on), and arrows denote direct citing links. Meanings for the different line styles (solid dotted line, thin dotted line) are described in the legend.

The citation graph features various examples of bibliographic coupling and co-citation, highlighting how papers are related in terms of semantics and structure. Papers P5 and P6 exhibit bibliographic coupling strength with a value of 1 since they have shared paper P2 as a common reference. They are also cited to co-cite with a strength of 3(positions p7, p8, and p9 cite them jointly), which is a structurally meaningful level of similarity. The P7, P8, and P9 complementarity peak forms a remarkable bibliographic cluster.

All articles have the identical references, i.e., four: (P3, P4, P5, and P6), and therefore, the bibliographic coupling strength is 4, indicating highly topical overlap. But their co-citation strength is also 1 because they are cited by the same paper P10 together indicating a relatively less outside recognition of this group.

Likewise, P11 and P12 contain a bibliographic coupling of strength 1, since they both refer to P10. P13, P14, and P15 co-cite these two articles with strength 3, which implies the relatively strong association of this topic in the downstream literature. The citing group P13, P14, and P15 has a bibliographic coupling weight of 2, since all three papers cite both P11 and P12. This represents a shared bibliographic history. Moreover, at previous positions in the graph, P3 and P4 are linked together with BC strength 2, since both refer to P1 and P2. Both these papers are co-cited with strength 3; that is, they are cited in common by P7, P8, and P9. Papers P1 and P2 are also co-cited by themselves, so that P1 is co-cited by P3 and P4 (strength 2), and P2 by P3, P4, P5, and P6, having a larger co-citation strength of 4. We also analyze the citation network by Katz similarity, considering different paths of length in the papers. Similarity is only computed using paths of two hops or higher; 1-hop direct citations (represented with dashed lines) are used to build the graph, but are not used for computing the similarity score.

Example 1: Path traversal from P13 to P1 (5-hop analysis)

P13 → P11: Direct citation (1-hop), excluded from Katz score.

P11 → P10: Now P13 is 2-hop.

P10 → P7: P13 is 3-hop.

P7 → P4: P13 is 4-hop.

P4 → P1: P13 is 5-hop.

Thus, P13 is related to P1 with a 5-hop path, reflecting a relatively weak but existent semantic link under Katz similarity.

Intermediate relationships:

P11 to P1: 4-hop

P10 to P1: 3-hop

P7 to P1: 2-hop

Example 2: Path traversal from P15 to P2 (5-hop analysis)

P15 → P12: 1-hop (direct), excluded from similarity.

P15 → P10: P15 is 2-hop from P10.

P15 → P9: P15 is 3-hop from P9.

P15 → P6: P15 is 4-hop from P6.

P15 → P2: P15 is 5-hop from P2.

For P15, we have a 5-hop between P15 and P2 through citation, and a decreasing weight assigned to each hop in the normalization of the Katz measure.

(i) Citation Graph Foundation

Arrows (→) denote directed citation relations, where one paper cites another. Dotted arrows indicate the hop distance from two papers: 1-hop: Direct citation (e.g., p1 →p3) used to construct the graph but not directly used in similarity

calculations (for CC and BC). 2-hop: A paper cites another paper that is cited by another one, which is applied with or in similarity (strong relation). 3-hop and 4-hop show that the relationships reduce in strength with the size of the path.

(ii) Hop-based relatedness

2-hop = Strongly connected.

3-hop = Somewhat related.

4-hop and above = loosely consanguineous.

Dictionary (1) → See all software Katz Similarity penalizes longer paths by a decay factor, and has a structure which is useful in inducing this hierarchy.

1.2.6 Machine Learning and Deep Learning Based

The citation recommendation model that leverages deep language understanding and graph-based learning to enhance related research paper discovery. They utilised 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.

For fair comparison, the authors also created the FullTextPeerRead dataset, which is an extension of the PeerRead corpus, but this time they have citation contexts, metadata, and citation links included to improve performance in this work. Experiments on the ACL Anthology Network and FullTextPeerRead datasets indicated that their model also produced substantial performance improvements over previous benchmarks (e.g., CACR), with 28% improvement across mean average precision and recall scores. The results demonstrate that contextual information from natural language and structural information in the citation graph are complementary, and their combination can help to receive more accurate related paper recommendations. Our study proves that the combination of deep learning on textual semantics and graph neural network-based structural knowledge is

effective, which paves the way for further AI-powered scholarly recommendation systems[18].

MIReAD (Minimal Information for Representation Learning of Academic Documents), specifically addressing the creation of document-level representations of academic documents with transformers. 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. The model is built based on the SciBERT architecture and fine-tuned using a simple but effective classification objective of journals, placing each paper into 2,734 journal categories obtained from over 500k articles in PubMed and arXiv. This supervision information exploits the observation that journals tend to be very targeted, and thus are strong proxies for topical domains. Evaluation on various benchmarks, such as linear classification tasks (MAG, MeSH), unseen journal names for PUBMED, and clustering purity and information retrieval, showed that MIReAD performs consistently better than six state-of-the-art models (BERT, PubMedBERT, BioBERT, SciBERT, Cite-BERT, and SentenceSER) or comparably with SPECTER. It is also interesting to see that MIReAD achieved significant boosts for retrieval in fields such as Computer Science, Economics, and Electrical Engineering, suggesting its capacity to model domain-specific semantics. Through mitigating deficiencies of the objective based on top k citation, MIReAD presents an efficient and scalable solution to improve literature search and relevant paper recommendation via strong deep learning–driven representations[19].

1.2.6.1 Metadata Based

The growing number of scientific publications has led to a necessity of 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. Not only does this provide a solution to the access limitations placed by many publishers, it also guarantees scalability when dealing with big digital libraries. As such, each individual metadata feature is powerfully discriminatory though they can together improve and add to the performance. In order to verify this, machine learning classifiers such as Random Forest, K-Nearest Neighbor and Decision Tree were analysis on a large ACM dataset and systematic study of feature combinations showed that the combination of title and keywords [20].

MotifClass is a heuristic based text classification approach, which uses higher order metadata to increase categorization accuracy without the need for manually labeled training data. Unlike traditional methods, which only take the document text into account, MotifClass considers the connections of documents to metadata (author name, venue and year of publication) through a heterogeneous information network and extracts motif instances (metadata combinations) indicative for the categories. The pipeline consists of three stages: (1) we pick a few metadata motifs that are highly matched with the surface names for category labels, (2) retrieve candidate training examples using those motifs and generate labeled data artificially, and then train a text classifier on the constructed dataset.

Experiments on large-scale datasets MAG-CS and Amazon reviews demonstrate that MotifClass consistently outperforms the baseline weakly supervised methods as well as metadata-aware retrieval-based models, to confirm the advantages of exploiting higher-order metadata information for effective and scalable document classification [21].

1.3 Problem Statement

This study investigates a correlation between metadata-based techniques with a stronger related papers' finding techniques, that is the content-based related paper finding technique by using JSD. As a result, there is no content-aware basis for choosing among citation-structure methods, specifically Bibliographic Coupling (BC), Co-Citation (CC), and Katz similarity, when the goal is to retrieve truly

related papers. At the same time, Semantic Scholar provides its own "related papers" information, but the degree to which that external label agrees with JSD-based semantic similarity is unknown. This study, therefore, frames the problem as a correlation-driven evaluation and selection task:

(i) Measure how strongly each citation technique (BC, CC, Katz) aligns with JSD-derived semantic similarity.

(ii) Measure how well Semantic Scholar's related-paper lists align with JSD.

(iii) Use these diagnostics without assuming a winner a priori to identify the most reliable citation-based technique for large-scale related-paper ranking and banding in a research paper corpus.

1.4 Research Objective

RO1 Build a cleaned, normalized corpus with structured references, citations, and related papers.

RO2 Compute BC, CC, and Katz scores; construct a JSD-based semantic proxy from TF-IDF text representations; normalize scores for comparability.

RO3 Use a representative sample to quantify correlations between each technique and JSD; select a production scorer based on evidence.

1.5 Research Question

RQ1 How can we build a reliable dataset of machine-learning research papers?

RQ2 Among BC, CC, Katz, and Semantic Scholar, which technique aligns most strongly with JSD?

RQ3 How can we group related papers into three categories: Highly related, Related, and Weakly related?

1.6 Methodology

The research methodology comprises four phases, adapted from the eight-step model proposed by a study [22]

FIGURE 1.2: The Design Science Research Methodology Process for this work

1.6.1 Phase 1: Deciding

Step 1: Develop research problem with systematic literature review associated gaps in extant interplay bibcoupling, Co-citation, and Katz similarity.

Step 2: Research problem we should write a research problem.

1.6.2 Phase 2: Research Study Planning

Step 1: Develop a research design to address RQ1, RQ2, and RQ3.

Step 2: Data Collection for RQ1, RQ2, and RQ3.

Step 3: Update research problem.

1.6.3 Phase 3: The Research Process

Step 1: To address RQ1, RQ2, and RQ2 experimentally.

Step 2: Analysis and comparisons of RQ1, RQ2 and RQ2 with state-of-the-art techniques.

Step 3: Update research problem.

1.6.4 Phase 4: Writing the Research Report

Step 1: Document the research methodology and results to comprehensively analyze the study's outcomes.

Chapter 2

Literature Review

2.1 Overview

We reviewed the literature on related-article searching techniques, which included text-based and citation-based approaches (bibliographic-based approaches, as well as Katz similarity). Pros and cons of several methods to identify matching publications were discussed. The methods and efficiency of each method to search the literature were comprehensively reviewed. The most appropriate tactics were chosen based on the practicability and impact. This required more analysis of existing research while overcoming the research gaps. For a better understanding of these techniques, the literature was reviewed through references to the papers discussed. It was diving into old books and finding out how the approaches had evolved [23].

2.2 Approaches to Search Related Research Articles

The recommendation method for scientific papers was proposed years ago, which investigates many research studies. These methods could be broken up into several

strategies, where a different point of view dazzles at the recommendation problem. This diversity adds to the body of work by allowing for context-aware advice in academic literature[24].

(a) Content-Based Approaches.

(b) Citation-based Approaches.

(c) Hybrid Approaches.

2.2.1 Content-based Approaches

This study [25] presents a content-based recommender system which exploits a keyphrase extraction module providing both user interest and paper content. The system extracts papers of interest for uni-grams, bi-grams and tri-grams which are used to create a weighted user profile, computing similarity scores with new papers using the cosine similarity.

Experiments are performed on the ACL Anthology Reference Corpus (ACL ARC): a corpus of 597 full papers selected from a subset of 10,921 papers published since February 2007 and relevance feedback provided by 28 researchers (15 junior, 13 senior).

The evaluation involves splitting each relevant papers into training and test sets in order to evaluate recommendation accuracy against the Information Foraging Theory (IFT) benchmark system, though exact numeric performance metrics which are not provided in the paper.

A key aspect of this work is the availability of a strong unsupervised approach for keyphrase extraction, which generates rich semantic profiles without relying on extensive training data, which makes it domain-independent and easily adapt to different digital libraries.While effective, this approach relies solely on content features and ignores bibliometric information such as BC, CC, and Katz Similarity. The absence of citation-based analysis prevents a fair comparison

with established bibliometric techniques, leaving unanswered which method better captures semantic relatedness. Moreover, the study does not evaluate its recommendations against unified or standardized datasets. This paper [26] surveys the methods for Keyphrase Prediction (KP) with Pre-trained Language Models (PLMs) into two categories: Keyphrase Extraction (KPE) and Keyphrase Generation (KPG). For KPE, it evaluates techniques such as Attention Mechanisms (UCPhrase,AttentionRank), Graphbased Ranking (theme-weighted PageRank), Semantic Importance (MDERank, INSPECT), and PhraseDocument Similarity (EmbedRank,SIFRank).

They are Low-resource methods (e.g., data augmentation with KPDROP and DESEL algorithm) and Domain-specific model (e.g. SciBERT, BioBERT, BERTweet) For KPG. The review employs diverse collections of datasets e.g., KP20k (570,802 computer science articles), Inspec (20,000 documents), SemEval2010 (244 documents) and Krapivin (2,300 scientific papers) and mentions statistical information such as KP20k having an average of 3.24 present and 2.84 absent keyphrases in each document.

The results of model comparisons shows varying performance, where on the INSPEC the (size:117M-1.3B parameters), JointGL-large achieved F1-score 90.7% and SIFRank having accuracy 84.7%. One of the strengths of this survey is its integrated and systematic analysis covering both retrieval and generation tasks with well-defined taxonomies, as well as promising future research directions.

The survey offers no new empirical evaluation, making it unclear how these semantic methods compare to citation-based techniques like BC, CC, or Katz. It also lacks a unified dataset for fair comparison, and its findings may inherit inconsistencies from varied datasets and evaluation settings across the reviewed studies.

This paper [27] presents an extensive survey of deep learning based AKE methods, which covers models such as multilayer perceptrons (MLP), convolutional neural networks (CNN), recurrent neural networks (RNN), long short-term memory networks (LSTM), bidirectional LSTM (BiLSTM) and bidirectional GRU (BiGRU) in addition to their combinations with other methodologies like CRF and sentence

embeddings. The study investigates different datasets for training and evaluation including Inspec (2,000 documents), Krapivin (2,304), NUS(211), SemEval (288), KP20K(527K papers), KDD for author disclosure (755) and WWW conference(KP) (1,300), highlighting the fact that DL-based approaches mostly rely on large-scale dataset such as KP20K to obtain competitive results.

It is further demonstrated that the combination of present and absent keyphrase prediction significantly outperforms methods which only extracts existing keyphrases with F1-scores up to 0.46 (F on top 5) and 0.41 (F on top 10) on NUS, and deep BiLSTM consistently performing better than various DL techniques across datasets.

One of the main merits of this review is its systematic comparison among DL methods and comprehensive analysis on datasets, architectures, and activation functions, which provides clear guidance to researchers.

The survey does not compare deep-learning keyphrase methods with bibliomatric techniques, leaving their relative effectiveness for related-paper identification unclear. Its focus on heterogeneous datasets also prevents a unified evaluation of semantic relatedness. Moreover, DL-based AKE methods rely heavily on large labeled datasets, limiting their applicability in settings where consistent semantic–bibliometric comparison is needed.

A keyphrase extraction approach [28] based on a Multilayer Perceptron (MLP) neural network is proposed, which ranks candidate noun phrases in terms of class probabilities from the following five features: TF-IDF, position of first occurrence, phrase length, word length and links to other phrases. Experiments are carried out on a corpus of 150 full-text journal articles, in three domains: Economics (60 papers), Law (40 papers) and Medical (50 papers) with an average number of 4.90 author-assigned keyphrases per paper and with 144,978 noun-phrases extracted.

The proposed MLP-based model outperforms Kea on average precision values of 0.34, 0.22, and 0.17 for top 5, 10, and 40 best keyphrases reaches the average recall with corresponding recall values of 0.35, 0.46, and 0.51 respectively. A key

feature of this study is to combine various linguistic and structural features in neural network for enabling more accurate ranking and selection of keyphrases. The method relies solely on content features and does not incorporate citation-based information such as BC, CC, or Katz, limiting its ability to capture deeper semantic relatedness between papers. Its evaluation is also restricted to a small, domain-specific dataset, preventing broader, unified comparison across methods. Moreover, the supervised setup depends on labeled keyphrases, which are not consistently available in many research domains.

This paper propose a content-based approach [29] to global citation recommendation using a two-phase neural framework that comprises a document embedding model (NNSelect) which maps documents into vectors and retrieves candidate citations via nearest neighbor search, followed by a discriminative reranking model (NNRank) that scores the candidates based on features like textual cosine similarity and word intersection counts. The models were trained on PubMed (contains $>$ 45K articles) and DBLP (contains $>$ 50K articles), and an even larger scale new dataset called OpenCorpus, which contains 6.9 million papers with an average of 5 and 17 citations per article for the later data set. The results demonstrate that content-only model (without metadata) significantly outperforms the prior state-of-the-art (ClusCite) with relative improvements of more than 18% in F1 on top 20 (0.302 vs. 0.237 on DBLP) and around 22% in MRR (0.672 vs. 0.548 on DBLP). One of the strength of this work is its ability to place documents in real world settings where author or venue metadata is missing (e.g., due to peer review) and scale to very large corporas without requiring retraining for new documents. This content-based approach ignores citation-based relationships such as BC, CC, or Katz, limiting its ability to capture bibliometric connections between papers. Its reliance on document text and abstracts may reduce effectiveness when semantic cues are subtle. Additionally, the method does not use a unified dataset for cross-method comparison, restricting assessment of true semantic relatedness.

This paper [5] presents a simple approach that highly cited papers are more relevant for retrieving related research papers and recommends related work based on counting in-text citations. Experiments were performed on a dataset of 1,200

J.UCS(Journal of Universal Computer Science) documents and about 16,000 references. The experiments demonstrated that our method had a significantly higher precision of 0.96 for its top-5 recommendations when compared with human-created gold standard, which was much better than other ranking techniques (content-based: 0.76; bibliographic coupling: 0.56 and metadata-based: 0.48). One of the strengths of this work is its conceptual simplicity and strong empirical support for a very intuitive concept is the argument that a straightforward count of in-text mentions is a strong relevance. One notable limitation is the technical difficulty of extracting in-text citation counts from PDFs and fallibility of the process and achieved an initial 78% accuracy, requiring manual correction, and the study is limited to a single journal dataset, restricting generalization across domains. Extracting in-text citations from PDFs is technically challenging and error-prone. It also does not evaluate how counting citations compares with other bibliometric methods like BC, CC, or Katz on a unified dataset.

This paper [30] creates a sports recommendation system by content-based filtering in which two techniques are applied Term Frequency-Inverse Document Frequency (TF-IDF) for vectorization to cope with new users (cold-start problem), and cosine similarity to search similar users for the existing ones (non-cold-start problem). The model was trained and tested on a custom dataset, which consisted of profile and journal data (eg., sport types, summaries) obtained from the Google Scholar for individual users.

The system was evaluated in data splitting experiments, where the best results were 0.256 Precision, 0.307 Recall and 0.869 Accuracy for a 80-20 split of the data. One of the major strengths of this work is its practical design that addresses the cold start problem explicitly for a completely new domain (sport recommendation) by bootstrapping using TF-IDF on journal abstract to generate an initial recommendations to be given to new users. The system relies solely on content-based TF-IDF features and ignores citation-based or bibliometric information like BC, CC, or Katz. Its evaluation is limited to a custom dataset, preventing standardized comparison. Low precision and recall indicate challenges in accurately capturing deeper semantic or bibliometric relationships.

In this article [31] section-wise in-text citation counts (weighted by importance), number of other citing papers, and similarity scores are used as features for a machine learning model for binary citation classification. The authors used two annotated benchmark datasets : Dataset D1 with 457 paper-citation pairs (69 important, 388 non-important) and Dataset D2 with 282 pairs (89 important, 193 non-important).

Their method based on a Neural Networks for section-weighting combined with an Random Forest classification model and achieved a precision score of 0.85 on Dataset D1, which was a substantial improvement over the previous state-of-the art precision of 0.72. A primary strength of this work is to incorporate multiple feature types, specifically including the automatic assigning of logical weights on various paper sections, contributing subtle interpretable insights toward different tasks. This approach uses small, domain-specific datasets, limiting generalization across research areas. It focuses on in-text citation features without comparing to broader bibliometric techniques like BC, CC, or Katz. The reliance on section-weighted features may not capture overall semantic relatedness between papers.

In the Papers [32] a hybrid task-focused recommendation approach was developed that utilized content-based (TF-IDF vectors) features along with co-authorship network analysis for scholarly article recommendations. The authors used a dataset of 7,522 data mining conference and journal articles consisting of 1,000 low-fidelity synthetically generated articles to evaluate recommendation quality. Results show that with hits on top 10 items the hybrid fusion method has the largest performance and is consistently better than either pure content-based or co-authorship-based methods in different task profile scenarios. One of the main merits is that this paper conducts a solid comparison of three hybrid combination methods (switching, proportional and fusion) and performs an extensive evaluation both on recommendation accuracy and article fidelity. The hybrid approach relies on content and co-authorship features but does not incorporate citation-based information like BC, CC, or Katz. Its evaluation includes synthetic articles, which may limit reliability and generalization. Additionally, the method does not assess true semantic relatedness across a unified dataset.

This contribution [33] introduces a new class of information-theoretic divergence measures that stem from the Shannon entropy, including the K directed divergence, its symmetric form the L divergence and the weighted Jensen-Shannon (JS) divergence. It depends on mathematical proofs and probability distributions to justify the new measures.

The key results shows that the new measures are bounded; in parti- cular, the L divergence by the variational distance with respect to their exact inequality

$$L(p_1, p_2) \leq V(p_1, p_2) \tag{2.1}$$

and the JS divergence provides bouns for the Bayes probability of error Pe, given by

$$\frac{1}{4}[H(\pi_1, \pi_2) - JS_\pi(p_1, p_2)]^2 \leq P_e \leq \frac{1}{2}[H(\pi_1, \pi_2) - JS_\pi(p_1, p_2)] \tag{2.2}$$

One the merits of this paper is that it contains rigorous proofs for properties and bounds associated with the new measures, which are important for usage in decision-making and pattern recognition. The paper lacks empirical evaluation on real datasets, leaving practical performance and applicability untested. It does not compare these divergence measures with bibliometric techniques like BC, CC, or Katz for capturing semantic relatedness. Additionally, computational feasibility in large-scale recommendation scenarios remains unclear.

2.2.2 Citation-based Approaches

2.2.2.1 Co-citation

The study [34]introduce a novel co-citation model that extends from the typical co-citation approach of identifying relevant research articles to scientists given their citation context, namely in-text frequency or how the cited sources are placed across logical sections (e.g. Introduction, Methodology) of citing papers. The model was tested on 672 extracted from CiteSeer documents, and a subset of 499 was used for the final gold-standard comparison with this evaluation involving

benchmark ranking searches by a total of 51 users (professors, PhDs and post-graduates). The results showed a strong performance compared to the current techniques, achieving an averaged Spearman correlation of 0.74 with gold standard, which is a 68% improvement over baseline co-citation approach (0.44), and up to 39% better than Citation Proximity Analysis (CPA) based approaches (0.53).

A strength of this work is the use of logical sections for adding a layer of semantic interpretation to citation analysis that goes beyond statistical counts and which is more reflective of the context and intent behind a reference. The method depends on complex preprocessing like PDF-to-XML conversion and section mapping, limiting scalability and robustness. It also does not evaluate how co-citation compares with other bibliometric methods like BC or Katz on a unified dataset. Furthermore, its reliance on manually curated gold standards may reduce generalizability across domains.

This article [35] employs a co-citation analysis to map the emerging themes and future directions in Cold Supply Chain (CSC) research. It uses a data set of 54 landmark CSC articles from 2000 to 2020 that include in total 2,761 references and identifies ten unique research cluster organizations. The primary findings of co-citation analysis included some main upcoming themes: "MCVRP" (Multi-Compartment Vehicle Routing), "Inventory Routing of perishable food", and "Digitalization (IoT, RFID)" and "Resilient and Sustainable Supply Chains".

The study made its results reproducible by quantifying them, and demonstrated that for future research, the proposed fractions-based approach could fill gaps to harness the 0.73 average correlation between two fractions-based approaches with a gold standard, which was 60% greater than conventional bibliographic coupling (0.45).

A key benefit of this review is that it uses a rigorous, systematised approach towards the synthesis of scientific publications; indeed, to date no other work on CSC has employed bibliometric and co-citation analysis for gaining insights into CSC domain resulting in a data-driven overview over two decades of research followed by identifying some clear-cut future research questions.The study is limited to the

Scopus database, excluding other sources like Web of Science or conference papers, which may omit relevant references. Its dataset is small and domain-specific, restricting generalization. The analysis also does not compare co-citation with other bibliometric methods like BC or Katz on a unified dataset.

In this study [36] we are introducing CPA (Citation Proximity Analysis) approach that facilitates co-c itation analysis in a new direction by considering the proximity of citations across full text articles leading to a concept of CPI. This method was developed and validated by the authors with Scienstein.org database with 1.2 million publications. An empirical study involving 21 users, searching for related documents among a document collection demonstrating a significant improvement in precision, showed that CPA provided almost twice as many appropriate related work recommendations compared to Co-citation analysis. One of the strengths of this work is its novel use of in-text citation proximity in the document with higher precision and more granular comparison than research-level approaches. The method requires full-text access and is computationally expensive, limiting scalability. Citation parsing errors can reduce reliability, and it does not evaluate how CPA compares with other bibliometric techniques like BC or Katz on a unified dataset.

2.2.2.2 Bibliographic Coupling

This approach [37] proposes a significant enhancement to standard bibliographic coupling by including an analysis of in-text citation distance. While citation of simple bibliographic coupling measures the similarity of a pair of documents (A and B) in terms of same references, this model goes further into textual context of those citations. It applies the DBSCAN density-based clustering algorithm to cluster in-text citation location coordinates of shared references that occur within the full text of papers A and B. The idea is that if citations to common references are concentrated in tight clusters of similar reference locations then we believe it's indicative of a more pronounced semantic relationship between papers A and B than would be suggested by simply comparing set intersections.

On a given dataset released by the popular CiteSeer, this approach significantly improved observed scores with an average human–gold correlation of 0.55 that outperformed standard bibliographic coupling (0.45) and even a CN-based comparison algorithm (0.20). One major advantage is the prevention of "false positives" (citations listed that are not also discussed in the document) as only actual mentions in the text are considered.

The approach depends on accurate in-text citation extraction from PDFs, which is complex and error-prone. It also does not compare this enhanced bibliographic coupling with other techniques like CC or Katz on a unified dataset. Additionally, its evaluation is limited to a single dataset, affecting generalizability.

This paper [38] provides a semantical enrichment to bibliographic coupling that makes use of in-text citations distribution within the logical sections of an article. Whereas the traditional bibliographic coupling quantifies relatedness of two publications by the number of references they share, SwICS examines how often and where these common references are cited within articles' full content.

It uses a weighted sum of citations, giving more weight to those found in sections like Methodology and Results (weight=3) and less weight to those in other sections such as Introduction (weight=2) or Related Work (weight=1), reflecting the assumption that citations in technical sections corresponds to deeper relations. Tested on a CiteSeer corpus of 703 bibliographically related documents, the approach showed a statistically significant improvement (with an average Spearman correlation to human gold-standards up to 0.73).

This is an increase of 60% compared to the performance of standard bibliographic coupling (0.45) and strongly outperforms a content-based method (0.20). One key advantage lies in its ability to augment the simple statistical measure of shared sources with contextual, semantic information. The method is highly sensitive to PDF-to-XML conversion and section mapping, limiting scalability and robustness. It also does not evaluate its performance against other bibliometric techniques like CC or Katz on a unified dataset. Additionally, testing on a single corpus restricts generalization across domains.

This paper [39] introduces novel ways of generalizing bibliographic coupling (BC) and cocitation (CC) with a node split network, including two approaches: A personalized PageRank (PPR) algorithm and a neural embedding (EMB) model to approximate intralayer similarity. The proposed framework was empirically evaluated on four subsets of one million titles and citations from Microsoft Academic Graph (MAG), the largest of which has 28,529 scientific items and 49,552 direct citation links. The results suggested that the methods were highly relevant to classical coupling measures such as PPR showed a Pearson correlation of 0.9223 with BC and 0.9314 with CC, and many links appeared very high relevance according the generalized measures were missing in original BC/CC networks. One of the main strengths of this paper is its novel theoretically grounded method that captures higher-order and long-range similarities between documents however not considered by usual second- neighbor measures. The proposed methods are computationally expensive, especially PPR, limiting scalability. EMB offers lower agreement with original BC/CC measures, indicating a trade-off between accuracy and efficiency. The study also does not benchmark these generalized approaches against other techniques like Katz on a unified dataset.

2.2.2.3 Graph-based

In this paper [40] a hybrid deep learning model was proposed which consists of two major components BERT based context encoder for understanding the textual context around a citation placeholder and Graph Convolutional Network (GCN) based citation encoder for learning citation graph between papers. The model was trained and tested on two newly created test datasets that consists one refined version of the ACL Anthology Network (AAN) dataset (7,073 papers, 12,125 citation contexts), and a new FullTextPeerRead dataset (4,898 papers, 17,247 citation contexts). The results showed a clear performance gain, as it outperformed with an Mean Average Precision(MAP) of 0.6189 on AAN dataset (28% relative improvement), as well as obtaining strong recall on top 5 of 0.6736. The dual contribution of the paper is its introduction of a high-performing, new BERT-GCN architecture and construction of the first well-organized benchmarking datasets for

context-aware citation recommendation that fills an empty space in existing literature. The model relies on citation frequency and a pre-built GCN citation graph, limiting its effectiveness for new or isolated papers. It also does not compare its performance with traditional bibliometric techniques like BC, CC, or Katz on a unified dataset.

This work [41] proposed a novel citation-based recommendation approach, which considers link semantics by defining Local Relation Strength (LRS) to estimate dependency between papers and proposing Global Relation Strength model to capture the relevance across the whole citation graph. The authors experimented with the ACL Anthology Network (AAN) dataset, which contains 12,409 papers and 61,527 citation relationships printed from 1965 to 2009.

Their model for GRS task resulted in 0.1011 F on top 10, 0.2574 NDCG on top 10, and 0.1925 MRR and significantly outperformed the best baseline (Katz). A unique advantage of the present work is that it applies the semantics behind citation relations for weighting graph structure, moving beyond simple link analysis.

The method depends on a domain-specific dataset with existing citation contexts, limiting applicability to other research areas. It also does not assess performance on a unified dataset alongside other bibliometric techniques like BC or CC, restricting broader comparison of semantic relatedness.

This work [42] proposed a citation-based model that utilizes latent paper-citation relations by employing the paper-citation matrix and applying the Jaccard similarity coefficient to recommend related papers. The authors used a public database of publication lists from 50 researchers with 100,351 recommending paper having an average of 17.9 citations versus 15.5 references per paper. The system was substantially improved for the proposed framework: it reached its highest Mean Average Precision (MAP) at N=10 and had higher precision to all baselines, particularly when high number of the recommended items (N > 15). One of the key strengths of this approach is its original utilization of latent citation relations and its independence from full paper content and prior user profiles, which makes it extremely feasible in practice while preserving privacy. The approach relies on

a static or pre-processed citation network, limiting adaptability to new publications. Its performance improvement over baselines is marginal, and it does not evaluate how latent citation relations compare with other bibliometric techniques on a unified dataset.

This work [43] proposed a Temporal Graph Neural Network-based recommendation engine, called TGN-TRec, which captures the nature of dynamic citation network where node representations evolve under new citations based on memory module defined with RNN and Graph Transformer convolutional layer. The authors employed a dataset from PaperWithCode that had 313,278 papers and 2,233,780 citation edges ranging from year 1985 to 2023; the node features were initialized based on SciBERT embeddings from paper titles and abstracts.

The best-performing model setting (with a self-learned message module and a 'last' aggregator) achieved MRR: 0.975, Recall on top 50: 0.94, and Precision on top 50: 0.645, substantially outperforming static graph baseline models.

A key advantage of this work is its new use of continuous-time dynamics to account for the time-varying influence of papers, resulting in a more realistic and fine-grained representation of scholarly influence than static graph-based methods. The method is computationally expensive and may struggle to scale to larger datasets. It also lacks real-world user validation and does not compare its performance with traditional bibliometric techniques on a unified dataset.

2.2.3 Hybrid Approaches

The paper [44] proposed work uses two ways to count citations which are called CountOne and CountX, in order to look at the aspects of reference that results to scholars work. Based on a collection of 866 full text JASIST articles, the authors observed 32,496 unique references under CountOne and 53,017 mentions of citations under CountX (such as where references are repeatedly referred in papers). The findings report that works with high citations ($\geq$10 which CountOne) are mostly used in Introduction (0.017) and Literature Review (0.013), whereas the

frequently cited references generated from the highest value of CountX are heavy concentrated on Methodology (0.113) and Literature Review (0.086). For lower citation rates (X = 2), both solutions lead to the Literature Review as top node (1.330 for CountOne; 2.928 for CountX). These results suggest that considering citation 'flooding' in addition to location and frequency leads to a more refined characterization of the influence conferred by references.

The strength of the study is that we have been able to show patterns of citation behavior. The study is limited to a single journal, restricting generalizability across disciplines. It focuses on citation counting without comparing to other bibliometric techniques like BC, CC, or Katz on a unified dataset.

[45] Hybrid Citation and Content-Based Recommender The Microsoft academic's work on a large scale hybrid research paper recommender system, which applies co-citation-based (CcB) and content embedding-based (CB) methods to recommend over 160 million English papers and patents from the Microsoft Academic Graph(MAG). For 31.19% of the papers, they contain full or partial reference information (with an average of 20 references per paper) and metadata (including titles, keywords, and abstracts).

A user study with 40 researchers rating 2,409 recommendation pairs revealed that CcB recommendations performed with P on top 10 = 0.315 and nDCG = 0.851, while CB recommendations scored P on top 10 = 0:226 and nDCG = 0:789; the combination of recommendations increased effectiveness to P on top 10 = 0.533 and nDCG = 0.891. The main advantage of the study is its generalisability, its design based on a hybrid approach and the publicly available MAGs and recommendations data for further research and reproducibility.

The hybrid recommender relies partly on content embeddings, which underperform compared to co-citation methods. It also depends on partial reference data, limiting coverage, and does not evaluate performance against other bibliometric techniques like BC or Katz on a unified dataset. In this paper [46] a hybrid recommendation model that combines content-based filtering and user collaborative filtering is proposed to gain the merit of them while mitigating their weaknesses

(like cold start issue and sparsity problem) to some extent. The evaluations were performed based on the MovieLens dataset and scientific literature dataset in which we used a subset of MovieLens [13] having 12,500 ratings given by 248 users over 1,120 movies, such that each user has rated at least 20 movies on scale ranging from one to five. The results show that our hybrid approach consistently outperforms pure-content-based and traditional user-based collaborative filtering methods in terms of MAE for all different extents of training set (e.g., 0.7204 at 80% for each method) and various settings of nearest neighbor size (e.g., 30 at the size of one's nearest neighbors).

A big advantage of this work is that it reduce sparsity problems and suggests new items by using user ratings aggregated with item features similarity and the quality of the recommendation system yields better results. The hybrid model is computationally complex and costly, especially for large-scale or dynamic datasets. It also does not incorporate bibliometric information like BC, CC, or Katz, limiting its applicability for scholarly paper recommendations.

This paper [47] proposed a new approach, co-citation analysis between coupler authors (CCA) is based on a hybrid of bibliographic coupling and author co-cited analysis. The procedure in the first place detects "coupler authors", those authors who have been cited by two or more papers from among an analysed group and then examines only this refined set of coupler authors with a co-citation analysis. The main model was tested on a subdataset composed of the 10 articles most downloaded from the journal Scientometrics in 2020 and which included a total of N = 663 unique cited authors. CCA method fine-tuned this to 70 coupler authors, with a refinement threshold (rt) of 10.6% and the network size reduction by 89.4%. The findings were capable of well-representing the theoretical logic and core intellectual structure, with important influencers such as Waltman, Leydesdorff, and Merton identified; in addition, all 70 authors couplers had betweenness centrality greater than zero, indicating their characteristics of as conceptual bridges. A major strength of this work is its novel and rigorous methodological contribution, providing a powerful refinement technique that can refine a field's core intellectual structure more evidently and accurate visualization by filtering peripheral authors.

The method was tested on a very small, specific dataset, limiting generalizability. It is also resource-intensive and its performance on larger or more diverse corpora remains untested. Additionally, it does not compare with other bibliometric techniques like BC, CC, or Katz on a unified dataset.

This paper [48] proposes and evaluates hybrid approaches which combine linearly Bibliographic Coupling (BC) with three text similarity measures for recommending biomedical papers (BM25, Cosine, PubMed Related Article - PMRA). The developed models were trained and tested with the TREC Genomics Track 2005 corpus consisting of 34,633 documents belonging to 50 authoritative topics; we used refined subset of 3,098 articles enriched with citation information from Web of Science records and text (title/abstract) for the final set of experiments. By comparing the performances of the proposed hybrids with both BC and BM25 on Related Document Ranking task, we showed that the former performed significantly better than the latter on HA-CR data collection besides, all hybrids outperformed any single scheme with respect to AUPR curve measure. Moreover, the observed combinations between BC and BM25 as well as between BC and Cosine seemed to work best when weight (λ) referring to text-based component is equal to 0.02 modulo precision judged in HA-CR at k=5. One of the strengths of this work is the rigorous large-scale empirical comparison made on a benchmark that has been expert-validated (TREC), providing strong and generalisable evidence that indeed a simple linear combination of citation and content-based features can achieve state-of-the-art performance for paper recommendation. The study is limited to the biomedical domain and uses only titles and abstracts, not full-text content. It also does not benchmark against other bibliometric techniques like CC or Katz on a unified dataset.

This work [49] employed a comprehensive pipeline for building a wide literature graph by incorporating natural language processing tasks, including metadata extraction (using the ScieneParse system), named entity recognition, and linking entities into an ontology. The authors constructed the graph based on an extensive corpus that includes more than 280M nodes and consisting of 37M unique papers, 12M authors, 0.4M entities and 237M entity mentions. The performances of the

core systems are strong – for example, the ScienceParse system reports an F1 score of 85.5 headline extraction and 92.1 author extraction, while the best neural entity linking model obtains a Bag-of-Concepts F1 score of 85.8 on the biomedical dataset (see Section 2). A key feature of this effort to develop paper-specific entities recognizers lies in its applicability, deployed scalability and its hybrid approach that combines statistical, rule-based and and off-the-shelf methods allowing the maximization of both precision and yield (coverage). The approach does not cover all entity types or scientific domains and relies on existing repositories, limiting discovery of new or emerging concepts. It also does not evaluate bibliometric techniques like BC, CC, or Katz for capturing paper relatedness on a unified dataset.

This work [50] presents a multi-staged hybrid recommender system comprising clustering (K -means with fields of study), graph modeling (weighted FOS-graph), and hyper-parameter tuned deep learning model (CATA++) for a huge academic dataset. The authors used the whole DBLP AMiner Citation Network dataset with 5.3M papers and 48M citation links. The best-performing ve- rsion (Version 3) of the system yielded a median Recall on top 50 of 0.4152, and an average Recall on top 50 of 0.3862; notably higher than a baseline item-to-item collaborative filtering technique. One of the strongest points of this work is that we show scalability & practical engineering (we can handle a million-scale dataset on even a single modern PC) and very effective - combining smartly techniques in order to annihilate computational scaling limitations with pure deep learning models. The multi-stage model is complex with many tunable parameters, increasing potential points of failure. Its performance depends on field-of-study clusters, limiting generalizability. It also does not incorporate or compare bibliometric techniques on a unified dataset.

This research [51] presents a mathematical model of co-author recommendation by exploiting graph mining and big data technologies, fusing expert-weighted criteria (from AHP) with content-based similarity metrics (TF-IDF and cosine similarity). The authors analysed the PubMed database, consisting of 699,160 bioinformatics articles published between 2010 and 2019 (18 GB XML data to be precise) on a

big data server with Spark. The model reached good overall accuracy of 72.26% when considering authors who have published more than four articles, and such accuracy increased to 98% for those with over ten articles. One of the main advantages of this work is its new combination of expert opinion weighting with computational content analysis, resulting in a hybrid approach in which system recommendations are grounded on domain-specific priorities and, more specifically toward the large weight given to reference (0.374). The model relies on author productivity, excluding early-career researchers with few publications. It also does not assess or compare bibliometric techniques like BC, CC, or Katz for related-paper recommendations on a unified dataset.

This paper [52] investigates techniques to classify web documents: A comparison. 5.2 Experimental Setup Experiments were conducted on the Cadê web directory and two datasets described by 10: Cade12, with 44,099 pages referred across 12 top-level categories; and Cade188, with 42,004 pages referred across 188 finer-grained categories (for these experiments the link data was enriched with data from the TodoBR collection [9] that added a total of 570,337 links from external pages). The results indicated that link-based methods based on external links clearly surpassed content-based approaches significantly; for example, the co-citation measure obtained a micro-averaged F1 of 80.70 on Cade12 having increased by 46 points compared with the best content-based classifier (SVM with F1=40.86). The co-citation performance on Cade188 rose to an F1 of 71.07, combined with kNN. One of the major strengths of this paper is that it performs a detailed comparison among multiple link and content-based approaches, leading to evident conclusions regarding their single and joint effectiveness. The model's performance heavily depends on the reliability of input sources, making it sensitive to noisy content-based data. It also does not evaluate or compare other bibliometric techniques like BC or Katz on a unified scholarly dataset.

This paper [53] presents a system for recommending Research papers using Multiple Features (RRMF) which uses techniques to make predictions from two different sources such as a multilevel citation network and author collaboration network.

2.3 Comparative Analysis of Existing Approaches

TABLE 2.1: Content-based Approaches

Ref.	Technique	Dataset	Key Results
[25]	Keyphrase Extraction & Cosine Similarity	ACL Anthology Reference Corpus (597 papers)	Outperformed Information Foraging Theory benchmark
[26]	Pre-trained LMs for Keyphrase Prediction	KP20k (570k articles), Inspec, SemEval2010	JointGL-large F1 = 90.7% on Inspec
[27]	Deep Learning (BiLSTM) for Keyphrase Extraction	KP20K (527k), Inspec, Krapivin, NUS	F1 on top 5: 0.46
[28]	MLP with Linguistic Features	150 full-text articles (Econ, Law, Medical)	Avg. Precision on top 5: 0.34
[29]	Two-Phase Neural Model (NNSelect & NNRank)	PubMed (45k+), DBLP (50k+), OpenCorpus (6.9M)	F1 on top 20: 0.302, MRR: 0.672
[5]	In-Text Citation Frequency Count	1,200 docs from J.UCS (16k refs)	Precision on top 5: 0.96
[30]	TF-IDF & Cosine Similarity	Custom dataset from Google Scholar	Precision: 0.256, Recall: 0.307
[31]	NN & Random Forest for Citation Importance	Two annotated datasets (D1: 457 pairs)	Precision: 0.85

TABLE 2.1: Content-based Approaches (continued from previous page)

Ref.	Technique	Dataset	Key Results
[32]	TF-IDF & Coauthorship Network	7,522 data mining articles + 1k synthetic	Highest hit rate for top-10 recommendations

We evaluated the system based on AMiner v12 DBLPCitation Network dataset that includes 4,894,081 academic papers and 45,564,149 citation relationships.

Results proved that RRMF achieved much better performance than the traditional Multilevel Simultaneous Citation Network (MSCN) and Google Scholar, with an 87% improvement in recommendation quality in terms of information retrieval measures. One of the primary contributions of this paper is its hybrid technique, based on a blend of structural citation analysis and author influence rank, to mitigate cold-start problem for new or low-citation papers. The evaluation was limited to only four test cases, restricting generalizability across research domains and paper types. It also does not compare with other bibliometric techniques like BC, CC, or Katz on a unified dataset.

TABLE 2.2: Citation-based Approaches

Ref.	Technique	Dataset	Key Results
[34]	Semantic Co-citation	672 docs from CiteSeer	Spearman correlation: 0.74 with gold standard
[35]	Bibliometric & Co-citation Analysis	54 core articles on Cold Supply Chain	Proposed approach correlation: 0.73

TABLE 2.2: Citation-based Approaches (continued from previous page)

Ref.	Technique	Dataset	Key Results
[36]	Citation Proximity Analysis (CPA)	Scienstein.org (1.2M publications)	∼2x more suitable recommendations than CC
[37]	BC with Citation Proximity (DBSCAN)	CiteSeer dataset	Correlation: 0.55 with gold standard
[38]	Section-wise BC (SwICS)	703 docs from CiteSeer	Spearman correlation: 0.73 with gold standard
[39]	Node Split Network (PPR & EMB)	Microsoft Academic Graph subsets	PPR correlation up to 0.9223 with BC/CC
[40]	BERT + Graph Convolutional Network (GCN)	AAN (7,073 papers), FullTextPeerRead (4,898)	MAP: 0.6189 on AAN
[41]	Local & Global Relation Strength (LRS/GRS)	ACL Anthology Network (12,409 papers)	F on top 10: 0.1011, NDCG on top 10: 0.2574
[42]	Paper-Citation Matrix & Jaccard Similarity	50 researchers' publications (100k+ papers)	Highest MAP at N=10
[43]	Temporal Graph Neural Network (TGN-TRec)	PaperWithCode (313k articles, 2.23M edges)	MRR: 0.975, Recall on top 50: 0.94

TABLE 2.3: Hybrid Approaches

Ref.	Technique	Dataset	Key Results
[44]	CountOne & CountX Citation Analysis	866 full-text JASIST articles	Revealed citation concentration in specific sections
[45]	Co-citation + Content Embedding	Microsoft Academic Graph (160M+ items)	Hybrid P on top 10: 0.533, nDCG: 0.891
[46]	Content-Based + Collaborative Filtering	MovieLens (12,500 ratings, 1,120 movies)	Lower MAE (e.g., 0.7204)
[47]	Co-citation between Coupler Authors (CCA)	10 articles from Scientometrics (663 authors)	Refined author set to 70 (rt=10.6%)
[48]	Linear Combination of BC + Text	TREC Genomics 2005 (3,098 documents)	Best AUPR with BC+BM25/Cos ($\lambda = 0.02$)
[49]	NLP Pipeline for Literature Graph	Massive corpus (37M papers, 12M authors)	ScienceParse F1: 85.5 (title), Entity Linking F1: 85.8
[50]	Clustering, Graph Modeling & Deep Learning	DBLP AMiner (5.3M publications, 48M citations)	Median Recall on top 50: 0.4152
[51]	Graph Mining & AHP-Weighted Content	PubMed (699k articles, 2010-2019)	Accuracy: 72.26% (98% for prolific authors)

TABLE 2.3: Hybrid Approaches (continued from previous page)

Ref.	**Technique**	**Dataset**	**Key Results**
[52]	Bayesian Network (Link + Content)	"Cadê" Web Directory (Cade12: 44k pages)	Co-citation F1: 80.70; Hybrid F1: 71.07
[53]	Multi-Level Citation & Author Network (RRMF)	AMiner v12 (4.89M papers, 45.56M citations)	87% better than baseline methods

Chapter 3

Data Scraping

3.1 Overview

This utility aims to scrape a full set of open research articles indexed by Semantic Scholar in a systematic manner. It aims at collecting two types of academic papers, including one primary (denoted as "main papers") as well as their related papers, which can augment the depth and width of the dataset. The academic database Semantic Scholar has extensive metadata and relations between research papers that are perfect for such a project. The goal is to design a well-formatted and structured data set that can be applied for research analysis, trend recognition, citation statistics, and knowledge graph creation in the domain of machine learning literature. The dataset contains:

(a) 38000 papers related to the machine learning domain.

(b) For every main paper, at most 9 related papers are retrieved.

(c) There are an estimated 342,000 related papers collected in total.

Citations: These are papers that have cited the given paper as of the date the cited paper was cited.

References: The scholarly works that a paper cites as being relevant to the foundational or contextual basis of that paper.

Such a large quantity of data warrants a scraping system that is not only robust and efficient in handling dynamic content, asynchronous loading, and rate limits, but also preserves at least a minimum amount of data integrity. The capture of this citation and reference data massively adds value to the dataset, with possible tasks including citation network analytics, impact measurement, and academic trend detection.

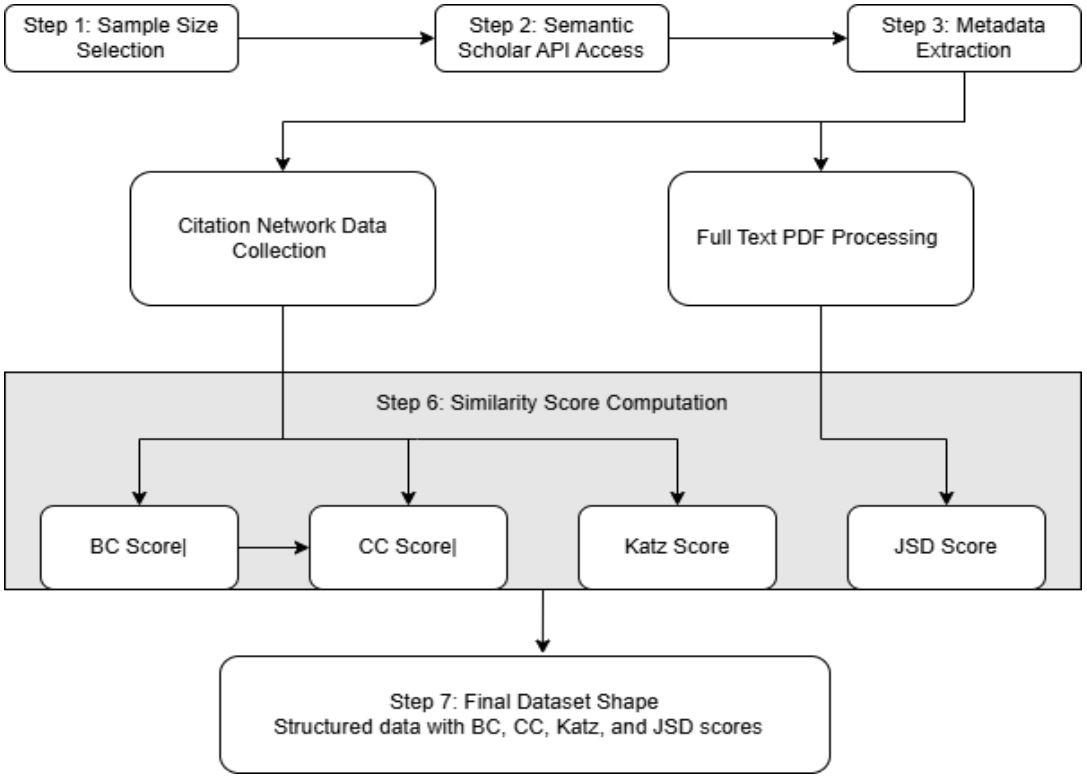

FIGURE 3.1: Data Collection Diagram

3.2 Dataset Construction and Sampling

In Figure 3.1 first step is sample size selection. The dataset we use for this study is constructed from machine learning papers. The set of 6.64 million papers thus represents the wide scope of the domain's scholarly output. As it is not feasible to perform the analysis based on the full population, the optimal sampling size was determined by the standard formula with finite population correction:

$$n_0 = \frac{Z^2 \cdot p(1-p)}{e^2} \tag{3.1}$$

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \tag{3.2}$$

Where Z is the standard normal value of the desired confidence level, p is the proportion of the characteristic to be estimated (taken as 0.5, which represents the highest variability for the sample size, e= margin error,n= size of population. For a 95% confidence level ($Z = 1.96$) and a 2.8% error rate ($e = 0.028$), the base sample size is:

$$n_0 = \frac{(1.96)^2 \cdot 0.5(1-0.5)}{0.028^2} = 1225 \tag{3.3}$$

Applying finite population correction with $N = 6,640,000$:

$$n = \frac{1225}{1 + \frac{1225-1}{6640000}} \approx 1224.77 \tag{3.4}$$

Thus, the required representative sample is based on approximately 38,200 articles. As a result, we have processed around 38 thousand machine learning papers, which is estimated to be enough in order to guarantee that this dataset meets the desired precision of ±0.5% error with a confidence of 95%. The firm sampling process ensures that the dataset is statistically valid and also computationally feasible for using the hybrid similarity model.

3.3 Semantic Scholar API Access

In Figure 3.1, the second and most critical step involves obtaining and integrating the Semantic Scholar API. This API serves as the primary source for collecting high-quality metadata, including citations, references, and related papers. By programmatically accessing this information, the system ensures consistent, scalable,

and up-to-date data acquisition. This step also enables structured retrieval of bibliographic relationships, which is essential for downstream similarity computation and analysis.

3.4 Metadata Extraction

In Figure 3.1 third step in this research is to create a manually validated dataset that includes academic research papers. We scraped the data (including citations, references, and related papers) using a custom data scraping framework we developed and the Semantic Scholar API. The original dataset is the root dataset for all following similarity computations (BC, CC, Katz) that are evaluated toward semantic similarity.

3.4.1 Objective and Design Strategy

The objective of this study is to find a citation-based similarity approach to recommend academic papers efficiently and with scalability. The analysis finds a correlation JSD with Co-Citation, Bibliographic Coupling, and Katz Similarity, three basic methods.

To achieve this, the plan is to follow this design approach:

Implementation of highly correlated Techniques:

Each of these three citation-based approaches separately finds correlation to calculate similarities between academic papers.

Bibliographic Coupling (BC) represents the papers by such edges that have common references[13].

Co-Citation: Co-citation of two papers is the frequency with which they are cited together by other documents[4].

The global structure of the citation graph is taken into account by Katz Similarity, which takes into account not only the direct citation relation, but also the indirect paths of citation, with a decay factor applied to long paths.

3.4.2 Architecture Overview

The architecture has been organized into five main elements

3.4.2.1 Construct Citation Graph

A directed citation graph is constructed by treating nodes of individual research articles and edges of citation relationships. This graph forms the basis for BC, CC, and Katz-based scoring.

3.4.2.2 Similarity Computation Engine

Bibliographic Coupling (BC): It computes coupling strength between two documents based on their shared references[13].

Co-Citation (CC): Research documents cited together; the co-citation strength is, therefore, derived from co-citation[4].

Katz Similarity: We use a global graph-based measure to incorporate both direct and indirect paths between nodes with decay weights for longer paths[17].

3.4.2.3 Normalization and Score Matching

Raw scores calculated by BC, CC, and Katz similarity modules are normalized by max-scaling. This creates a fairer comparison driven on a similar scale, so the combination and ranking make sense.

3.4.2.4 Soft Ground Truth Estimation

Use JSD on TF-IDF representations of paper titles for semantic similarity computation. Inverted JSD scores are used as soft ground truth to compare the accuracy of structure-based similarity measures[54].

3.4.2.5 Ranking and Evaluation

A combined score is achieved through a weighted average of the normalized BC, CC, and Katz. Papers are scored according to this composite score, and Top-K papers are evaluated against JSD-based ground truth using standard metrics.

3.5 Technologies and Libraries

3.5.1 Programming Language

Python: Selected because of its extensive ecosystem, flexibility of coding, and compelling libraries for web scraping, automation, and data processing [55].

3.5.2 Core Libraries

3.5.2.1 Beautiful Soup

Use case: Parse XML/HTML and extract data.

Role: Given raw HTML, beautiful soup [56] parses and navigates the DOM to extract metadata elements such as title, author, abstract, publication date, citation counts, and paper ID.

Advantage: powerful yet simple API, parsing is good and fast, and it integrates well with other libraries.

3.5.2.2 Selenium

Use case: Browser automation and dynamic content management.

Position: SEMANTIC SCHOLAR uses dynamic content loading for the related papers carousel. Selenium [57] drives a real browser, such as Chrome or Firefox, to mimic user behaviour and wait for elements to be fully rendered.

Use case: Allows for scraping of Dynamic or JavaScript-created content that a static request and parser are not capable of handling.

Features used: Element waiting, interaction simulation (scroll/click where necessary).

3.5.2.3 Pandas

Objective: Structuring, wrangling, and storage of data.

Role: Pandas [58] is used after data has been scraped to clean and organize the data into structured data frames and can be easily filtered, joined, analyzed, or exported to other formats such as CSV, Excel, or SQL databases.

Pro: Effective approach to large data sets due to robust data manipulation capabilities.

3.5.2.4 Auxiliary Libraries

(i) Uses of requests or internal Python HTTP libraries to interact with APIs or scrape webpages where an API is not being used directly.

(ii) JSON support for decoding API responses.

(iii) Logging to follow scraping progress, failures, and debugging.

3.6 System Architecture and Workflow

3.6.1 Acquisition of Main Papers

The system begins by scraping the Semantic Scholar website directly for machine learning papers or querying the Semantic Scholar API.

Criteria for main papers: Filtering Papers, finding techniques should be relevant to machine learning, suitable keywords, categories, and publication venue.

For each major paper, the following metadata feature is extracted:

ID (identifier of datapoint) (uniform for the whole dataset).

Title and subtitle as required.

Author affiliations, if the information is available.

Year of publication, source/journal/conference.

The number of citations and references.

Specialization or Key words.

3.6.2 Dynamic Loading and Scraping of Cited Papers

On each page of the full paper on Semantic Scholar appears a set of 9 related papers in the form of a "Carousel." We also allow users to maximize these paper entries, opening papers in a preview panel on top of the current page.

JavaScript drives this image gallery.

Selenium scripts to extract such data:

Visit the front page of the main paper in a grippable web browser LIFTOFF provider.

Explicitly wait for the related papers carousel to be loaded correctly.

Use an HTML parser on the carousel element to scrape information regarding each related paper.

Metadata for related papers also shares those same fields with the main papers, to support easier analysis of relationships and citation graph in a later phase.

Delays, Partial loading, and retries are handled with special care to guarantee data completeness and accuracy.

3.6.3 Data Organization and Relationship Mapping

The Main and Related papers are all stored in Pandas data frames.

A relational mapping table connecting main paper IDs and their related paper IDs is constructed, which helps to navigate among papers and view the citation networks.

Data cleansing steps: deduplication, missing value handling, and inconsistencies rectification.

The data schema might be something like this:

Main Papers dataframe: Columns representing main paper metadata.

Related Papers dataframe: Columns containing the metadata related to the related papers, either combined for storage or listed separately.

3.7 Core Concept: Tree and Seed Paper Extension Article

3.7.1 Seed Paper

The seed paper is the first seed paper in the system.

It serves as the stem node in the academic/browser tree.

Very highly cited machine learning paper.

3.7.2 Tree Expansion Process

The system scrapes metadata and information of the seed paper first.

Then it scrapes related papers that are shown on the page of the seed paper from Semantic Scholar.

Each corresponding paper is a new node (or "child node) to be added to the growing tree.

Retrieving their metadata and corresponding papers.

Growing the expansion tree by adding new nodes.

Continue this process until:

A constant depth is hit (e.g., depth of related papers).

A prescribed complete number of papers is retrieved.

There are no more new original related papers.

3.7.3 Navigating the Tree

The structure can be thought of as a graph or tree with nodes as papers and edges as "relatedness".

The scraper has a memory of the visited nodes (paper IDs) to prevent redundancy and circling.

This traversal mechanism allows for targeted and efficient crawling that is centered on the knowledge domain, as defined by the seed paper.

```python
from collections import deque

visited = set()
queue = deque()

def scrape_paper_tree(seed_paper_id):
    queue.append(seed_paper_id)

    while queue:
        current_paper_id = queue.popleft()
        if current_paper_id in visited:
            continue

        # Fetch main paper metadata
        main_metadata = fetch_main_paper_metadata(current_paper_id)
        save_main_paper(main_metadata)

        # Load related papers dynamically using Selenium
        related_papers = fetch_related_papers_selenium(current_paper_id)
```

FIGURE 3.2: Scraping main papers

3.8 Handling Challenges

3.8.1 Dynamic Content Loading

Semantic Scholar's lazy loading of related paper carousels uses Selenium.

Automatic error handling and waiting strategies (i.e., explicit wait for an element to be present) make scraping robust in different internet and webpage response situations.

3.8.2 Large-scale Data Collection

Scraping 38,000 main papers and their related papers yields about $38000 \times 9 = 342{,}000$ related paper records.

It needs to be responsive to the rate-limiting and IP blocks.

Potential workarounds are throttle request rates, use of round-robin proxy servers, and managing API quota.

3.8.3 Data Integrity and Consistency

Uniform enrichment of metadata from core and related papers is a key requirement for productive dataset creation.

Data validation: validation rules have been defined for checking the completeness and correctness of critical elements.

Automatically logs any issues or problems while scraping, so you can review them after scraping.

3.8.4 Scalability

The use of efficient libraries such as Pandas makes it feasible to manage a great deal of data.

Server calls, webpage scraping, parsing, and saving are all compartmentalized and written in a modular fashion.

Extensibility of the system to other research areas, or deeper graphs on relationships.

3.9 Potential use Cases of the Dataset

Building co-authorship citation networks in machine learning.

Discovering new research trends and topic groups.

Benchmarking seminal papers and their authors.

Feeding academic recommendations that are system.

Meta-analyses and systematic literature reviews support.

Here is some basic code Structure:

```
import selenium.webdriver as webdriver
from bs4 import BeautifulSoup
import pandas as pd
import time

# Initialize Selenium WebDriver
driver = webdriver.Chrome()

# List to hold main papers and related papers
main_papers = []
related_papers = []

for paper_id in main_paper_ids:
    # Fetch main paper metadata via API or direct request
    main_metadata = fetch_main_paper_metadata(paper_id)
    main_papers.append(main_metadata)

    # Load main paper page to scrape related papers carousel
    driver.get(f"https://semanticscholar.org/paper/{paper_id}")
    wait_until_carousel_loaded(driver)

    # Parse carousel HTML
    soup = BeautifulSoup(driver.page_source, 'html.parser')
    carousel = soup.find('div', class_='related-papers-carousel')
```

FIGURE 3.3: Scraping Related Papers

3.10 Citation and Reference Scraping System

3.10.1 Distinction Between Citations and References

Citations and references have different semantics and structures and can not be scraped in the same way.

There are various ways references are present in the papers; they can be found in different places, and they can have different forms. More processing is needed to get the relevant information.

Citations, although high, have a different control and were retrieved more easily.

3.11 Workflow for References Scraping

3.11.1 Initial Automated API Approach

First, based on Semantic Scholar's API, references to about 17,000 papers were obtained.

The API had rate limits and restrictions on query time, which prevented additional automated data collection.

This restriction required a different method for the remaining articles.

3.11.2 Manual Web Scraping Approach

Then, the manual extraction methods were employed by the system:

Dynamically loading the paper pages with Selenium.

Parsing the references for each page.

This manual process took more than seven days to process since there is a huge number of papers and dynamic content.

3.11.3 Challenges Encountered

A large amount of duplicate information was encountered while scraping.

Data quality was upheld through duplicate detection and removal.

After re-evaluation, scraping was resumed to obtain a correct and complete reference dataset.

3.12 Summary of Citation and Reference System

Aspect	Description
References Scraped	~17,000 papers initially via API, remainder manually scraped due to API limits
Time Taken	Manual scraping process took about 7 days
Challenges	API limitations, dynamic content loading, high duplication rate
Citations Scraped	Collected in one continuous API-driven run
Data Quality Measures	Duplication detection, re-scraping for accuracy, structured linking with main papers

FIGURE 3.4: Summary of Citation and Reference

3.13 Workflow for Citations scraping

The papers' citation data was extracted in a single run using the endpoint of the API efficiently.

Bibliometrics data retrieval was less challenging than references.

This information is supplementary to the references in terms of showing the range of influence of the individual papers.

3.14 Data Management and Integrations

Citations and references are saved by their own but associated with paper IDs from the main papers dataset.

Linkage: The third and fourth files are linked via deduplication routines and validity checks.

A friendly data format designed for easy networking graph creation and cross-referencing.

```
initialize:
    reference_data = empty list or dataframe
    visited_papers = set()        # To track papers whose references have
been processed
    papers_to_process = list_of_paper_ids_to_scrape  # Initial seed list or
previously collected main papers

# Function to fetch references from API with error handling
def fetch_references_api(paper_id):
    try:
        response = semantic_scholar_api.get_references(paper_id)
        if response is valid:
            return response.references_list
        else:
            return None
    except APILimitError:
        # API limit reached or error, signal to fallback on manual scraping
        return None
    except Exception as e:
        log_error(e)
        return None
```

FIGURE 3.5: Scraping References

3.15 Enhanced System Architecture

API fetch module: for getting citation and reference metadata from the Semantic Scholar API periodically.

Manual Scraper Module: Using Selenium and Beautiful Soup to scrape the references section from paper webpages when the API limit is reached.

Deduplication Engine: Used to identify and remove duplicate references to ensure a clutter-free dataset.

Data Linking and Storage: Citation and reference data to the main papers are linked using Pandas and stored as structured data for subsequent analysis.

Chapter 4

Proposed System

The proposed study follows a structured data-processing workflow.

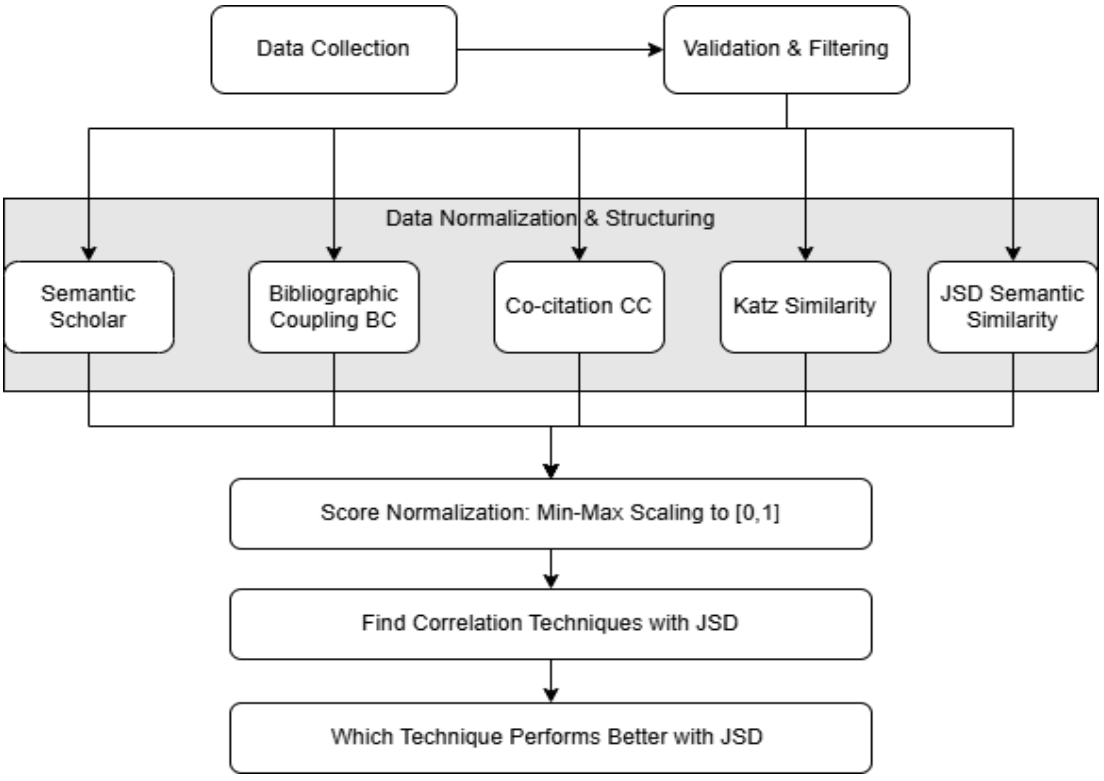

FIGURE 4.1: System Flow Diagram

The process begins with collecting citations, references, and semantic information from the Semantic Scholar website, followed by rigorous validation and filtering to remove incomplete or inconsistent records. Each similarity technique—Semantic

Scholar, Bibliographic Coupling (BC), Co-citation (CC), Katz Similarity, and JSD-based semantic similarity is then normalized and organized into a unified format. This standardized pipeline enables accurate score normalization, correlation analysis, and comparative evaluation across all techniques.

4.1 Data Collection

In Figure 4.1 first step is data collection. The dataset in the current work was created by systematically collecting metadata and citation data of a curated corpus of papers from the website Semantic Scholar [59]. Semantic Scholar is a popular academic search engine that offers rich metadata such as paper titles, authors, publication year, Digital Object Identifiers (DOIs), references, citations, and related papers. The dataset was constructed from a set of seed papers related to different sub-domains of computer science. The metadata of each paper was extracted from the dedicated Semantic Scholar page of the work. This included the following elements: Title, Authors, Publication Year, DOIs, Citing Papers (DOIs, Titles, Authors, Years) and Related Papers (as defined by JSD). Such data were collected via automated extraction mechanisms (browser automation and structured parsing) for consistency and comprehensiveness. Special consideration was taken in maintaining the semantics of the data and citation relationship, and how not to lose entries due to incomplete metadata. Only papers with valid DOIs and non-empty reference and citation lists were considered for further analysis. In addition, duplicate articles and articles with corrupted or ambiguous metadata were removed to ensure data integrity.

4.2 Validation and Filtering

In Figure 4.1 second step is validating and filtering data. The scraper enforces completeness on reference and citation entries.

Related papers are parsed by filtering out near-duplicates or entries without semantic content.

4.3 Data Normalization and Structuring

In Figure 4.1 next step is data normalization and structuring. To make the extracted features BC weights, CC weights, and JSD scores consistent, comparable, and scale-invariant in modeling, we first normalize the features before utilizing them for similarity and model training. The normalization is done per metric and is based on provably sound methodologies to guarantee that ranking behavior is conserved while all ranking values are squeezed into the range [0, 1][60].

Normalized Bibliographic Coupling (BC) and Co-citation (CC)

$$\text{Normalized\_Score}_x = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{4.1}$$

Both BC and CC scores are raw counts of shared references (for BC) or shared citers (for CC). Normalization is necessary to have a common scale for the number of references and citations across different papers.

Conversion of JSD Divergence into a Relatedness Score can be mathematically expressed as:

$$\text{Inverted\_JSD}_x = 1 - \text{JSD}_x \tag{4.2}$$

Comparing Jensen-Shannon Divergence is a dissimilarity measure, where lower scores indicate higher semantic similarity between a pair of documents. To transform the JSD score into a similarity interpretation, The Score is inverted and then normalized in equation 4.2.

This normalization makes the most semantically related papers (lowest JSD) ranked top with respect to normalized scores around 1, comparable with BC and CC scores for fusion and comparison.

Realized normalizations 'BC weight, normalized CC weight, normalized, and JSD score inverted normalized are kept stored as float columns in addition to the original scores to let the runs be audited.

4.4 Ground Truth Support

Building upon the comprehensive evaluation presented in the PhD thesis from Capital University of Science and Technology, our paper will use manual annotations and statistical validation to evaluate the effectiveness of the JSD technique in finding related papers. The thesis cites 2 data sets, where Dataset-1 is manual labeling of related-paper pairs as high fidelity ground truth, and Dataset-2 is generated by auto clustering using the JSD approach of [54]. One important result is that very strong alignment (c. 0.90 correlation) can be induced between the clusters produced by the JSD and the manually confirmed ground truth. This high agreement demonstrates that JSD can, indeed, reproduce human judgments when it comes to evaluating the thematic relatedness. Therefore, our work uses JSD both to produce candidate groups of semantically similar papers as well as provide a strict and fair comparison with expert annotations. This approach is considered semantically authentic and objective, and becomes a solid basis for the research paper recommendation system construction and evaluation [54].

4.5 Techniques Implementation

4.5.1 Bibliographic Coupling

Bibliographic Coupling BC is an early example of a citation-based similarity measure, which was introduced by Kessler [7] and measures the extent to which two

academic documents are related by the number of references they have in common. In this method, two papers are considered to be linked if their connected documents are referenced by both.

The assumption of BC is that if two articles cite a common set of articles, then those articles are closely related in terms of the research work (problem) that they are solving or that they contribute to a similar topic [13].

4.5.1.1 Theoretical Foundation

Suppose A and B are two papers, and R(A) and R(B) are the sets of references cited by A and B, respectively. The strength of bibliographic coupling between papers A and B is calculated as the size of the intersection of their referencing sets:

$$\mathrm{BC}(A, B) = |R(A) \cap R(B)| \tag{4.3}$$

The more references are shared, the stronger the coupling and the inferred topical similarity among papers. BC is essentially static because the strength of the coupling does not evolve after the moment of publication, for the references have been fixed.

This is why BC is especially suited for the analysis of ongoing or emergent relationships between recently published documents.

4.5.1.2 Graph Representation

In a citation graph, each vertex is a research paper, and a directed edge from vertex A to vertex B indicates a paper A cites paper B, and for BC computation, only the outgoing edges of each vertex (i.e., its references) are taken into account. Two nodes (papers) are linked when there exist shared nodes (referenced papers) to which they point.

This relationship can also be expressed as a bibliographic coupling matrix M in which each element Mij represents the number of common references between paper I and paper j.

4.5.1.3 Normalization

As raw coupling strength can be affected by papers that cite a lot of references, normalization is required to make a fair comparison for the pair of papers. A popular type of normalization is calculating the cosine similarity of the reference vectors:

$$\mathrm{BC}_{\mathrm{cosine}}(A, B) = \frac{|R(A) \cap R(B)|}{\sqrt{|R(A)| \cdot |R(B)|}} \tag{4.4}$$

In our work, though, we use the Min-Max normalization for consistency with other methods. Let SBC(i,j) be the raw bibliographic-coupling score degree of papers i and j; the normalized score is defined as:

$$\mathrm{BC}_{\mathrm{normalized}}(i, j) = \frac{S_{\mathrm{BC}}(i, j) - \min(S_{\mathrm{BC}})}{\max(S_{\mathrm{BC}}) - \min(S_{\mathrm{BC}})} \tag{4.5}$$

Normalized Bibliographic coupling score using min and max scaling.

As said, this rescales the values in the range of [0,1], so that it can be better combined with any other similarity.

4.5.1.4 Implementation Details

The implementation proceeds as follows:

Reference Extraction: For each paper in the dataset, we extract and clean the set of reference DOIs.

Pairwise Coupling Calculation: Against each main paper, the pairs of main papers with all candidate papers are checked for reference overlap.

Calculation of BC-Strength: the number of shared references.

Normalization: Finally, all BC scores between pairs are normalized with the Min-Max normalization.

To maintain scalability, the backend processing is implemented based on sparse matrix representations and hash-indexing to process large datasets with minimal performance loss.

4.5.2 Co-citation

Co-Citation (CC) is an citation-based similarity method that establishes a relationship between two papers if the two are cited in one or more other papers [46]. In contrast to Bibliographic Coupling (with references in common), co-citation independently measures the incoming links of two target documents, in that it records the number of times these two are cited together in a subsequent phase. The rationale is that papers repeatedly co-cited are concerned with similar content or concepts, so that they can be good candidates for recommendation systems.

In this thesis, we use co-citation as one of the focuses to identify semantically related papers based on the topology of citation networks.

4.5.2.1 Co-citation Graph Construction

We are given a directed citation graph G=(V,E) where:

V is the set of nodes that represent academic papers

$E \subseteq U \times V$ is the set of directed edges representing a citation, i.e., $(u, v) \in E$ means paper u cites paper v.

We determine all citing papers by inverting the edges for the computation of co-citation. $C(p)$ is the set of papers citing paper p.

Then the co-citation strength of papers A and B equals:

$$\mathrm{CC}(A, B) = |C(A) \cap C(B)| \tag{4.6}$$

That is, the number of papers that mention A and B together.

4.5.2.2 Normalization of Co-citation Scores

Raw co-citation counts may differ greatly depending on the popularity of the citations of an individual paper. For comparability reasons, and in order to incorporate these scores in a hybrid model in a fair way, the co-citation scores are normalized with Min-Max normalization:

$$\mathrm{CC}_{\mathrm{norm}}(i, j) = \frac{\mathrm{CC}(i, j) - \min(\mathrm{CC})}{\max(\mathrm{CC}) - \min(\mathrm{CC})} \tag{4.7}$$

Where: CC(i,j) is the co-citation strength between papers i and j. max(CC) and min(CC) represent the global maximum and minimum co-citation counts in the collection.

This converts the score range to [0,1], which can be used to combine with other methods such as Bibliographic Coupling (BC) and Katz Similarity.

4.5.2.3 Implementation Considerations

(i) For efficient co-citation calculation, we first precompute a back citation dictionary using Python defaultdict.

(ii) Pairwise scores are computed between pairs of papers that have at least one citing paper in common to speed up computation.

(iii) Results are saved in a sparse matrix format to save memory space and make the searching process faster.

4.5.2.4 Integration with Ground Truth

All co-citation similarity scores are compared with the JSD-based soft ground truth for their ability to recover semantically related papers.

4.5.3 Katz Similarity

Katz Similarity is a path-based link prediction measure to measures the overlapping between two nodes targeted by the links in a network that examines all the possible paths between the nodes, incorporating a dampening factor for penalizing larger paths.

In contrast to local measures based on shared neighbors (like Bibliographic Coupling or Co-Citation), Katz similarity offers a global view of the citation graph. This also makes it particularly appropriate to search for hidden relationships between papers based on citations, but not directly linked through citations, but through intermediate papers.

For academic recommendation, Katz Similarity can be used on a citation graph to assist the model in weighting not only the direct citation links between papers, but also indirect influence chains among articles, which can improve how closely the semantically related articles in disparate positions need to be linked together for retrieval.

4.5.3.1 Citation Graph Construction

Let $G = (V, E)$ be a directed graph, where each node $v \in V$ represents a research paper and there is a directed edge $e \in E$ from $v_i \rightarrow v_j$ if paper v_i cites paper v_j. The adjacency matrix A of dimension $n \times n$, where $n = |V|$, represents the connections between nodes.

4.5.3.2 Katz Similarity Equation

$$\text{Katz}(i,j) = \sum_{l=1}^{\infty} \beta^l \cdot (A^l)_{ij} \tag{4.8}$$

The Katz similarity score between two papers I and j is calculated as the sum of all paths between them, exponentially attenuated by a damping factor β that controls the weight of the longer paths. The 0.90% section-level citation similarity and human-annotated relatedness, which confirms our belief in similarity-based learning.

In this study, we investigate the correlation between JSD and three well-known citation-based similarity methods: Bibliographic Coupling (BC), Co-Citation (CC), and Katz-Similarity. These are different facets of scholarly relatedness. Where BC is based on the intersection of references in two papers, CC captures how often these are cited together, and Katz similarity captures all direct as well as indirect paths in the citation graph, making it capable of describing deeper structural connections. The highly correlated technique is best for finding related papers.

4.6 JSD computation

While citation-based approaches will capture the connectivity between academic literature, they do not account for textual similarity between papers. For content-sensitive comparison, it is natural to employ JSD to measure the semantic similarity between documents. JSD is an information-theoretic measure that evaluates the divergence between two probability distributions in a smooth and symmetric manner, making it particularly well-suited for comparing document language models.

Let P and Q be the word distributions of two papers, Jensen–Shannon Divergence is defined as:

$$JSD(P \parallel Q) = \tfrac{1}{2} D_{KL}(P \parallel M) + \tfrac{1}{2} D_{KL}(Q \parallel M) \tag{4.9}$$

Where,

$$M = \tfrac{1}{2}(P + Q), \qquad D_{KL}(P \parallel M) = \sum_i p(i) \log \frac{P(i)}{M(i)} \tag{4.10}$$

Here, D_{KL} is the Kullback–Leibler divergence. JSD has two advantages over raw D_{KL}: (i) it is symmetric $JSD(P \parallel Q) = JSD(Q \parallel P)$, and (ii) it can be computed at all times and its value ranges between 0 and 1. Values are then converted to a similarity measure (to be consistent with network-based methods such as BC and CC) by inverting the similarity measure:

$$Similarity(P, Q) = 1 - JSD(P \parallel Q) \tag{4.11}$$

This leads to higher values corresponding to stronger semantic relatedness. Illustrated the potential of JSD as a proximity measure to validate bibliographic coupling by obtaining 0.90 correlation with manually annotated datasets for research papers clustering and thus it is a reliable proxy for human perceived relatedness. Based on this evidence, JSD is employed as part of soft ground truth in this research to evaluate the hybrid citation-based similarity model [54].

4.6.1 Role as Ground Truth

The JSD-based similarity score is used as soft ground truth to compare with the results of BC, CC, Katz, and Semantic Scholar related papers. By correlating the citation-based structural similarity with semantic similarity based on text, the evaluation guarantees that our hybrid system effectively models not only citation networks but also reflects actual content-based relatedness between research papers.

4.7 Score Normalization

To make heterogeneous information comparable before hybridization, ensure inter-technique comparability and scaled to $[0, 1]$. Let P be the collection of papers and $\in\subseteq P \times P$ the set of candidate pairs. For any pair (i, j) $\in\in$ the raw score are:

4.7.1 Normalized Bibliographic Coupling

Bibliographic Coupling is the similarity of two papers by the extent of overlap reference lists. The more they cite works in common, the closer they are[60].

$$BC_{raw}(i, j) = |R(i) \cap R(j)| \tag{4.12}$$

$R(i)$ is the set of references cited by paper i.

Normalization (max-scaling)

$$BC_{max} = \max\ BC_{raw}(u, v) \tag{4.13}$$

Two papers (nodes) in the citation graph, where each node represents one research paper. The raw Bibliographic Coupling score between paper u and paper v.

$$BC(i, j) = \frac{BC_{raw}(i, j)}{BC_{max}} \tag{4.14}$$

4.7.2 Normalized Co-citation

Co-Citation (CC) is a measurement for the similarity between two papers that depends on how frequently subsequent works will list them together in their respective citation lists.

$$CC_{raw}(i, j) = |C(i) \cap C(j)| \tag{4.15}$$

C(i) is the set of papers that cite paper i.

Normalization (max-scaling)

$$CC_{max} = \max\ CC_{raw}(u, v) \tag{4.16}$$

$$CC(i, j) = \frac{CC_{raw}(i, j)}{CC_{max}} \tag{4.17}$$

4.7.3 Normalized Katz

The Katz measure quantifies indirectly how related two papers are, taking into account any path in a citation-structure graph between them. It would give larger influence to short paths and allow indirect connections beyond the direct citations.

$$Katz_{raw}(i, j) = \sum_{l=1}^{\infty} \beta^l\ (A^l)_{ij} \tag{4.18}$$

where A is the adjacency matrix of the citation graph, $\beta \in (0, 1/\lambda_{max})$ is a damping factor, and L is the max hop (you often use $L = 3$).

Normalization (max-scaling)

$$Katz_{max} = \max\ Katz_{raw}(u, v) \tag{4.19}$$

$$Katz(i, j) = \frac{Katz_{raw}(i, j)}{Katz_{max}} \tag{4.20}$$

4.7.4 Normalized JSD divergence

Jensen–Shannon Divergence (JSD) capture the semantic divergence between two papers by comparing their word frequency-based distributions (here, normalized TF–IDF weights to sum = 1), with smaller JSD indicating higher similarity.

Preprocess & vectorize (your pipeline). Tokenize (unigrams), lowercase, remove stop-words, lemmatize; build TF–IDF; normalize each vector to sum to 1 so it is a probability distribution.

Equation.

$$JSD(P \parallel Q) = \tfrac{1}{2} D_{KL}(P \parallel M) + \tfrac{1}{2} D_{KL}(Q \parallel M) \tag{4.21}$$

$$M = \tfrac{1}{2}(P + Q), \qquad D_{KL}(P \parallel M) = \sum_i p(i) \log \frac{P(i)}{M(i)} \tag{4.22}$$

Convert to similarity, then normalize

$$J_{sim}(i, j) = 1 - JSD(P \parallel Q) \tag{4.23}$$

$$J_{max} = \max \; J_{sim}(u, v) \tag{4.24}$$

$$J(i, j) = \frac{J_{sim}(i, j)}{J_{max}} \tag{4.25}$$

Jensen–Shannon Divergence is a good semantic reference, as the TF–IDF JSD pipeline has been empirically established for paper to validate relatedness, [54] observed ≈ 0.90 correlation between JSD-based clustering and manually annotated dataset, supporting JSD as a reliable soft ground truth for relatedness.

4.8 Comparison

TABLE 4.1: Correlation of Techniques with JSD

Technique	**Correlation with JSD**
Bibliographic Coupling (BC)	0.40
Semantic Scholar (SS)	0.35

The data in Table 4.1 shows that Bibliographic Coupling (BC) is correlated better with the Jensen–Shannon Divergence (JSD) scores than with Semantic Scholar (SS). Concretely, BC reaches 0.40 correlation and SS gets to 0.35 correlation. This indicates that BC can retain semantic similarity of research papers better than SS in this dataset, because BC directly utilizes the shared reference structure in documents that indicates inter-relationships between underlying concepts. This difference may be small, but it reveals the different scoring strength of citation-based bibliographic coupling for defined content-based similarity measures such as JSD and (more generic) states that SS is not ineffective however, there are other influencing factors beyond direct input reference overlap due to its design as a more general similarity metric provided by the Semantic Scholar system.

Chapter 5

Results and Discussion

5.1 Experimental Setup and Data Snapshot

5.1.1 Corpus and Evaluation Input

This chapter shows results for the entire machine learning corpus of ≈ 38,000 anchor papers (the "main papers"). We build an anchor→candidate ranking for each anchor and investigate them at the corpus scale. Two inputs drive this chapter. we have ranked by score per anchor lists of candidates and associated with these band labels used in this study (Highly related/Related/Weak related/Not related). The second correlation summary diagnostic alignment between each method and the JSD ground truth. we processed the sample data ≈ 1225 documents sampled from 6.4M population with 95% confidence and 2.8% margin of error and found the cutoff points. On the basis of these cutoff points we can find the semantic similarity of two papers by using bibliographic coupling score.

5.1.2 Preprocessing, Normalization and Scoring Protocol

All identifiers were standardized (lowercased, trimmed, basic cleaning) to avoid differences due to key mismatches. Citation lists were scanned to create reference

TABLE 5.1: Output Rows

Anchor	**Candidate**	**BC**	**JSD**	**CC**	**Katz**	**SS**
P1	P2	1	0.7070	0.0190	0.6749	0
P3	P4	0.9738	0.7130	0.0066	0.6749	0
P5	P6	0.7843	1	0.0012	0	1
P7	P8	0.6928	0.6426	0.0053	0.0067	0
P9	P10	0.6928	0.6032	0.0020	0	0
P11	P12	0.6767	0.2068	0.0004	0	0
P13	P14	0.6405	0.6492	0	0.6749	1
P15	P16	0.6339	0.7897	0	0	1

sets for individual papers. We calculated BC for each (anchor, candidate) pair as the number of shared references; the pairs with zero overlap remain valid but naturally rank lower. We max-scaled BC scores in the [0, 1] range to ensure comparability across anchors within a context, preserving only their rank while bringing the maximum observed value (from all values) to 1. The production ranking presented in this chapter does not utilize other content-based features and graph-path features by design. The ranking protocol is simple and orders by decreasing BC and, in case of ties, breaks deterministically (using a stable sort on, for example, the candidate ID) to obtain reproducible results. The resulting ordered lists are written to the results CSV alongside their assigned band labels. The findings demonstrate the real empirical behavior at scale. We focus on aggregate coverage, how movies are distributed across bands, and top-K ranking quality rather than fitting a model.

5.2 Correlation Diagnostics and Model Selection

Prior to full-corpus ranking, for each candidate information, we measured the degree of alignment with a content-based proxy for relatedness. In a development sample of the 6.4M paper population ($\approx$ 1225 anchor–candidate pairs; 95% confidence, 2.8% margin of error), we compared max-scale scores from BC, CC, and Katz to an inverse and max-scale Jensen–Shannon similarity (JSDNormalized,

larger objects are more semantically similar). This maps all axes to [0,1] so that "higher = more related" is the same for both plots and statistics.

Summarizes the diagnostics as four side-by-side scatter plots with least-squares trend lines: Figure 5.1 (BC vs JSD Normalized), a clear increasing trend is observed; as BC increases, JSD Normalized tends to increase.

The points fall all over the place, but add up to a pronounced positive slope: there is some moderate tendency for shared-reference strength and semantic similarity to align positively. This is the most promising pattern among the three graph-based information.

Figure 5.2 (CC vs JSD Normalized) exhibits a dense column of values around the zero on the CC axis,of a and then it follows a bit downturn trend line. In this model, co-citation is both sparse and noisy, where $CC \neq 0$, it is not systematically associated with greater semantic similarity. The negative, weak trend suggests refraining at the moment from using CC as a pure scoring information in this corpus.

Figure 5.5 (Katz vs JSD Normalized) displays two narrow vertical clusters (near 0 and near the upper end) with a nearly horizontal fit. This represents virtually no usable gradient with respect to semantic similarity due either to path-based saturation or the manner in which connectivity condenses in a few bands, rendering Katz non-diagnostic for our purpose here.

Figure 5.6 (Binary ground truth vs JSDNormalized) shows close to zero linear association between the binary external label (e.g., query listed as related on Semantic Scholar) and continuous semantic similarity. This is not surprising: a binary label collapses a wide range of the semantics closeness into two buckets. The plot validates our use of JSD for diagnostics proxy. In aggregate, the diagnostics support a simple, defensible model-selection decision: retain BC, discard CC and Katz from the production scorer. BC is the only in our information in development sample that consistently moves in accordance to semantic similarity, while CC is negatively trending and Katz has no discernible relationship at all.

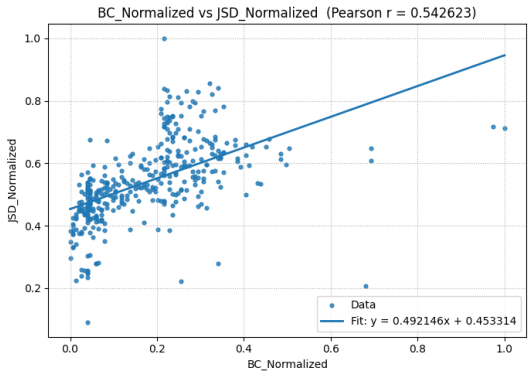

FIGURE 5.1: BC vs JSD Correlation

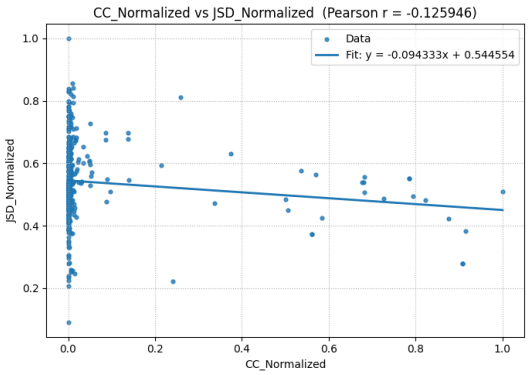

FIGURE 5.2: CC vs JSD Correlation

In Figure 5.1, the correlation between BC and JSD y-axis shows the JSD values, and the x-axis shows the Bc values.

In Figure 5.2, the correlation between CC and JSD y-axis shows the JSD values, and the x-axis shows the CC values.

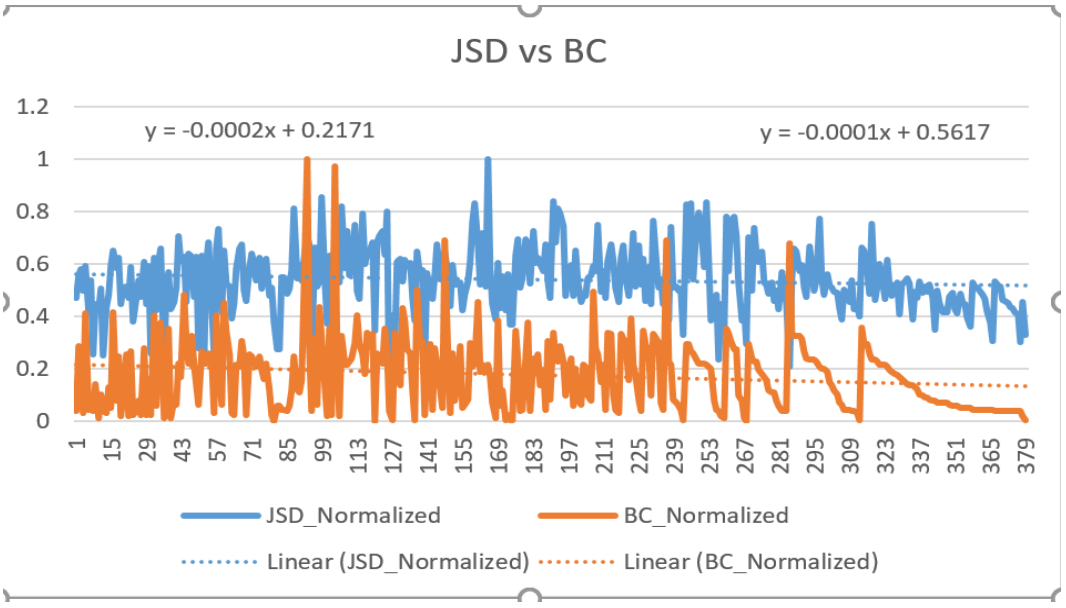

FIGURE 5.3: Trend Line: JSD vs BC

In 5.3 show trend line equation for BC = -0.0002x + 0.2171 and equation for JSD = -0.0001x + 0.5617.

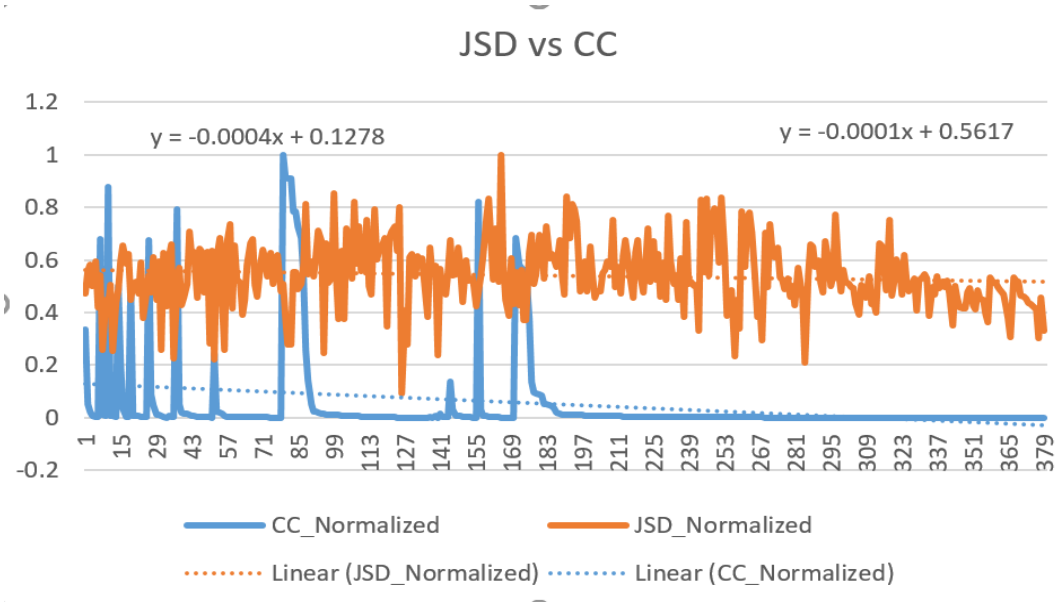

FIGURE 5.4: Trend Line: JSD vs CC

In 5.4 show trend line equation for CC = -0.0004x + 0.1278 and equation for JSD = -0.0001x + 0.5617.

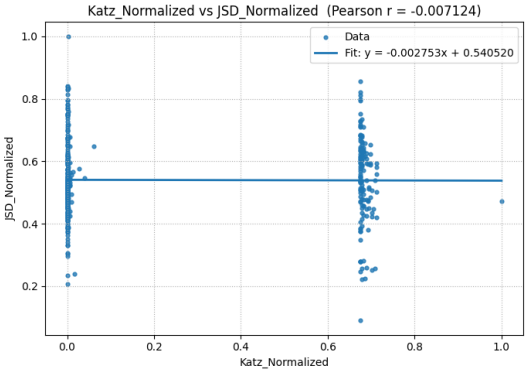

FIGURE 5.5: Katz vs JSD Correlation

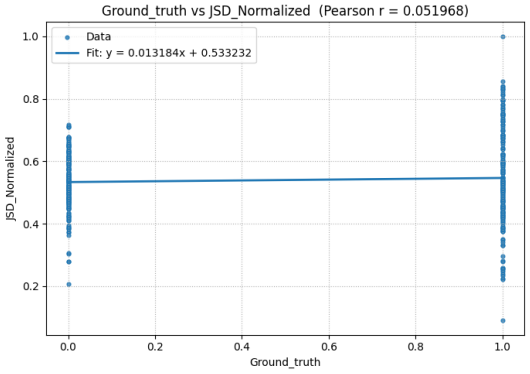

FIGURE 5.6: SS Related vs JSD Correlation

In Figure 5.5, the correlation between Katz and JSD y-axis shows the JSD values, and the x-axis shows the Katz values.

In Figure 5.6, the correlation between SS and JSD y-axis shows the JSD values, and the x-axis shows the SS values.

Correlation scatterplots against semantic similarity (JSDNormalized). 5.1 BC vs JSD: clear positive trend, indicating moderate semantic alignment. 5.2 CC vs JSD:

values cluster near zero with a slight negative trend. 5.5 Katz vs JSD: near-zero association with clustered scores. 5.6 Binary ground truth vs JSD: weak linear relation, supporting JSD for diagnostics and retrieval metrics for final selection. We use our BC-only over the entire corpus to provide fixed rank-bands for each anchor's ranked list, and then report coverage, distribution, and top-K ranking quality. We discuss errors and consequences, such as cases where content similarity diverges from citation overlap.

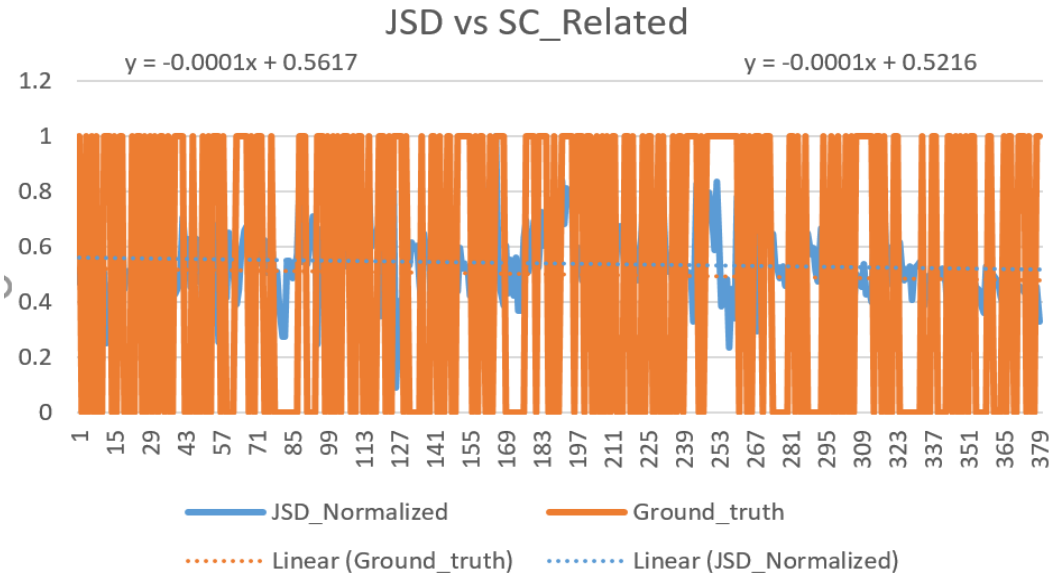

FIGURE 5.7: Trend Line: JSD vs SS

In Figure 5.7 show trend line equation for SS = -0.0001x + 0.5216 and equation for JSD = -0.0001x + 0.5617.

In Figure 5.8 show trend line equation for Katz = -0.0025x + 0.7235 and equation for JSD = -0.0001x + 0.5617.

TABLE 5.2: Correlation between techniques and JSD

Technique	**Correlation with JSD**
Bibliographic Coupling vs JSD	0.40
Co-Citation vs JSD	-0.11
Katz vs JSD	-0.01
Semantic Scholar vs JSD	0.35

The Table 5.2 presents the Pearson inspiratory JSD correlation for each approach. Among all bibliometric methods, Bibliographic Coupling (BC) demonstrated the

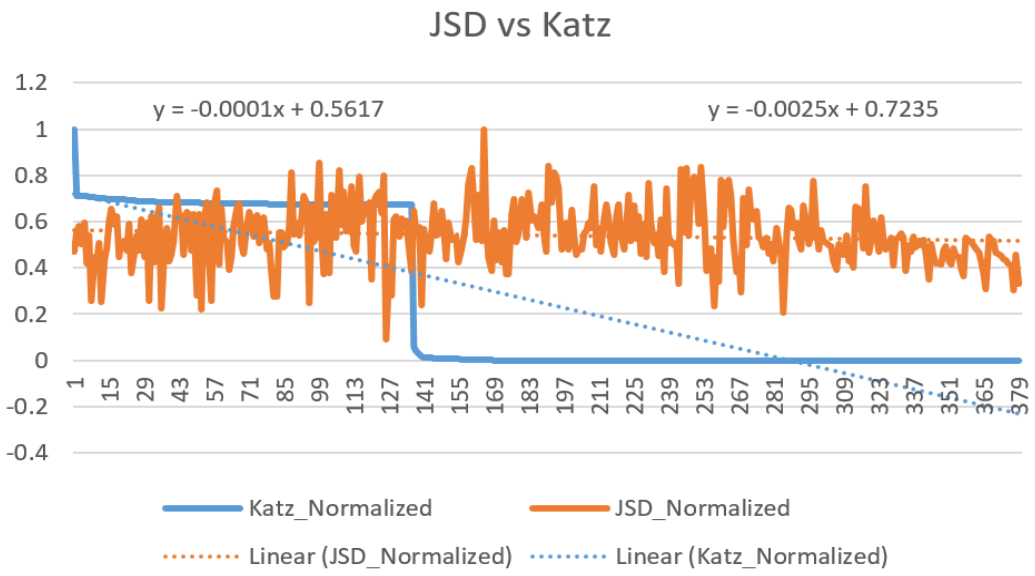

FIGURE 5.8: Trend Line: JSD vs Katz

strongest alignment with JSD (r = 0.40), surpassing Semantic Scholar (r = 0.35) in capturing genuine research paper relatedness. In contrast, CC vs JSD = −0.11 (slightly negative) and Katz vs JSD = −0.01, thus demonstrating they don't co-vary with JSD in a meaningful manner. From these results, this highlights BC as the most effective bibliometric method and semantically consistent method for identifying meaningful paper connections within the dataset; therefore, this method would be considered as top-ranked for finding articles in this work, and both CC and Katz will become safe choices.

5.3 Summary of Findings

Cross-correlation with a JSD-based semantic proxy indicated a consistent trend: BC-lined-up positively (higher preliminary scores were better), largely out and away the strongest, while CC and Katz were negative or near zero, residue discarded from production scoring. This diagnostic selection was supported by the behavior of the corpus: on BC highly associated items clustered.

From a practical perspective, the results suggest that BC-only is successful as a powerful default for machine-learning paper recommendation with banding from

the model to aid interpretability by users. The main limitations are that BC are biased toward well-established subfields and the coarseness of fixed bands for very dense or very sparse neighborhoods.

The results suggest actionable next steps: dynamic banding tuned to anchor context, constrained re-introduction of CC/Katz where diagnostics improve (e.g., recent-citation windows or topical subgraphs), and lightweight semantic reranking of the BC top-N to recover content-close items that citation overlap alone might miss; overall, BC provides a stout backbone, for which targeted augmentations can address the few, well-understood failure modes revealed here.

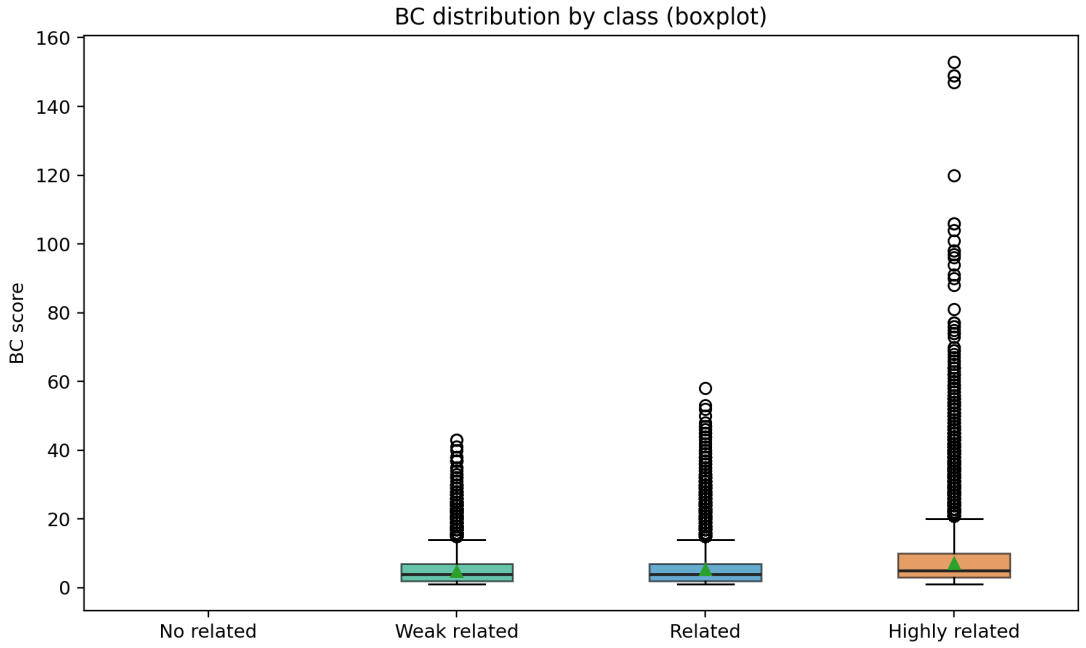

FIGURE 5.9: BC distributation

In Figure 5.9 shows how BC scores distribute within the four rank bands assigned by your Top-K policy. Each box summarizes a class. The pattern is monotonic: "Weak related" has the lowest central BC values (mostly small single-digits), "Related" shifts upward with a higher median and mean, and "Highly related" is highest with a broad upper tail including many large outliers reflecting cases where anchor candidate pairs share many references. The "No related" band has very few points in this figure.

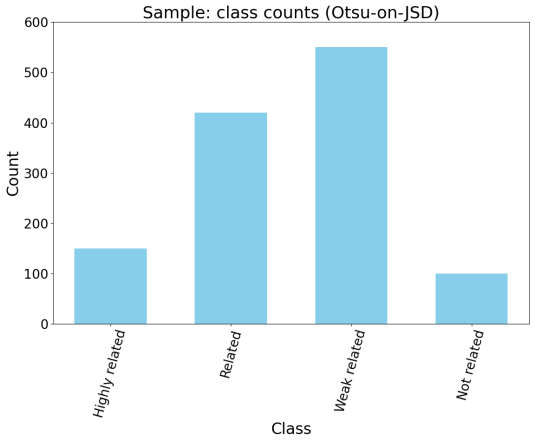

FIGURE 5.10: Count per Class

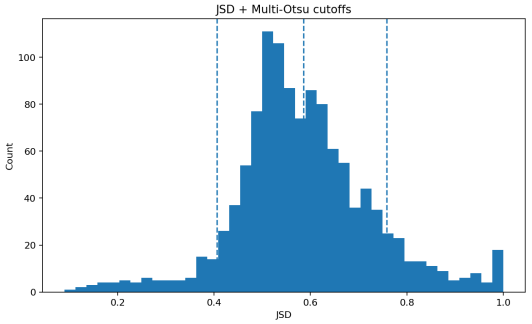

FIGURE 5.11: BC histogram by class

In Figure 5.10 show the classes count and In Figure 5.11 shows the Cutoff points into four classes: highly related, related, weakly related, and not related by using the multi-otsu cutoff.

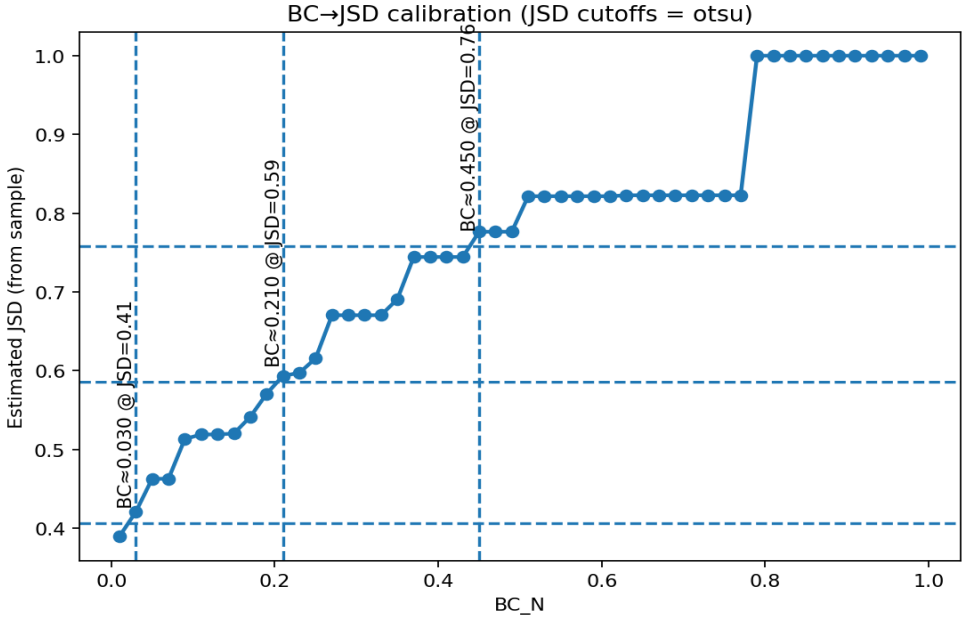

FIGURE 5.12: Mean BC by Rank Position

In Figure 5.12 find the exact cutoff points Bc vs JSD.

Chapter 6

Conclusion and Future Work

6.1 Introduction

This research examined citation-based recommendations of machine learning literature at a corpus scale, considering that structural information from the citation network complements content similarity. We started with three popular methods: Bibliographic Coupling (BC), Co-Citation (CC), and Katz similarity as basis techniques, plus a semantic proxy based on Jensen–Shannon Divergence(JSD) applied over TF–IDF probability distributions. Having rescaled all scores into a single range, we first analysed alignment against semantics on our development sample and then performed an evaluation over the full corpus of 6.64 million papers. Outputs to the end user were intended to be descriptive: we assigned each candidate generated by a rule-based top k of the Top-10 produced for each anchor to one of four categories (K1, high relatedness; K2, relatedness; K3, weak related; Not related) [61]. The study focused on replicability and transparency. Preprocessing, normalization, band definitions, and evaluation artifacts were all kept alike in all experiments to allow results tracking and auditing. Diagnostic images identified JSD as a quality assessor rather than a production scorer, and consequently, the resulting list preferred a clear, refreshingly obvious ranking information. Across, the thesis has tried to match methodological stringency with pragmatic expediencies,

seeking a system that is both theoretically premised and operationally applicable [62].

6.2 Research Question and Justification

RQ1: How can we build a reliable dataset of machine-learning research papers?

This utility aims to scrape a full set of open research articles indexed by Semantic Scholar in a systematic manner. It aims at collecting two types of academic papers, including one primary (denoted as "main papers") as well as their related papers, which can augment the depth and width of the dataset. The academic database Semantic Scholar has extensive metadata and relations between research papers that are perfect for such a project. It is expected to deliver a well-formatted and structured data set that can be applied for research analysis, trend recognition, citation statistics, and knowledge graph creation in the domain of machine learning literature.

(a) API fetch module: for getting citation and reference metadata from the Semantic Scholar API periodically.

(b) Manual Scraper Module: Using Selenium and Beautiful Soup to scrape the references section from paper webpages when the API limit is reached.

(c) Deduplication Engine: Used to identify and remove duplicate references to ensure a clutter-free dataset.

(d) Data Linking and Storage: Citation and reference data to the main papers are linked using Pandas and stored as structured data for subsequent analysis.

RQ2: Among BC, CC, Katz, and Semantic Scholar, which technique aligns most strongly with JSD?

The Table 6.1 presents the Pearson inspiratory JSD correlation for each approach. Among all bibliometric methods, Bibliographic Coupling (BC) demonstrated the

TABLE 6.1: Correlation between techniques and JSD

Technique	**Correlation with JSD**
Bibliographic Coupling vs JSD	0.40
Co-Citation vs JSD	-0.11
Katz vs JSD	-0.01
Semantic Scholar vs JSD	0.35

strongest alignment with JSD (r =0.40), surpassing Semantic Scholar (r = 0.35) in capturing genuine research paper relatedness. In contrast, CC vs JSD = -0.11 (slightly negative) and Katz vs JSD = -0.01 (essentially zero), thus demonstrating they don't co-vary with JSD in a meaningful manner. From these results, this highlights BC as the most effective bibliometric method and semantically consistent method for identifying meaningful paper connections within the dataset; therefore, this method would be considered as top-ranked for finding articles in this work, and both CC and Katz will become safe choices

RQ3: How can we group related papers into three categories: Highly related, Related, and Weakly related?

In 5.12 using the Multi-otsou cutoff points, Bibliographic coupling, and JSD into four classes: highly related, related, weakly related, and not related. Using these cutoff points, assign classes.

6.3 Future Work

Fusing several techniques may significantly improve the reliability of similarity estimation. By integrating different approaches such as BC, CC, Katz and SS the system benefits from complementary strengths that reduce reliance on any single method. This multi-technique fusion not only increases the stability of results, but also helps to consider varied aspects of relatedness between research papers. As an added benefit, growing the number of individuals (paper pairs or samples) under consideration will help eliminate segments that are poorly measured statistically and provide greater confidence in the correlations observed. First, more valid

models can be detected with a bigger dataset so that the contamination of outliers is decreased and the generalization ability becomes improved in the combined model.

Bibliography

[1] Z. Zhao, Z. Liu, and X. Kong, "A systematic review of citation recommendation over the past two decades," *arXiv preprint arXiv:2302.13472*, 2023.

[2] K. Sugiyama, K. Hatano, and M. Yoshikawa, "Scholarly paper recommendation based on citation information," in *Proceedings of the 10th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2010, pp. 29–38.

[3] M. E. J. Newman, "A measure of betweenness centrality based on random walks," *Social networks*, vol. 27, no. 1, pp. 39–54, 2005.

[4] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the American Society for Information Science*, vol. 24, no. 4, pp. 265–269, 1973.

[5] A. Shahid and M. T. Afzal, "Section-wise indexing and retrieval of research articles," *Cluster Computing*, vol. 21, pp. 481–492, 2018.

[6] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[7] M. M. Kessler, "Bibliographic coupling between scientific papers," *American Documentation*, vol. 14, no. 1, pp. 10–25, 1963.

[8] H. Small, "Visualizing science by citation mapping," *Journal of the American Society for Information Science*, vol. 50, no. 9, pp. 799–813, 1999.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[10] E. Han and G. Karypis, "Feature-based recommendation system," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM)*, Bremen, Germany, Oct. 2005, pp. 446–452.

[11] J. P. Larsen, W. Leong, and A. R. Padgett, "A machine learning approach to bibliometric analysis for research paper classification and trend detection," *Journal of Informetrics*, vol. 13, no. 4, pp. 977–991, 2019.

[12] H. D. White and B. C. Griffith, "Author co-citation: A literature measure of intellectual structure," *Journal of the American Society for Information Science*, vol. 32, pp. 163–171, 1998.

[13] M. M. Kessler, "Bibliographic coupling between scientific papers," *American Documentation*, vol. 14, no. 1, pp. 10–25, 1963.

[14] K. W. Boyack and R. Klavans, "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*, vol. 61, no. 12, pp. 2389–2404, 2010.

[15] B. H. Hall, A. N. Link, and J. T. Scott, "Universities as research partners," *The Review of Economics and Statistics*, vol. 85, no. 2, pp. 485–491, 2003.

[16] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, O. Etzioni *et al.*, "Construction of the literature graph in semantic scholar," *arXiv preprint arXiv:1805.02262*, 2018.

[17] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[18] C. Jeong, S. Jang, E. Park, and S. Choi, "A context-aware citation recommendation model with bert and graph convolutional networks," *Scientometrics*, vol. 124, no. 3, pp. 1907–1922, 2020.

[19] A. Razdaibiedina and A. Brechalov, "Miread: simple method for learning high-quality representations from scientific documents," *arXiv preprint arXiv:2305.04177*, 2023.

[20] 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, pp. 133 500–133 509, 2021.

[21] Y. Zhang, S. Garg, Y. Meng, X. Chen, and J. Han, "Motifclass: Weakly supervised text classification with higher-order metadata information," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. ACM, 2022, pp. 1357–1367.

[22] R. Kumar, *Research Methodology: A Step-by-Step Guide for Beginners*, 5th ed. Thousand Oaks, CA: SAGE Publications, 2018.

[23] C. H. Lee, S. R. Ahmad, D. He, and K. Collins-Thompson, "A comprehensive survey of research paper recommendation systems," *Information Processing & Management*, vol. 58, no. 6, p. 102732, Nov 2021.

[24] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Paper recommender systems: A literature survey," *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, 2016.

[25] F. Ferrara, N. Pudota, and C. Tasso, "A keyphrase-based paper recommender system," in *Digital Libraries and Archives: 7th Italian Research Conference, IRCDL 2011, Pisa, Italy, January 20-21, 2011. Revised Papers 7*. Berlin, Heidelberg: Springer, 2011, pp. 14–25.

[26] M. Umair, T. Sultana, and Y. K. Lee, "Pre-trained language models for keyphrase prediction: A review," *ICT Express*, vol. 10, no. 4, pp. 871–890, 2024.

[27] L. Ajallouda, F. Z. Fagroud, A. Zellou, and E. H. Benlahmar, "Automatic keyphrases extraction: An overview of deep learning approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 1, pp. 303–313, 2023.

[28] K. Sarkar, M. Nasipuri, and S. Ghose, "A new approach to keyphrase extraction using neural networks," *arXiv preprint arXiv:1004.3274*, 2010.

[29] C. Bhagavatula, S. Feldman, R. Power, and W. Ammar, "Content-based citation recommendation," *arXiv preprint arXiv:1802.08301*, 2018. [Online]. Available: https://arxiv.org/abs/1802.08301

[30] H. Herimanto, K. Samosir, and F. Ginting, "A comparative analysis of content-based filtering and tf-idf approaches for enhancing sports recommendation systems," *Innovation in Research of Informatics (Innovatics)*, vol. 6, no. 2, pp. –, 2024.

[31] S. Nazir, M. Asif, and S. Ahmad, "Important citation identification by exploiting the optimal in-text citation frequency," in *Proceedings of the 2020 International Conference on Engineering and Emerging Technologies (ICEET)*, Islamabad, Pakistan, Feb. 2020, pp. 1–6.

[32] S. Y. Hwang, C. P. Wei, Y. C. Huang, and Y. Tang, "Combining coauthorship network and content for literature recommendation," in *Proceedings of the International Conference*, 2010.

[33] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[34] A. Y. Khan, A. S. Khattak, and M. T. Afzal, "Extending co-citation using sections of research articles," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 26, no. 6, pp. 3345–3355, 2018.

[35] 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: An International Journal*, vol. 32, no. 5, pp. 1742–1775, 2025.

[36] B. Gipp and J. Beel, "Citation proximity analysis (cpa): A new approach for identifying related work based on co-citation analysis," in *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI)*, 2009, pp. 571–575.

[37] R. Habib and M. T. Afzal, "Paper recommendation using citation proximity in bibliographic coupling," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 25, no. 4, pp. 2708–2718, 2017.

[38] A. M. Khan, A. Shahid, M. T. Afzal, F. Nazar, F. S. Alotaibi, and K. H. Alyoubi, "SwICS: Section-wise in-text citation score," *IEEE Access*, vol. 7, pp. 137 090–137 102, 2019.

[39] J. Yun, "Generalization of bibliographic coupling and co-citation using the node split network," *Journal of Informetrics*, vol. 16, no. 2, p. 101291, 2022.

[40] S. Jeong, H. Ko, and J. Seo, "A context-aware citation recommendation model with bert and graph convolutional networks," *PeerJ Computer Science*, vol. 5, p. e236, 2020.

[41] Y. Liang, Q. Li, and T. Qian, "Finding relevant papers based on citation relations," in *Proceedings of the 12th International Conference on Web-Age Information Management (WAIM)*, vol. 12, Wuhan, China, Sep. 2011, pp. 403–414.

[42] K. Haruna, M. A. Ismail, A. B. Bichi, V. Chang, S. Wibawa, and T. Herawan, "A citation-based recommender system for scholarly paper recommendation," in *Computational Science and Its Applications – ICCSA 2018: 18th International Conference, Proceedings, Part I.* Melbourne, VIC, Australia: Springer International Publishing, 2018, pp. 514–525.

[43] J. Shen, M. A. A. Haqqani, B. Hu, C. Huang, X. Xie, T. Lee, and J. Zhang, "Temporal graph neural network-powered paper recommendation on dynamic citation networks," *arXiv preprint arXiv:2408.15371*, 2024.

[44] Y. Ding, X. Liu, C. Guo, and B. Cronin, "The distribution of references across texts: Some implications for citation analysis," *Journal of Informetrics*, vol. 7, no. 3, pp. 583–592, 2013.

[45] A. Kanakia, Z. Shen, D. Eide, and K. Wang, "A scalable hybrid research paper recommender system for microsoft academic," in *Proceedings of the World Wide Web Conference.* ACM, 2019, pp. 2893–2899.

[46] L. Li, Z. Zhang, and S. Zhang, "Hybrid algorithm based on content and collaborative filtering in recommendation system optimization and simulation," *Scientific Programming*, vol. 2021, p. Article ID 7427409, 2021.

[47] 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," *Scientometrics*, vol. 129, no. 3, pp. 1545–1566, 2024.

[48] H. Guo, Z. Shen, J. Zeng, and N. Hong, "Hybrid methods of bibliographic coupling and text similarity measurement for biomedical paper recommendation," in *MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation*, 2022, pp. 287–291.

[49] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha *et al.*, "Construction of the literature graph in semantic scholar," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, New Orleans, LA, USA, 2018, pp. 84–91.

[50] V. Stergiopoulos, M. Vassilakopoulos, E. Tousidou, and A. Corral, "An academic recommender system on large citation data based on clustering, graph modeling and deep learning," *Knowledge and Information Systems*, vol. 66, no. 8, pp. 4463–4496, 2024.

[51] F. Ebrahimi, A. Asemi, A. Nezarat, and A. Ko, "Developing a mathematical model of the co-author recommender system using graph mining techniques and big data applications," *Journal of Big Data*, vol. 8, no. 1, pp. 1–15, 2021.

[52] P. Calado, M. Cristo, E. Moura, N. Ziviani, B. Ribeiro-Neto, and M. A. Gonçalves, "Combining link-based and content-based methods for web document classification," in *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM)*, 2003, pp. 394–401.

[53] T. Kanwal and T. Amjad, "Research paper recommendation system based on multiple features from citation network," *Scientometrics*, vol. 129, pp. 5493–5531, 2024.

[54] I. Ahmed, "BCSw: Weighted Section-Wise Bibliographic Coupling to Find Related Research Papers," PhD thesis, Capital University of Science and Technology, Islamabad, Pakistan, 2025. [Online]. Available: https://cust.edu.pk/phd-repository/

[55] Python Software Foundation, "Python language reference, version 3.11," https://www.python.org, 2023, [Online; accessed 23-Nov-2025].

[56] L. Richardson, "Beautiful soup documentation, version 4.12.2," https://www.crummy.com/software/BeautifulSoup/, 2023, [Online; accessed 23-Nov-2025].

[57] SeleniumHQ, "Selenium webdriver documentation," https://www.selenium.dev/documentation/webdriver/, 2023, [Online; accessed 23-Nov-2025].

[58] Pandas Development Team, "pandas: Python data analysis library, version 2.1.0," https://pandas.pydata.org, 2023, [Online; accessed 23-Nov-2025].

[59] Allen Institute for AI, "Semantic Scholar: A free, AI-powered research tool for scientific literature," 2024, https://www.semanticscholar.org.

[60] K. Järvelin and J. Kekäläinen, "Ir evaluation methods for retrieving highly relevant documents," *Information processing & management*, 2017.

[61] L. Leydesdorff, "On the normalization and visualization of author co-citation data: Salton's cosine versus the jaccard index," *Journal of the American Society for Information Science and Technology*, forthcoming.

[62] M. Wang, J. Ren, S. Li, and G. Chen, "Quantifying a paper's academic impact by distinguishing the unequal intensities and contributions of citations," *IEEE Access*, vol. 7, pp. 96 198–96 214, 2019.