CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

BCSw: Weighted Section-Wise Bibliographic Coupling to Find Related Research Papers

by

Ibrar Ahmed

A dissertation submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Faculty of Computing
Department of Computer Science

2025

BCSw: Weighted Section-Wise Bibliographic Coupling to Find Related Research Papers

By

Ibrar Ahmed

(DCS171001)

Dr. Kashif Naseer Qureshi, Associate Professor

University of Limerick, Limerick, Ireland

(Foreign Evaluator 1)

Dr. Naeem Ramzan, Professor

University of West of Scotland, Paisley, UK

(Foreign Evaluator 2)

Dr. Muhammad Abdul Qadir

(Research Supervisor)

Dr. Abdul Basit Siddiqui

(Head, Department of Computer Science)

Dr. Muhammad Abdul Qadir

(Dean, Faculty of Computing)

DEPARTMENT OF COMPUTER SCIENCE

CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY

ISLAMABAD

2025

To my parents, family and especially to the healthcare workers and doctors working in the difficult time of COVID19 era

CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY
ISLAMABAD
Expressway, Kahuta Road, Zone-V, Islamabad
Phone: +92-51-111-555-666 Fax: +92-51-4486705
Email: info@cust.edu.pk Website: https://www.cust.edu.pk

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the dissertation, entitled "**BCSw: Weighted Section-Wise Bibliographic Coupling to Find Related Research Papers**" was conducted under the supervision of **Dr. Muhammad Abdul Qadir**. No part of this dissertation has been submitted anywhere else for any other degree. This dissertation is submitted to the **Department of Computer Science, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the field of **Computer Science.** The open defence of the dissertation was conducted on **December 30, 2024**.

Student Name: **Ibrar Ahmed** (DCS171001)

The Examination Committee unanimously agrees to award PhD degree in the mentioned field.

Examination Committee:

(a) External Examiner 1: Dr. Ehsan Ullah Munir
Professor
CUI, Wah Cantt Campus

(b) External Examiner 2: Dr. Aman Ullah Yasin
Professor
Bahria University, Islamabad

(c) Internal Examiner: Dr. Nayyer Masood
Professor
CUST, Islamabad

Supervisor Name: Dr. Muhammad Abdul Qadir
Professor
CUST, Islamabad

Name of HoD : Dr. Abdul Basit Siddiqui
Associate Professor
CUST, Islamabad

Name of Dean: Dr. Muhammad Abdul Qadir
Professor
CUST, Islamabad

AUTHOR'S DECLARATION

I, **Ibrar Ahmed (Registration No. DCS171001)**, hereby state that my dissertation titled, '**BCSw: Weighted Section-Wise Bibliographic Coupling to Find Related Research Papers**' 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/ world.

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

(**Ibrar Ahmed**)

Dated: 30 December, 2024

Registration No: DCS171001

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the dissertation titled "**BCSw: Weighted Section-Wise Bibliographic Coupling to Find Related Research Papers**" 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 dissertation 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 dissertation declare that no portion of my dissertation 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 dissertation even after award of PhD Degree, the University reserves the right to withdraw/ revoke my PhD 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 dissertation.

(**Ibrar Ahmed**)

Dated: 30 December, 2024

Registration No: DCS171001

List of Publications

It is certified that following publication(s) have been made out of the research work that has been carried out for this dissertation:-

1. **I. Ahmed** and M. T. Afzal, "A systematic approach to map the research articles' sections to IMRaD," *IEEE Access*, vol. 8, pp. 129 359–129 371, 2020.

Ibrar Ahmed

Registration No: DCS171001

Acknowledgement

First and foremost, I express my gratitude to Almighty Allah for granting me the health, wisdom, and strength to embark on this PhD research journey and successfully complete it. I wholeheartedly dedicate this accomplishment to my esteemed supervisors, Dr Muhammad Tanvir Afzal and Dr Abdul Qadir, without their guidance and support, this achievement would have been impossible. I consider myself extremely fortunate to have had the opportunity to work under the supervision of both these renowned academics, who possess extensive experience and knowledge in this field, and whose invaluable guidance has been instrumental in my success.

I would also like to express my sincere appreciation to my family for their unwavering support throughout my doctoral studies. This accomplishment is a testament to the prayers and blessings of my late parents, who have been a source of inspiration for me throughout my life. Additionally, I am grateful to all of my employer, for their generous financial and moral support, which has helped me to focus on my research work with ease.

Ibrar Ahmed

Abstract

Identifying related academic papers is crucial in advancing scientific research, allowing scholars to build upon existing knowledge and explore new frontiers. This endeavor has sparked a significant research community interest in exploring different techniques. These techniques can be divided into two major classes: content-based and Metadata-based (bibliographic coupling and co-citation analysis). Each technique offers unique advantages in discovering related literature concerning a query paper. Co-citation identifies related papers by analyzing how often two papers are cited together in other works. Bibliographic coupling discovers related papers by examining shared references between them. This research focuses on bibliographic coupling. Recent results showed a significant improvement in the discovery of related papers by employing weighted section-wise coupling. However, three key challenges must be addressed to use this technique effectively.

The first challenge involves the limited precision in mapping actual section headings to the logical section as per IMRaD (Introduction, Methodology, Results, and Discussion) structure. This research has addressed this problem by combining already used (existing) features and with some newly discovered features, such as counts of figures and tables, varying in-text citation counts, and the detection of sub-headings wrongly matched with IMRaD. The evaluation of the proposed approach and comparisons with state-of-the-art approaches revealed an improvement of 18.96%, 21.77%, and 9.50% in average precision the state of the art techniques by **Ding** [1], **Shahid** [2], and **Habib** [3] respectively.

The second challenge pertains to the arbitrary assignment of weights to sections for weighted section-wise coupling . To address this, a model was trained using Artificial Neural Networks (ANN) with backpropagation to dynamically allocate section weights on a reasonably large dataset of 5,000 papers. Experiments demonstrated an improved correlation with JSD rankings at 0.86, surpassing the previous score of 0.75.

In conclusion, by addressing key challenges, the research successfully develops and validates innovative methods that significantly improve the findings of related

papers. This work contributes to the field by offering an accurate and efficient means of discovering related research and setting a precedent for future studies aiming to refine the exploration of related literature in digital libraries and scholarly search engines.

Contents

Author's Declaration v

Plagiarism Undertaking vi

List of Publications vii

Acknowledgements viii

Abstract ix

List of Figures xv

List of Tables xvi

Abbreviations xvii

Symbols xix

1 Introduction 1

1.1 Overview . . . 1
1.2 Introduction . . . 1
1.3 Background . . . 3
1.3.1 IMRaD Mapping of Research Article Sections to Logical Sections . . . 5
1.3.2 Techniques for Identifying Related Papers . . . 6
1.3.2.1 Content-Based . . . 6
1.3.2.2 Co-Citation . . . 9
1.3.2.3 Bibliographic Coupling . . . 13
1.3.3 Co-Citation vs. Bibliographic Coupling . . . 15
1.4 Section-Based Bibliographic Coupling . . . 16
1.4.1 Weighted Section-Based Bibliographic Coupling . . . 16
1.4.2 Dynamically Tuning of Section's Weights . . . 16
1.4.3 Significance of research paper sections . . . 17
1.5 Research Problem . . . 17
1.6 Problem Statement . . . 17

1.7 Research Questions and Objectives . . . 18
1.7.1 Research Questions . . . 18
1.7.2 Research Objectives . . . 18
1.8 Research Scope . . . 18
1.9 Research Methodology . . . 19
1.10 Dissertation Organization . . . 20

2 Literature Review **22**
2.1 Overview . . . 22
2.2 Inclusion and Exclusion Criteria . . . 23
2.3 Techniques for Logical Section Extraction in Scientific Research Papers 24
2.3.1 Dataset for Extraction of Sections and Mapping to IMRaD . 25
2.4 Literature Survey on Techniques for Finding Related Papers . . . 26
2.5 Techniques to Find Related Research Papers . . . 26
2.5.1 Content-based Strategies . . . 27
2.5.2 Metrics for Evaluating Research Paper Similarity . . . 33
2.5.3 Techniques Based on Metadata . . . 35
2.6 Strengths and Limitations of Content-Based and Metadata-Based Approches . . . 48
2.6.1 Advantages of Content-Based Approaches . . . 48
2.6.2 Issues with Content-Based Approaches . . . 49
2.6.3 Advantages of Metadata-Based Approaches . . . 50
2.6.4 Issues with Metadata-Based Approaches . . . 51
2.6.5 Collaborative Filtering Approaches . . . 51
2.6.6 Graph-Based Approaches . . . 52
2.6.7 Data Mining-Based Approaches . . . 52
2.6.8 Citation-Based Approaches . . . 53
2.6.9 Dataset for Related Papers . . . 57
2.7 Discussion . . . 58
2.8 Summary . . . 60

3 Extraction of Section-Headings and Mapping to IMRaD Structure **62**
3.1 Overview . . . 62
3.2 Introduction . . . 63
3.2.1 Features for Section Extraction and IMRaD Mapping . . . 64
3.3 System Architecture . . . 68
3.3.1 Schema Generation Engine (SGE) . . . 71
3.4 Data Extraction Engine (DEE) . . . 76
3.5 Data Mapping Engine (DME) . . . 76
3.5.1 Subheadings Mappings . . . 77
3.5.2 Sections Sequences . . . 78
3.5.3 Sections Known Names . . . 79
3.5.4 Citation Count as a Metric for Section Identification . . . 84
3.5.5 Object For Section Identification . . . 84

3.5.5.1 Figures as a Metric for Section Identification 84
3.5.5.2 Algorithms as a Metric for Section Identification . 85
3.5.5.3 Graphs as a Metric for Section Identification 85
3.5.5.4 Tables as a Metric for Section Identification 85
3.6 IMRaD Section Mapping 86
3.7 Execution and Processing Time 88
3.8 Results 90
3.8.1 Precision and Recall Calculations for IMRaD Sections 91
3.9 Evaluation 94
3.9.1 Comparative Analysis 94
3.10 Discussion 96

4 Section's Ranking and Weights Adjustment to Discover Bibliographically Coupled Papers 102
4.1 Overview 102
4.2 Introduction 103
4.3 Methodology 104
4.3.1 Data Collection 104
4.3.2 Dataset Validation Methodology 105
4.3.2.1 Step-by-Step Validation Process 105
4.3.2.2 Preprocessing 106
4.3.2.3 Vectorization 107
4.3.2.4 Jensen-Shannon Divergence Calculation 108
4.3.2.5 Algorithm in Pseudocode for JSD 108
4.3.3 Results and Discussion of Validation of Dataset-2 111
4.4 System Architecture to Find Dynamic Weights of Sections 113
4.4.1 DNN: Deep Neural Network with Backpropagation 113
4.5 Results and Evaluations 119
4.6 Complete Example: JSD and Spearman's Correlation 120
4.6.1 Step 1: Data Preparation 120
4.6.2 Step 2: Weighted Common References 121
4.6.3 Step 3: Probabilities 121
4.6.4 Step 4: Jensen-Shannon Divergence (JSD) 122
4.6.5 Step 5: Spearman's Rank Correlation 123
4.7 Conclusion 125
4.7.1 Limitations and Challenges 127
4.8 Summary 129
4.8.1 Applications of the Proposed Method 130

5 Overall Conclusion and Future Work 132
5.1 Overview 132
5.2 Conclusion 132
5.2.1 RQ1: Section Mapping to IMRaD 133
5.2.1.1 Results Comparison 134

5.2.1.2 Issues in PDF-to-XML Conversion and Annotation 134
5.2.2 RQ2: Optimizing Section Weights for Bibliographic Coupling 136
5.3 Discussion 137
5.4 Future Work 138
5.5 Summary 139

Bibliography **141**

List of Figures

1.1 Bibliographic Coupling and Co-Citation [26] 10
1.2 The Design Science Research Methodology Process for this work . . 20

3.1 Subheading example in the form of PDF file. 66
3.2 Subheading Example-2 66
3.3 Subheading Example-1 67
3.4 Figures in research publications. 67
3.5 Tables in research publications. 68
3.6 Example of a figure caption. 69
3.7 System Architecture 70
3.8 Table Contain the Paper Information 71
3.9 Table Contains the Sections 72
3.10 Table Contains Section Mapping 72
3.11 Table Contain Section Mapping Relations 73
3.12 Table Contains the Figures Information 73
3.13 Table Contains the Table Information 73
3.14 Table Contain the Citaion Information 74
3.15 Table Contains the Algorithms Information 74
3.16 Table Contains the Graphs Informations 75
3.17 ERD of Database Schema 75
3.18 Section-Wise comparison of Precision 95
3.19 Section-Wise comparison of Recall 96
3.20 Section-Wise comparison of F-Measure 97
3.21 Comparison of Precision of combined sections 97
3.22 Comparison of Recall of combined sections 98
3.23 Comparison of F-Measure of combined sections 98

4.1 ANN with Backpropagation for Section Weight Tuning 115
4.2 IMRaD Sections Weights (Dataset-1) 124
4.3 IMRaD Sections Weights (Dataset-2) 124
4.4 Results: Comparison of correlation for dataset-1 126
4.5 Results: Comparison of correlation for dataset-2 127

5.1 Influence of Citation Context on Bibliographic Coupling 138

List of Tables

2.1 Strengths and Limitations of Scientific Paper Recommendation Techniques 55

3.1 Processing Time Comparison of Techniques 89
3.2 Comparison of Precision of Combined Sections 90
3.3 Comparison of Recall of Combined Sections 90
3.4 Comparison of F-Measure of Combined Sections 91

4.1 IMRaD Sections Normalized Weights 121
4.2 The section's Ranking and Weights - (Dataset-1) 123
4.3 The section's Ranking and Weights - (Dataset-2) 123
4.4 Cross Sections Weights Paper A and Paper B - (Dataset-1) 123
4.5 Cross Sections Weights Paper A and Paper B - (Dataset-2) 125
4.6 Correlation (%) for Dataset-1 125
4.7 Correlation (%) for Dataset-2 125

5.1 Comparison of Methodologies for Section Mapping Accuracy 134
5.2 Correlation Scores for Different Recommendation Approaches (Dataset-1) 136
5.3 Correlation Scores for Different Recommendation Approaches (Dataset-2) 136

Abbreviations

AI: Artificial Intelligence

ANN: Artificial Neural Networks

API: Application Programming Interface

AWS: Amazon Web Services

BC: Bibliographic Coupling

CAD: Citation Anchor Detection

CBF: Content-Based Filtering

CC: Co-Citation

CF: Collaborative Filtering

CI: Citation Influence

CNN: Convolutional Neural Network

DB: Database

DCS: Distributed Consensus System

DME: Data Mapping Engine

DEE: Data Extraction Engine

ERD: Entity Relationship Diagram

ETCD: Highly Available Key-Value Store for Distributed Systems

F1: F-Measure

GPU: Graphics Processing Unit

HPC: High-Performance Computing

HTML: HyperText Markup Language

HTTP: HyperText Transfer Protocol

IMRaD: Introduction, Methodology, Results, and Discussion

IoT: Internet of Things

ITC: In-Text Citation

JSD: Jensen-Shannon Divergence

KPI: Key Performance Indicator

LSA: Latent Semantic Analysis

MLPNN: Multilayer Perceptron Neural Network

MVE: Mapping View Engine

NLP: Natural Language Processing

PDF: Portable Document Format

RBM: Restricted Boltzmann Machine

RNN: Recurrent Neural Network

RQ: Research Question

SGE: Schema Generation Engine

SQL: Structured Query Language

SwICS: Section-wise In-Text Citation Score

TF-IDF: Term Frequency-Inverse Document Frequency

XML: eXtensible Markup Language

xPath: XML Path Language

xQuery: XML Query Language

Symbols

$\mathbf{W}^{(i)}, \mathbf{b}^{(i)}$	Weights and biases for layer i
$\mathbf{Z}^{(i)}$	Activation for layer i
σ	Activation function
$\hat{y}$	Output representing paper relatedness
$\mathbf{X}$	Input representing tuned section weights
$\odot$	Represents element-wise multiplication
α	Learning rate for the update
$W^{(i)}_{\text{tuned}}$	Represent the tuned section weights
S_w	Section-wise bibliographic coupling strength
II	Introduction-Introduction coupling weight
MM	Methodology-Methodology coupling weight
RR	Results-Results coupling weight
DD	Discussion-Discussion coupling weight
IM, IR, ID, MR, MD, RD	Cross-section coupling weights between different sections
TF	Term Frequency
IDF	Inverse Document Frequency
JSD	Jensen-Shannon Divergence
M	Average probability distribution
$KLD(P\|\|M)$	Kullback-Leibler divergence of P relative to M
P, Q	Probability distributions for two documents
P_A, P_B	Normalized TF-IDF vectors for Papers A and B
$M(i)$	Intermediate average distribution for JSD calculation
$iter$	Current iteration in training

max_iter	Maximum allowed iterations for optimization
err	Error in predicted relatedness

Chapter 1

Introduction

1.1 Overview

This chapter introduces all the concepts in the research domain, including finding the related research papers. Highlights different techniques, which include content-based and metadata-based, to find related research papers with the scope and characteristics of these techniques. Then, it introduces the research problem, questions, Methodology adopted to answer the questions, and an overview of the thesis.

1.2 Introduction

The ability to discover related papers is crucial for advancing science and knowledge. It enables researchers to build upon existing work and contribute to the scientific community. Over the past two decades, this challenge has garnered immense attention from the research community, resulting in many proposed approaches and techniques to address the issue of finding related papers. Researchers continuously strive to improve the precision and recall of schemes to discover related research papers by leveraging metadata and the contents available in digital libraries. This research focuses on analyzing the problems in existing schemes and formulating a

scheme to improve the precision and recall of finding related papers. Many parameters have been explored to identify related papers using content and metadata approaches. Contented-based and meta-data-based are two common approaches used to find related research papers. In the content-based approach, text is the main ingredient in finding the relatedness, while in meta-data, bibliographic information, particularly citations, serves as a valuable indicator of related research papers. Two common approaches, namely co-citations and bibliographic coupling, are frequently employed to address this challenge. Co-citations refer to papers cited by the same citing paper, while bibliographic coupling identifies papers cited by multiple papers. Bibliographic coupling holds promise for discovering related papers and offers potential for improvement. When two papers cite the same paper(s), the citing papers are considered bibliographically coupled. The strength of bibliographic coupling depends on the number of common citations between papers. Researchers have observed that an increase in the strength of bibliographic coupling correlates with higher relevancy of the cited papers. However, it is essential to recognize that contextual information is an important parameter in bibliographic coupling and co-citation.

The research proposes section-wise bibliographic coupling by dynamically tunning section weights to find relevancy using bibliographic coupling. By considering the citations within specific sections of papers, the research aims to filter out related papers from a pool of citations. Certain sections, such as results and analysis, are likely to contain pertinent information than others, like the background section. The possibility of section-wise bibliographic coupling by dynamically assigning weights to different sections based on their significance in finding related research papers is explored to emphasize the importance of different sections in determining relevancy. The central focus of the research revolves around section-wise bibliographic coupling by assigning weights to sections and its potential to improve accuracy in identifying related papers. One key objective is to accurately recognize sections within research papers and align them with a predefined logical structure, such as IMRaD (Introduction, Methodology, Results, and Discussion). Although existing systems exist for mapping to IMRaD headings, they suffer from suboptimal precision and

recall. To address the limitations of previous research, Ding et al. [1] proposed a dictionary-based method for section identification, relying on predefined terms to map sections to the IMRaD structure. Although their approach was practical in some instances, it lacked flexibility due to its static nature, which made it challenging to adapt to varying paper formats. Building upon this, Shahid et al. [2] introduced a template and dictionary-based method that offered some improvements but struggled to distinguish between main sections and subheadings, often treating subsections as independent entities. This limitation led to lower precision and recall in mapping sections accurately. In contrast, Habib and Afzal [3] focused on a citation-based mapping technique, utilizing in-text citation counts to enhance section identification. While this method showed promise, its reliance solely on citation density proved inadequate, especially in papers where the frequency of citations did not correspond to the significance of the sections.

This research aims to enhance the accuracy of identifying related papers by first identifying sections and mapping them to the IMRaD structure, followed by fine-tuning the section weights. The initial focus is on accurately classifying sections and aligning them with the logical IMRaD framework. The ultimate objective is to improve bibliographic coupling by dynamically adjusting section weights to better identify the relationships between research papers.

1.3 Background

The challenge of identifying related research papers has evolved significantly over time, driven by the rapid proliferation of academic publications [4]. Early approaches, such as bibliographic coupling and co-citation analysis, were pioneered in the 1960s. Garfield [5] introduced the concept of co-citation, emphasizing the potential of co-citation to reveal connections between scholarly works. Building on this idea, Kessler [6] formalized bibliographic coupling, establishing a method to link documents based on shared references. These foundational techniques played a pivotal role in the development of citation analysis, which remained the dominant approach for identifying relationships among academic publications for decades.

With the rapid growth of academic publications, finding relevant research papers has become increasingly difficult. Traditional methods, such as keyword searches and counting citations, can no longer identify the most appropriate work in a given field. By the late 1990s and early 2000s, researchers realized that not all citations are equally influential, making simple citation counts insufficient for measuring a paper's relevance [7]. As a result, researchers began developing advanced methods to recommend relevant papers.

Gipp and Beel [8] introduced models that used natural language processing (NLP) to uncover key topics and themes in research articles. This made it possible to recommend papers accurately by understanding the content of the papers, not just the citations.

As the field advanced, researchers focused on improving bibliographic coupling, which links papers sharing common references. Habib and Afzal [3] pointed out that traditional bibliographic coupling does not consider the importance of different sections within a paper. For example, a paper's methodology section might be relevant to a researcher than the results section, but older models did not account for this. Newer models now adjust the importance of sections dynamically, improving the precision of recommendations.

Khan et al. [9] found that analyzing which sections of a paper are cited can reveal deeper connections between papers, making recommendations accurate. Khan et al. [10] proposed a model that uses in-text citation patterns and frequencies across different sections to enhance co-citation analysis. This section-aware approach provides detailed insights into research relationships and trends.

Habib and Afzal [11] also looked at how closely papers are cited together (citation proximity) to improve recommendations. Their work shows that the closer two papers are cited within a document, the likely they are to be related. Expanding on these ideas, Khan et al. [12] introduced the Section-wise In-Text Citation Score (SwICS). This method measures the importance of sections based on how often they are cited, helping to identify relevant papers effectively.

In conclusion, the field of research paper recommendation has come a long way—from basic citation counts to advanced models that use NLP techniques. However, there is still a need for models that can adapt to the importance of different sections within papers. This study aims to fill this gap by introducing a new section-aware bibliographic coupling model. By adjusting section weights based on the query, the proposed model will improve the identification of related research papers, providing precise recommendations.

1.3.1 IMRaD Mapping of Research Article Sections to Logical Sections

The IMRaD structure—comprising **Introduction, Methodology, Results, and Discussion**—has become the standard framework for organizing scientific research articles. This format offers clarity and logical progression, improving the readability and coherence of academic writing [13–15]. The IMRaD model promotes consistency in reporting research findings, helping readers and researchers efficiently follow the narrative and extract meaningful insights. Additionally, mapping diverse article sections to the IMRaD framework is vital in enhancing **bibliographic coupling** and finding related papers.

In their study, **Shahid et al.** [2] identify key challenges in accurately mapping heterogeneous section titles across research articles to the IMRaD structure. This problem arises because journals and researchers often employ varied section headings such as *Background*, *Experimental Setup*, or *Related Work*, which do not directly align with the IMRaD framework. For instance, *Background* sections typically align with the **Introduction**, while *Experimental Setup* corresponds to the **Methodology** [1, 9]. These inconsistencies complicate automated section classification and create hurdles for comparative analysis of section-wise **bibliographic coupling**, where proper alignment across similar research sections is crucial.

The work by **Shahid et al.** [2] underscores several challenges in IMRaD mapping:

1. **Inconsistent Section Naming**: The lack of standardized section titles across disciplines makes it difficult for automated tools to align sections accurately.

2. **Cross-Domain Variability**: Fields such as computer science and medicine structure their papers differently, further complicating uniform classification.

3. **Ambiguity in Section Roles**: Some papers combine or split sections (e.g., merging *Background* and *Related Work*), creating further confusion for automated methods.

4. **Dependency on Domain Knowledge**: Effective mapping often requires specialized domain knowledge, limiting the effectiveness of general-purpose solutions.

1.3.2 Techniques for Identifying Related Papers

This section introduces two major classes of techniques for identifying related research papers: content-based and metadata-based. Each class includes specific methods with its own strengths and limitations, which are explained below.

The discussion is grouped into subsections as follows:

- **Content-Based**
- **Metadata-Based**
  - **Co-Citation**
  - **Bibliographic Coupling**

1.3.2.1 Content-Based

Content-based filtering (CBF) has emerged as a pivotal method in the research paper recommendation process, leveraging advanced weighting mechanisms to extract key terms and features from documents. One of the earliest and most

widely used approaches is the Term Frequency-Inverse Document Frequency (TF-IDF) scheme [13]. TF-IDF assigns weights to terms within a document based on their frequency in the text while reducing the influence of commonly used terms across the entire corpus. This weighting system improves the relevance of terms that are unique or particularly significant to a specific document, aiding in accurate identification of related research papers.

Building upon the foundation of TF-IDF, sophisticated methods have emerged. Latent Semantic Analysis (LSA) provided a breakthrough by analyzing co-occurrence patterns in the data to identify hidden structures in term-document relationships [16]. This approach captures underlying semantic meanings of terms, offering a deeper understanding of document content beyond simple keyword matching. Techniques like LSA laid the groundwork for dynamic and nuanced recommendation methodologies.

Further advancing the field, deep learning techniques have been integrated to refine semantic feature extraction. Neural networks, particularly word embeddings like Word2Vec [14], capture contextual relationships between terms by learning word vectors from large corpora. This approach moves beyond simple keyword matching, improving the ability to identify relevant content. Moreover, advanced models such as Bidirectional Encoder Representations from Transformers (BERT) have been employed to comprehend complex language structures in research papers [15], enhancing the accuracy of recommendations.

The Jensen-Shannon Divergence (JSD) is a symmetric and bounded metric for measuring the similarity between probability distributions, often used in text-relatedness tasks. It compares term frequency or topic distributions between documents, capturing semantic relatedness. A lower JSD value indicates higher similarity.

The application of CBF in personalized recommendation systems has been explored in several studies. For instance, [17] developed a personalized research paper recommendation system by building user profiles based on keyword extraction and calculating similarity using cosine metrics. Similarly, [18] employed bisociative

information networks (BisoNets) to recommend papers from distinct research domains, utilizing TF-IDF for weighting. These approaches highlight the flexibility of CBF in various contexts, including serendipitous discovery and cross-domain recommendations.

However, CBF is not without its challenges. While it performs well in identifying thematic elements within papers, the approach is computationally intensive, particularly when building personalized profiles for each user [19]. Additionally, CBF systems often rely heavily on structured textual data, making them less effective for papers with inconsistent formatting or unstructured content [20].

Another limitation is the difficulty in incorporating popularity metrics, which can hinder the ranking of equally relevant papers. Hybrid methods, such as those proposed by [21] and [22], attempt to overcome this limitation by combining content-based metrics with bibliometric data to improve recommendation quality.

Despite these challenges, the advantages of CBF lie in its ability to provide personalized recommendations without requiring large collaborative datasets. Techniques such as ontology-based similarity [23] and context-rich network mining [24] further demonstrate the adaptability of CBF systems in various research scenarios. CBF plays a critical role as the field evolves, especially in systems focused on individual preferences and specialized academic needs.

Advantages

- **Personalized Recommendations:** Content-based filtering uses a user's previous reading history or preferences to suggest papers aligned with their interests.

- **No Cold Start for Existing Users** The system performs well for users with an established profile since it already has sufficient data to generate accurate recommendations.

- **Explainability** Recommendations are transparent because they rely on the paper's content (e.g., keywords, abstracts, titles), making it clear why specific papers are recommended.
- **Domain Independence** This approach can be applied across various domains as it only depends on the content of the papers, such as abstracts or metadata, without needing external data.
- **Scalability** Algorithms such as TF-IDF (Term Frequency-Inverse Document Frequency) or cosine similarity allow for efficient and scalable solutions in recommendation systems.

Limitations

- **Cold Start Problem for New Users** If a user has no prior activity, the system struggles to generate relevant recommendations due to the lack of user-specific data.
- **Over-Specialization** The system may recommend papers that are too similar to what the user has already seen, limiting the diversity of suggestions.
- **Limited Novelty Discovery** It may miss recommending novel or interdisciplinary papers outside the user's typical reading pattern, reducing uncertainty.
- **Content Quality Dependency** The effectiveness of recommendations depends heavily on the quality and completeness of metadata (such as accurate keywords and abstracts).
- **Inability to Capture Implicit Interests** Content-based filtering struggles to account for implicit preferences, such as evolving interests or trends, as it relies purely on explicit textual data.

1.3.2.2 Co-Citation

Co-citation analysis identifies instances where two or more documents are cited together, suggesting a potential relationship or similarity between the cited works.

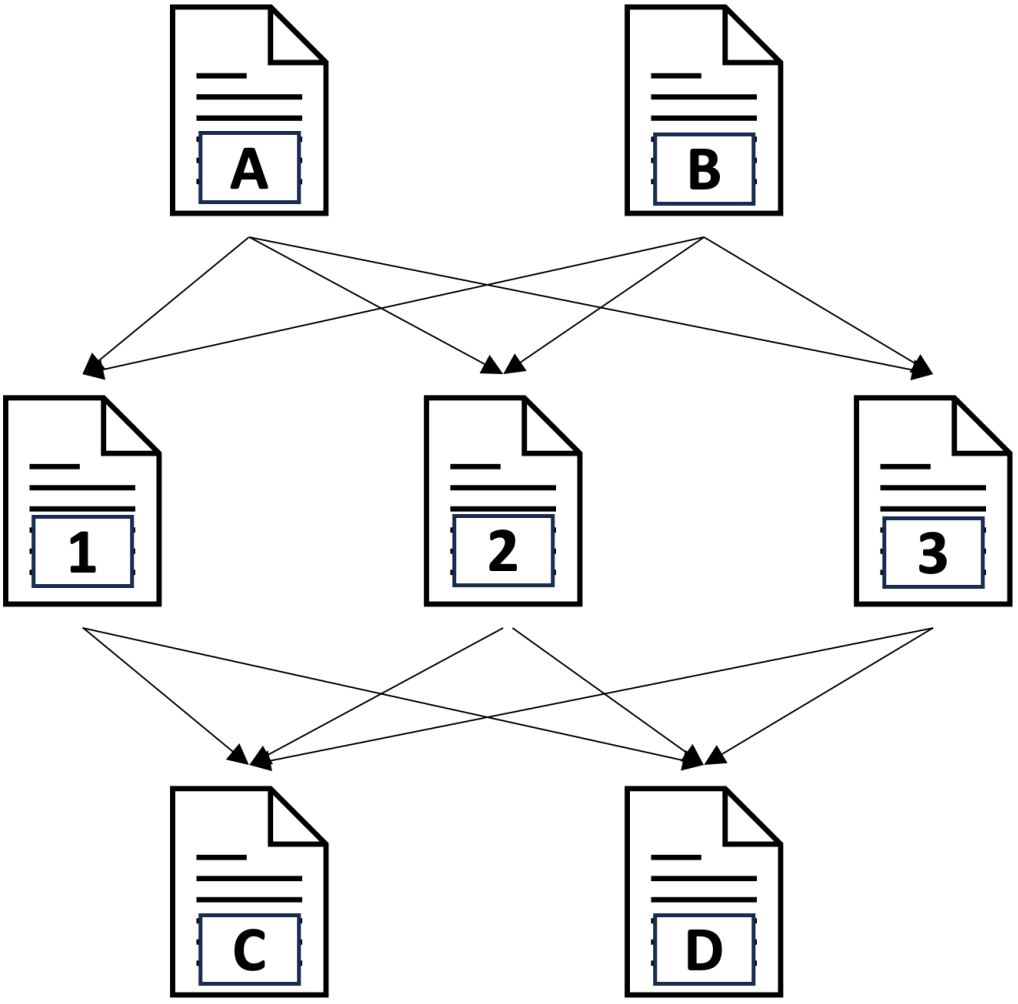


FIGURE 1.1: Bibliographic Coupling and Co-Citation [26]

This method has been instrumental in mapping research fields and uncovering connections among scholarly papers since its early development [25]. Traditional co-citation techniques rely on historical citation data, inherently making them static and less responsive to evolving research landscapes. Consequently, they may struggle to capture emerging trends or newly established research areas.

In Figure 1.1, the C.C. strength for papers C and D is also 3, as all three citing papers (1, 2, and 3) cite both C and D. For papers C and D, which are both cited by papers 1, 2, and 3, their co-citation strength (C.C. strength) can be calculated as:

$$\text{C.C. strength}(C, D) = \sum_{k=1}^{n} \text{C.C.}(C_k, D_k)$$

where:

$$\text{C.C.}(C_k, D_k) = \begin{cases} 1, & \text{if paper } k \text{ cites papers } C \text{ and } D, \\ 0, & \text{otherwise.} \end{cases}$$

Advancements in co-citation analysis over the years have sought to refine its applicability. In the 1980s, Small and Griffith [7] explored how clusters of co-cited documents could reveal intellectual structures within a discipline. Later, White and McCain [27] further demonstrated how author co-citation analysis could map research specialties. These traditional methods, however, were limited by their static nature.

Researchers have recently turned to dynamic co-citation models to address these limitations. For instance, Boyack and Klavans [28] proposed a method for mapping science that adjusts co-citation networks over time, providing a flexible approach to capturing changes in academic discourse. Recently statistical models enable even dynamic analysis by incorporating temporal changes in citation patterns. These models effectively detect shifts in academic discourse, providing insights into new areas of study and evolving intellectual landscapes.

Despite these improvements, co-citation analysis still faces several challenges. A significant issue is computational complexity, especially when handling large-scale bibliographic datasets. Processing extensive networks of scholarly citations requires substantial computational resources, which can hinder real-time analysis capabilities [29]. Moreover, co-citation analysis depends heavily on existing citation data, leading to the 'cold start' problem, where newer or lesser-known research papers are often excluded due to lacking citations [30]. This limitation can result in overlooking potentially groundbreaking but still developing research areas.

Advantages

- **Identifying Research Clusters** Co-citation analysis helps identify clusters of related works, revealing intellectual communities or research fields.

- **Mapping the Evolution of Research** It provides insights into the development and progression of research topics over time, making it useful for historical analysis of scientific fields.

- **Uncovering Relationships Across Disciplines** Co-citation analysis can reveal connections between research areas that may not be immediately obvious, encouraging interdisciplinary research.

- **Assessing Author Influence** It allows for measuring the influence of individual authors or research groups within a specific domain based on how frequently their works are co-cited.

- **Data-Driven Framework** Co-citation analysis offers an objective method for mapping knowledge structures using citation data.

Limitations

- **Bias Towards Older Works** Since co-citation depends on papers being cited together, older works tend to accumulate co-citations, limiting the visibility of newer research.

- **Dependence on Citation Practices** Co-citation patterns can vary across disciplines, and differences in citation behaviors may affect the accuracy of cross-disciplinary comparisons.

- **Incomplete Citation Data** The analysis relies on citation databases that may contain missing or erroneous data, leading to biased results.

- **Limited Context Understanding** Co-citation only captures numerical patterns of citations without providing qualitative insights into the specific nature of the relationship between the co-cited papers.

- **Computational Complexity** Analyzing large citation networks for co-citation patterns can be computationally intensive, requiring advanced tools and algorithms.

1.3.2.3 Bibliographic Coupling

Bibliographic coupling measures the similarity between research papers based on their shared references. First introduced by Kessler in 1963 [6], this technique suggests a thematic connection between papers that have overlapping references, offering an alternative to citation analysis for mapping research areas. Early on, it was primarily a static approach, relying on the number of shared references to infer the strength of the connection between papers.

Figure 1.1 illustrates the concept of bibliographic coupling and co-citation. It shows that papers A and B both cite papers 1, 2, and 3. The bibliographic coupling strength (B.C. strength) between papers A and B is calculated using the following equation:

$$\text{B.C. strength}(A, B) = \sum_{k=1}^{n} \text{B.C.}(A_k, B_k)$$

where:

$$\text{B.C.}(A_k, B_k) = \begin{cases} 1, & \text{if papers } A \text{ and } B \text{ cite paper } k, \\ 0, & \text{otherwise.} \end{cases}$$

In Figure 1.1, the B.C. strength for papers A and B is 3, as they both cite papers 1, 2, and 3. In the following decades, bibliographic coupling has seen numerous refinements. Small [25] expanded on the concept by exploring how coupling strength can reveal clusters of related scientific research. This clustering ability paved the way for a sophisticated techniques that began considering the age of references, suggesting that shared citations to recent works might indicate a substantial relationship [31]. Additionally, Glänzel and Czerwon [32] examined bibliographic coupling in the context of research collaboration, highlighting its use in identifying emerging research topics.Recent methods have further improved the precision of bibliographic coupling by integrating different weighting mechanisms. For instance, weighted bibliographic coupling [3] assigns varying degrees of importance to references, acknowledging that not all citations have equal significance in establishing thematic connections between papers. Section-based approaches have also been

developed to account for the placement of references within a paper. Khan et al. [9] proposed a section-based weighting model, suggesting that references in sections like Methodology or results carry influence than those in introductions, enhancing the overall relevance of the coupling. Dynamic bibliographic coupling has emerged as an important area of focus, emphasizing the evolving nature of research networks. Boyack and Klavans [28] introduced methods to incorporate temporal dynamics into coupling analysis, allowing for the detection of changing research trends over time. Recently, Gündoğan and Kaya [33] presented a novel hybrid paper recommendation system that utilizes deep learning, which offers an accurate and context-aware mapping of scientific literature. This approach addresses some of the limitations of traditional coupling methods, such as their reliance on historical citation data.

Despite these advances, bibliographic coupling still requires further refinement, particularly in incorporating the contextual weight of references within specific sections of research papers. For instance, a shared reference in the results section might indicate a stronger relationship between two papers than in the introduction. Addressing this contextual weighting is the central focus of this study, aiming to enhance the effectiveness of bibliographic coupling in identifying related research papers accurately.

Advantages

- **Identifying Related Research** Bibliographic coupling helps uncover relationships between papers, even if they do not cite each other directly.
- **Analyzing Research Trends** By clustering papers with shared references, it becomes easier to trace the development of specific research topics over time.
- **Useful for Newer Publications** Unlike co-citation analysis, which favors older papers, bibliographic coupling effectively identifies relationships among newer works, as it only requires shared references.

- **Mapping Intellectual Structure** It enables the identification of research fields, subfields, or academic communities through clusters of bibliographically coupled documents.
- **Data-Driven Insights** The method provides an objective framework for analyzing relationships between research papers using citation data.

Limitations

- **Dependence on Citation Practices** Citation behavior varies across disciplines, affecting the accuracy of field comparisons.
- **Limited Qualitative Insights** While bibliographic coupling identifies numerical relationships, it does not reveal the context or quality of the shared references.
- **Database Limitations** The completeness and accuracy of bibliographic coupling depend on the quality of citation data available in databases like Scopus or Web of Science.

1.3.3 Co-Citation vs. Bibliographic Coupling

Co-citation and bibliographic coupling are valuable methods for identifying connections between research papers. Co-citation involves recognizing that two papers are frequently mentioned together, indicating a historical association, much like when people discuss two related research topics. On the other hand, bibliographic coupling is comparable to discovering two papers that directly reference the same sources or ideas, signifying a strong content connection.

A notable advantage of bibliographic coupling is its dynamic nature, which updates as new papers with shared references emerge. In contrast, co-citation tends to remain static over time, reflecting past connections, while bibliographic coupling is dynamic, with its strength potentially increasing as new research papers with shared references emerge.

1.4 Section-Based Bibliographic Coupling

In Section-Based Bibliographic Coupling, what matters most is how closely related two papers are based on the specific parts or sections that cite the same source. If they mention a common source within the same section, like the Results section, it impacts their connection than citing the same source in different sections, such as the Introduction. For instance, if two papers refer to a paper in their Results section, it indicates a stronger relationship than if one paper talks about a source in the literature review and the other in the introduction. It's about understanding the closeness between papers by looking at where they share references within each section.

1.4.1 Weighted Section-Based Bibliographic Coupling

Weighted Section-Based Bibliographic Coupling emphasizes the significance of specific sections in research papers. When both Paper A and Paper B reference the same paper (C) in the results section, their connection is pronounced than if they cite a paper in the introduction. This highlights the need to assign weights to various sections, acknowledging that certain parts of a paper, like the results section, can significantly influence relatedness. In simple terms, it's about recognizing the importance of different sections and giving them appropriate importance when gauging how closely papers are connected.

1.4.2 Dynamically Tuning of Section's Weights

Given the recognition that different sections in research papers carry varying levels of importance, as highlighted by Weighted Section-Based Bibliographic Coupling, the concept of Dynamically Tuning Section Weights comes into play. Dynamically tuning section weights involves adjusting these importance levels dynamically, allowing for a nuanced and responsive approach to evaluating the significance of various sections in understanding the relatedness between research papers. It's

about fine-tuning the emphasis on different sections based on their impact on the connection between papers.

1.4.3 Significance of research paper sections

In Section-Based and Weighted Section-Based Bibliographic Coupling, a crucial consideration revolves around the sections in research papers. Since researchers use different headings and sections in their papers, correlating the sections of two papers becomes a challenge. To address this, a common theme for section naming is needed. IMRaD (Introduction, Methodology, Results, and Discussion) offers a solution with predefined sections. This facilitates the identification of actual section boundaries and mapping them to the IMRaD logical structure.

1.5 Research Problem

- Identify the research paper's sections and map them to IMRaD.The best technique developed to identify and map sections has an F-measure of 0.78.

- Assign weight to IMRaD sections. To discover related papers using static weights produce a 0.77 correlation with JSD.

1.6 Problem Statement

- Current literature lacks a method with better accuracy to map sections to IMRaD; therefore, there is a need to develop the method.

- Current literature lacks a method for discovering related papers using dynamically adjusted section weights in bibliographic coupling. Therefore, there is a need to develop a scheme to dynamically tune section weights.

1.7 Research Questions and Objectives

1.7.1 Research Questions

Here are the research questions that need to be answered to solve the problem mentioned in the problem statement.

- **RQ1**: How can a method be devised with improved accuracy to map the sections of research papers to the IMRaD structure, considering the variations of these sections?
- **RQ2**: How can sections' weights be tuned to maximize the correlation between Bibliographic Coupling strength and paper relatedness?

1.7.2 Research Objectives

- **RO1**: Extract sections from research papers and map them to the IMRaD structure to improve the precision of section mapping.
- **RO2**: Tune section weights to improve the correlation between section-wise bibliographic coupling strength and paper relatedness.

1.8 Research Scope

This research aims to enhance the identification of related research papers through dynamically tuned section-wise bibliographic coupling, focusing on:

- **Section Identification:** Improving the precision of mapping individual sections of research papers to their corresponding IMRaD logical sections.

- **Dynamic Weight Adjustment:** Implementing dynamic weight tuning to identify the section's importance in bibliographic coupling.

While this research aims to improve bibliographic coupling techniques, it is not focused on building a recommendation system.

Instead, the primary objective is to refine the processes and underlying mechanisms that enable the accurate identification of related research papers.

This distinction ensures that the research remains centered on enhancing bibliometric analysis and section-wise coupling, contributing foundational improvements that could indirectly support future recommendation systems.

1.9 Research Methodology

The research methodology comprises four phases, adapted from the eight-step model proposed by Kumar et al. [34]:

Phase 1: Deciding What to Research Chapter - 1

Step 1: Formulate a research problem through a systematic literature review, identifying gaps in current bibliographic coupling approaches.

Step 2: Write the research proposal, focusing on developing new bibliographic coupling methods that incorporate section weights.

Phase 2: Planning the Research Study Chapter 2

Step 1: Conceptualize a Research Design to answer RQ1 and RQ2

Step 2: Data Collection for RQ1 and RQ2

Step 3: Data Preprocessing for RQ2.

Phase 3: Conducting the Research Study Chapter 3, Chapter-4

Step 1: Perform experiments for RQ1 and RQ2.

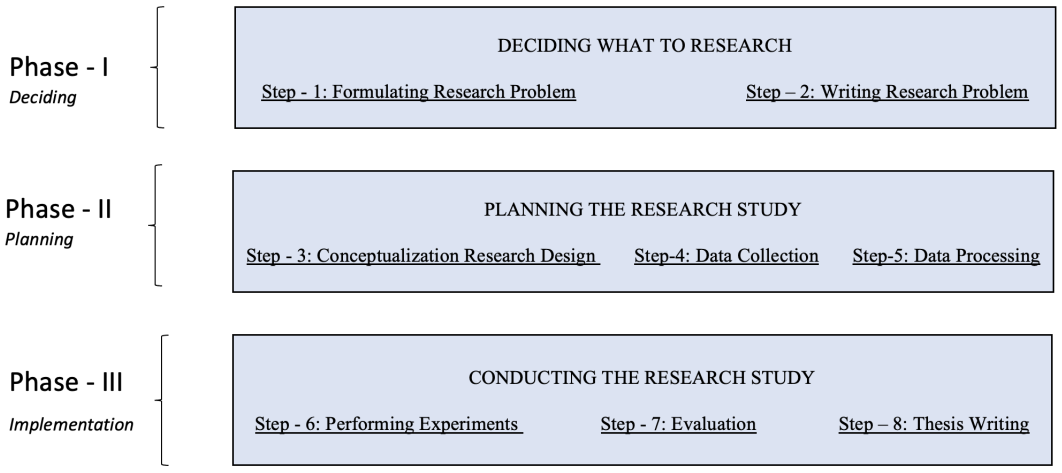


FIGURE 1.2: The Design Science Research Methodology Process for this work

Step 2: Evaluation and Comparisons of RQ1 and RQ2 with the state of art techniques.

Phase 4: Writing the Research Report

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

1.10 Dissertation Organization

- Chapter - 1: Introduction - This opening chapter sets the stage for the thesis by introducing the context of the study, stating the research problem, objectives, and research methodology. It serves as the groundwork for the reader to understand the research's scope and significance.

- Chapter 2: Literature Review - This chapter provides a comprehensive literature review of the existing literature in the field. It examines previous studies, frameworks, and methodologies established by the research community. The aim is to identify the gaps in the current body of knowledge that this thesis intends to fill.

- Chapter 3: Extraction of Section-Headings and Mapping to IMRaD Structure - This chapter presents the first key contribution of the thesis. It discusses the proposed Methodology for accurately identifying actual sections within research papers and mapping them to the logical IMRaD structure. The proposed strategy, its implementation, and the subsequent results are extensively discussed.

- Chapter 4: Section's Ranking and Weights Adjustment to Discover Bibliographically Coupled Papers - Chapter delves into the second main contribution of this research. It focuses on a novel technique for ranking sections and the dynamic assignment of weights in the context of discovering bibliographically coupled papers. The proposed strategy, its implementation, and the subsequent results are extensively discussed.

- Chapter 5: Overall Conclusion and Future Work - This final chapter summarizes the research findings, highlighting the contributions made in the field. It discusses the implications of the findings and their potential for future research. It concludes with reflections on the study and suggestions for possible directions for future research in the domain.

This systematic layout provides a step-by-step progression through the study, ensuring a coherent and comprehensive understanding of the research.

Chapter 2

Literature Review

2.1 Overview

The chapter provides a literature review of different techniques, such as content-based and metadata-based approaches, explaining the method, its primary goal, and the analysis of that technique. It also includes a literature review of methods for extracting sections from research papers and mapping them to the IMRaD structure (Introduction, Methodology, Results, and Discussion).

Additionally, it discusses the appropriate dataset to be used in this research.

The focus of the review is as follows:

1. Explore research on identifying sections in research papers and how they relate to the IMRaD structure.
2. Investigate existing techniques used to discover related research papers.
3. Explore the advantages of using bibliographic information to find related research papers.
4. Explore research on new methods to improve bibliographic coupling techniques.

5. Conduct a detailed analysis of existing datasets to identify the most suitable dataset of related papers.

2.2 Inclusion and Exclusion Criteria

To maintain focus and transparency, this literature review applied well-defined inclusion and exclusion criteria during the study's selection process. These criteria ensured the inclusion of the most relevant and high-quality research aligned with the study's objectives.

Inclusion Criteria

- Research publication published between the years 2001 to the year 2024.
- Peer-reviewed journal articles, conference papers, and book chapters relevant to scientific paper recommendations.
- Papers focused on methodologies such as bibliographic coupling, content-based filtering, metadata-based strategies, and hybrid systems. These keywords are used to find the relevant papers.
- The research publication is available in the English language.

Exclusion Criteria

- Any recommender system's publication unrelated to research papers relatedness, such as unrelated topics or non-academic domains.
- Duplicate studies or earlier versions of papers already included in the review.
- Articles with incomplete or unavailable data limit the ability to replicate or validate findings.
- Studies focused on irrelevant areas (e.g., non-computer science disciplines, clinical medicine, or niche industrial applications) that do not align with the scope of this review.

2.3 Techniques for Logical Section Extraction in Scientific Research Papers

The work by Ding et. al.[1] investigates the distribution of references across texts and its implications for citation analysis. Using a dataset of 866 articles from the Journal of the Association for Information Science and Technology (JASIST) published between 2000 and 2011, the study analyzes 32,496 references and 53,017 mentions. The authors employ algorithms to detect structural patterns and identify implicit sections based on content placement within the text. Regular expressions detect sections and their headings, while section boundaries are determined using linguistic and contextual information. However, the study concludes that the approach struggles with consistently identifying sections due to the formatting variability across papers.

The technique proposed by Shahid et al. [2] focuses on transforming unstructured or semi-structured scientific documents into documents with clearly defined logical sections. This method automates the tagging of sections such as Introduction, Methodology, and Results, improving the precision of information retrieval systems. It employs a two-pronged approach: a dictionary of key terms for accurate text segment identification and a structural layout template to guide classification. Tested on a dataset of 5,000 research papers from CiteSeer containing 39,420 sections, the method achieved a precision and recall rate of 0.78. Despite promising results, challenges include variability in section naming conventions and scientific terminology's evolving nature, necessitating regular dictionary and template updates. Habib et al. [3] proposed a method for section-based bibliographic coupling for research paper recommendation. Using a CiteSeer-based dataset (Dataset-1) comprising 320 papers divided into 32 subsets, the study extracts sections using XML tags. It maps them to generic types such as Introduction, Related Work, Methodology, Results, and Conclusion. The method successfully extracted and mapped sections with 90% accuracy for randomly selected papers. However, the research focuses solely on XML tags and does not consider the content of the sections, nor does it map sections to the IMRaD structure. Additionally, the study's evaluation is

limited to a small and non-representative dataset, which may not generalize well to larger collections.

2.3.1 Dataset for Extraction of Sections and Mapping to IMRaD

The dataset utilized in this study originates from the work of Shahid et al., sourced explicitly from the **Journal of Universal Computer Science (J.UCS)**, a multidisciplinary journal in Computer Science. The selection of this dataset was motivated by the diverse expertise of its contributors, providing a broad and representative foundation for investigation. Utilizing `pdfx`, documents from J.UCS were downloaded and converted into XML format. Sections were extracted in an initial dataset of 12,180 sections across 1200 research papers. However, a comprehensive preprocessing phase was conducted to mitigate noise and encoding issues, reducing the dataset to 7000 refined sections. From this, 329 research documents were randomly selected for detailed analysis, producing a subset of 1833 sections.

Two domain experts manually annotated the dataset, following the methodology described by Shahid et al. [2]. The experts classified sections into predefined logical categories—*Introduction, Related Work, Methodology, Results, Discussion, and Conclusion*—by examining section titles and consulting the document content for accuracy. This process aligns with established tasks in the literature, such as sentence classification into predefined categories. The annotation achieved a high level of reliability, with an **inter-annotator agreement kappa value of 0.91**, reflecting strong consistency between the experts. The dataset, derived from the structured content of J.UCS, provides a robust resource for examining the classification of academic content and its application in bibliographic coupling. Its combination of automated extraction, rigorous preprocessing, and expert-driven annotation makes it a reliable foundation for text classification and bibliometric analysis research.Therefore, the dataset can be utilized to identify various sections within academic documents and accurately map them to the IMRaD structure

(Introduction, Methodology, Results, and Discussion). This mapping provides a systematic way to categorize and analyze the content of academic papers. Furthermore, the results obtained from this analysis can be compared with findings from previous research, allowing for a deeper understanding of trends, patterns, and potential discrepancies in how different studies structure and present their findings. Such comparisons can also contribute to validating the methodology and identifying areas for further refinement.

2.4 Literature Survey on Techniques for Finding Related Papers

A literature review explored methods for identifying related academic papers, focusing on techniques like bibliographic coupling, co-citation analysis, and text-based approaches. The strengths and weaknesses of different techniques for finding related papers were then examined. The mechanism and effectiveness of each method in uncovering relevant literature were carefully assessed. The most notable techniques were selected based on their relevance and impact. This involved identifying those with practical approaches for identifying related papers and addressing research gaps.Papers cited in the literature were reviewed to gain deeper insights into the selected techniques. This included analyzing seminal works and understanding how methods have evolved. Acknowledging the limitations of relying solely on older literature surveys, the latest techniques were adapted by incorporating keywords such as "bibliographic coupling," "co-citation," and "text-based techniques" to identify and review the most cited and influential papers in the field.

2.5 Techniques to Find Related Research Papers

A comprehensive investigation by Beel et al. [35] found that hundreds of paper recommendation approaches were introduced. These approaches can be categorized

into multiple types. These diverse approaches contribute to the richness of the research paper recommendation landscape, each offering unique perspectives and methods to enhance the discovery of relevant scholarly articles.

- Content-based Strategies
- Metadata-based Approaches
- Collaborative Filtering Based Approaches
- User Profile-Based Approaches
- Data Mining Approaches

2.5.1 Content-based Strategies

Content-based strategies for research paper recommendation involve meticulously examining each document's intrinsic features and characteristics. Ding et al. [1] focuses on analyzing textual components, including titles, abstracts, and keywords, to discern the content similarities among papers. Researchers have introduced techniques to enhance the effectiveness of recommendation systems. The latest advancements include:

- Keyphrase-Based
- Keywords-Based
- Concept-Based
- Graph-Based
- Hybrid Approaches

Keyphrase-Based

Ferrara et al

The Keyphrase-Based approach was initially proposed by Ferrara et al. [36]. This method involves several steps: first, candidate phrases are extracted through POS tagging and n-gram extraction, followed by stemming and stopword removal. Next, statistical and linguistic properties characterize each candidate phrase, including frequency, POS value, depth, last occurrence, and lifespan. These features compute a score for each word, and the top-ranked keyphrases are selected based on a predefined threshold. A matching score between user profiles and document representations is calculated using cosine similarity. Results show that the proposed method outperforms with a precision of 0.93 compared to 0.83 precision of unigram. However, challenges include accurate keyphrase extraction, robust user profile construction, scalability for larger datasets, and generalization across domains.

Sarkar et al. 2010

A study by Sarkar et al. [37] proposed an innovative keyphrase extraction model using neural networks. Their work demonstrates improvements in precision and recall over traditional statistical methods. The model leverages large-scale datasets to capture contextual information in text processing, showcasing the potential of neural networks in enhancing keyphrase extraction.

Umair et al. 2022

Umair et al. [38] proposed a novel keyphrase extraction method using pre-trained language models (PLMs), combining attention mechanisms and semantic similarity. Their approach addresses the limitations of traditional extraction models by capturing deeper contextual relevance and enhancing performance on complex documents. Additionally, graph-based ranking and phrase-document similarity techniques have been integrated into PLMs to achieve superior results in keyphrase extraction, improving accuracy without needing labeled data, thus supporting low-resource settings and domain-specific tasks.

Ajallouda et al.'s Overview of Deep Learning in AKE (2023):

Ajallouda et al. [39] presented a comprehensive review of automatic keyphrase

extraction (AKE) using deep learning methods. Their study highlights how AKE evolved from traditional machine learning techniques to advanced neural networks, including RNNs, CNNs, and autoencoders. The paper discusses both keyphrase extraction (from the text) and keyphrase generation (predicting absent phrases) approaches. It identifies challenges in keyphrase extraction, such as semantic redundancy, and offers future directions for enhancing extraction precision through hybrid models that integrate supervised and unsupervised learning techniques.

Liu et al. 2024

Liu et al. (2024) [40] introduce AdaptiveUKE, an unsupervised keyphrase extraction model employing gated topic modeling to address semantic diversity challenges. The approach assigns topics independently based on document richness and uses a novel scoring algorithm considering topic importance and relatedness, ensuring both keyphrase relevance and diversity. Experimental results across datasets like Inspec and SemEval2010 demonstrate the model's superior performance, surpassing state-of-the-art baselines by notable margins. This study highlights the importance of adapting to topic variability for improved extraction quality in real-world documents.

Keyword-Based

Zhang et al. (2021)

In 2021, Zhang et al. [41] proposed a keyword-based recommendation system using a hybrid approach integrating keyword extraction with semantic analysis. This system addresses the limitations of earlier methods by incorporating contextual information, thereby enhancing the retrieval accuracy of research papers. Evaluation using datasets from various digital libraries shows a significant increase in recommendation precision.

Pohan et al.'s (2022)

Pohan et al. [42] conducted a systematic literature review on transformer models within recommender systems. Their study highlights the growing importance

of online transactions during the pandemic and the challenges of providing personalized recommendations among many products. The authors examine how transformers, with their parallel processing capabilities, outperform traditional models such as Recurrent Neural Networks (RNNs) in processing large-scale data. Their literature review identifies key applications of transformer-based systems, ranging from e-commerce platforms to academic research repositories, emphasizing recommendation speed and precision improvements. They also document the dominance of transformer-based methods in recent studies, demonstrating how these models have become essential for next-generation recommender systems.

Concept-Based/User Profile Based

Research in personalized recommendation systems has increasingly focused on improving user satisfaction by expanding user profiles through advanced techniques such as semantic analysis. The proposed APRPRS (Advanced Personalized Research Paper Recommendation System) by [43] aims to enhance the user profile with semantic keyword expansion, leading to improved recommendation accuracy. The system leverages WordNet to identify semantically similar keywords, thus enriching the user profile and increasing the relevance of recommended papers. Previous work on recommendation systems has used techniques like collaborative filtering and rule-based filtering, but these approaches lack the adaptability of semantic keyword expansion. Additionally, semantic expansion to enhance personalization supports the hypothesis that including contextually relevant keywords can improve recommendation outcomes. This paper contributes to the field by demonstrating a 9% increase in user satisfaction when using expanded semantic profiles, showcasing a substantial improvement over traditional methods.

Graph-Based

Huang et al. (2002)

The hybrid recommendation system proposed by Huang et al.[44] utilizes content-based and collaborative methods within a graph-based model to enhance book

recommendations. This system constructs a two-layer graph where books are linked by content similarity and customers by demographic similarity, with purchase history forming inter-layer links. Recommendations are generated using graph search techniques, such as the Hopfield net algorithm. The evaluation, which uses precision and recall metrics, shows that the hybrid approach outperforms pure content-based and collaborative methods. However, content-based recommendations excelled in some assessments, suggesting a heavy reliance on book content. Challenges include the computational complexity of managing high-degree associations and the difficulty in representing user interests solely based on purchase history.

Hybrid Approaches

Guo et al.'s Paper Discovery Technique

The approach by Guo et al. [45] uses structural and content information within papers to improve academic paper discovery. By constructing paper-author and paper-area heterogeneous information networks and using citation relationships, author collaborations, and research area details, the model employs a random walk-based strategy to uncover paper relevance. Despite its success in outperforming specific traditional and graph-based methodologies, it faces challenges like potential biases and the complexity of accurately modeling scholarly relationships.

Kanakia et al. [46] developed a scalable hybrid recommendation system for Microsoft Academic to manage approximately 160 million English research papers and patents. This system addresses challenges such as incomplete citation data and the cold-start problem in existing recommender systems. Integrating co-citation and content-based approaches balances the novelty and authority of recommendations. The evaluation through a user study indicated a strong correlation between participant scores and the system's similarity rankings, revealing areas for improvement in precision, particularly in content-based recommendations.

Liu et al.'s Enhanced Model (2020)

Recent advancements in keyphrase extraction have emphasized hybrid and neural

network-based approaches to enhance precision and adaptability. Traditional statistical models, while effective, often fall short of capturing the deeper semantic relationships within documents. Researchers like Zhang and Xu (2020) introduced hybrid keyword-based recommendation systems that combine statistical and semantic techniques to improve the relevance of paper recommendations. Guo et al. (2015) [45] explored heterogeneous information networks to enhance academic paper discovery, demonstrating the power of connected data. As mentioned in recent studies, transformer-based models are proving instrumental in contextual keyword extraction, enabling dynamic adjustments to varying document structures. This is complemented by research from Dong et al. (2019), highlighting the limitations of content-based filtering in paper recommendations, urging the need for advanced, adaptive systems. The integration of neural networks, as demonstrated by Liu et al. (2020), further refines keyphrase extraction by leveraging large datasets and contextual embeddings. Emerging trends point towards combining deep learning techniques with traditional methods, achieving higher precision and recall. With the evolving complexity of data, adaptive extraction systems are becoming increasingly crucial for personalized and context-aware paper recommendations.

Nair Machine Learning Approach (2021)

Nair et al. [47] propose a novel *Content-Based Scientific Article Recommendation* (C-SAR) model using a combination of deep learning and classical algorithms. The C-SAR model leverages Gated Recurrent Units (GRU) and the Apriori algorithm to provide article recommendations based on content similarity. GRUs capture sequential patterns in text data, while Apriori aids in frequent itemset mining, offering an additional filtration layer for relevant articles. This hybrid approach outperforms traditional methods by integrating content-based filtering with machine learning models.

Recent advancements highlight the importance of deep learning models for improving recommendation systems. Practical, collaborative filtering and citation-based methods often struggle with the cold-start problem and sparse data. Researchers have explored neural networks, such as Long-Short-Term Memory (LSTM), to

address these limitations for personalized recommendations. Additionally, concept-based models and graph-based techniques are increasingly being adopted to enhance the precision of recommendations.

Recurrent neural networks (RNNs) have gained traction for processing sequential data and modeling user behavior. Studies show that incorporating user feedback improves the relevance of recommendations. However, while offering personalization, purely content-based approaches often lack scalability for large datasets. The C-SAR model aims to overcome these limitations using a hybrid approach, making it suitable for scientific repositories with extensive datasets.

Su-Anne Teh et al. 2023

Su-Anne Teh et al. [48] explored a hybrid-based research article recommender system that integrates multiple recommendation techniques, including content-based filtering, collaborative filtering, and hybrid filtering approaches. Their work highlights the advantages of combining different recommendation methods to address challenges such as data sparsity, cold start, and scalability, ultimately enhancing performance and accuracy in personalized recommendations. The study demonstrated the effectiveness of recommender systems in e-commerce and academic contexts, where they help users save time and improve decision-making by providing customized suggestions.

In summary, the fusion of content-aware deep learning techniques [49] with data mining algorithms presents a promising direction for research paper recommendations. Models like C-SAR pave the way for personalized and efficient literature retrieval, addressing challenges related to information overload researchers today face.

2.5.2 Metrics for Evaluating Research Paper Similarity

Several similarity measures have been proposed and utilized in the literature for text similarity tasks. This review focuses on the most widely used measures.

Cosine Similarity

The Cosine Similarity formula, shown in Equation 2.1, measures the cosine of the angle between two vectors A and B. It provides a similarity score independent of the magnitude of the vectors, making it highly suitable for text similarity tasks with varying document lengths. This measure ranges from 0 to 1, where 1 indicates identical orientation, meaning the texts are aligned, and 0 indicates no similarity (orthogonality) [50].

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\|\|B\|} \tag{2.1}$$

Euclidean Distance

The Euclidean Distance formula (Equation 2.2) measures the straight-line distance between two points in n-dimensional space. It is straightforward but less effective in high-dimensional spaces due to the "curse of dimensionality." While it captures absolute differences, it does not consider vector direction, limiting its utility for text similarity tasks.

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \tag{2.2}$$

Jaccard Similarity

Jaccard Similarity, shown in Equation 2.3, measures the similarity between two sets A and B. It is defined as the size of the intersection divided by the size of the union of the sample sets. Jaccard Similarity is especially useful for comparing the diversity of sets but is limited in text similarity tasks as it does not account for term frequency [51].

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.3}$$

Manhattan Distance (MD)

Manhattan Distance (Equation 2.4), also known as L1 distance, is the sum of the absolute differences of the coordinates of two vectors. It is useful for capturing absolute differences but is sensitive to data scale [52].

$$\text{Manhattan Distance} = \sum_{i=1}^{n} |A_i - B_i| \tag{2.4}$$

Hamming Distance

Hamming Distance, presented in Equation 2.5, counts the number of positions at which two equal-length strings differ. It is particularly useful for categorical data but less effective for text similarity tasks [53].

$$\text{Hamming Distance} = \sum_{i=1}^{n} (A_i \neq B_i) \tag{2.5}$$

Jensen-Shannon Divergence (JSD)

Jensen-Shannon Divergence (Equation 2.6) measures the similarity between two probability distributions. It is a symmetric and finite measure based on the Kullback-Leibler divergence. The use of JSD is advantageous because it provides a bounded measure, making it easier to interpret [54].

$$\text{JSD}(P\|Q) = \frac{1}{2} D_{KL}(P\|M) + \frac{1}{2} D_{KL}(Q\|M) \tag{2.6}$$

where $M = \frac{1}{2}(P + Q)$ and D_{KL} is the Kullback-Leibler divergence.

2.5.3 Techniques Based on Metadata

In academic research, the quest to find related papers has evolved beyond traditional keyword searches and basic text parsing. Utilizing metadata—including

elements such as a paper's title, abstract, keywords, author information, publication date, references, and citation data—has become critical in uncovering intricate connections between research works. Metadata acts as a comprehensive framework that allows for a nuanced exploration of research literature, revealing links that would otherwise remain hidden when relying solely on content-based methods. Metadata-based features are not only used to find related papers but also help to Classify Research Paper Topics [55].

Employing metadata significantly transforms information retrieval processes. It allows computer programs to identify and analyze academic documents' structural and contextual elements, leading to a systematic approach to discovering related papers. This metadata-driven strategy enables researchers to uncover latent connections within the scholarly literature, offering new pathways for scientific inquiry [56]. However, this approach also introduces complexity, as navigating vast and varied metadata requires sophisticated algorithms and computational techniques. Metadata essentially serves as a key to unlocking the complex interrelationships in academic research. It simplifies grasping the core contributions of research papers, not just for specialized audiences but also for a wider readership. Despite the challenges in its application, particularly regarding data quality and consistency, the benefits of incorporating metadata in scholarly research are substantial. It facilitates bridging gaps between established knowledge and emerging insights, thus propelling the advancement of science.

Bethard et al. (2008)

Bethard et al. [57] proposed a system that learns the relative importance of various factors deemed crucial by researchers through citation patterns. These factors include term similarity, citation frequency, the recency of citations, usage of similar terms in citations, thematic similarity, and social patterns (such as collaboration networks). The system employs a linear classifier to combine these factors into a scoring function that ranks articles, with the classifier learning feature weights from the citation network. In their evaluation, using a dataset of 10,921 papers from the Anthology Reference Corpus, the system demonstrated a significant improvement

in mean average precision over methods that used only related work features or did not incorporate iterative learning.

Shahid et al. (2009)

Shahid et al. [58] introduced a method based on in-text citations to identify related papers within an article's body. Their approach delved into the internal links between citing and cited papers, proposing that the frequency of in-text citations serves as a critical indicator of paper relatedness. The underlying assumption was that papers cited more than five times within the text will likely have a significant thematic overlap with the citing paper. This method highlighted the importance of understanding how citations are distributed within an article, offering a detailed perspective on paper interconnectivity.

Nassiri et al. (2010)

Nassiri et al. [59] developed the Normalized Similarity Index (NSI) to measure the similarity between papers within a citation network. The NSI incorporates co-citations, bibliographic, and longitudinal coupling, providing a holistic measure of paper similarity. Their study found a high correlation between NSI results and peer reviews, suggesting that NSI could serve as an effective proxy for expert assessments. In comparative analysis, NSI outperformed combined linkage and weighted direct citation techniques to identify closely related research.

Krapivin et al. (2011)

Krapivin et al. [60] adapted the PageRank algorithm, designed initially for ranking web pages, to suit the academic domain for identifying related papers. Given that academic papers typically contain numerous outbound citations, they recognized the unique challenge this posed to traditional PageRank. To address this, they introduced the Focused PageRank approach, which modified the algorithm to capture the significance of references in academic documents better. In their evaluation using a dataset of 266,788 ACM papers, Focused PageRank demonstrated higher effectiveness than basic Citation Count and traditional PageRank, underscoring the value of tailored ranking algorithms in scholarly contexts.

Gori et al. (2012)

Gori et al. [61] further refined the use of PageRank in academic settings by developing a citation graph model. Their method utilized random walks with properties of attenuation and propagation to better represent paper relationships in the citation network. They introduced matrices to represent these relationships, achieving a 100% ranking accuracy for related papers in their experiments.

This work highlighted the potential of advanced graph-based techniques in enhancing the identification of relevant literature.

El-Arini et al. (2013)

El-Arini et al. [62] challenged the traditional reliance on keyword-based queries for academic paper searches. Instead, they proposed using a paper's references to yield relevant search results, introducing the concept of influence flow to capture the transmission of ideas from cited to citing papers. Their approach outperformed existing systems such as Google Scholar, demonstrating the superiority of reference-based searching in understanding the intellectual lineage and thematic continuity of research.

Strohman et al. (2015)

Strohman et al. [63] introduced a novel system where users could submit incomplete documents to retrieve related papers. Their method employed text analysis, citation analysis, and feature-based similarity measures. By incorporating six distinct features in a two-step ranking process, they significantly improved the accuracy of paper recommendations, offering a tailored search experience for researchers in drafting new research manuscripts.

Reyhani et al. (2017)

Reyhani et al. [64] proposed SimCC, a method for calculating the similarity between two papers based on the contribution score of the cited paper to the citing paper. By combining content analysis and citation analysis, SimCC aimed to capture references' nuanced contributions to the citing work. Their evaluation demonstrated that SimCC outperformed other similarity metrics, such as cosine

similarity and Dice's coefficient, highlighting the importance of considering the contextual impact of citations.

Bichteler et al. (2018)

Bichteler et al. [65] explored the combined use of bibliographic coupling and co-citation analysis to recommend related papers. Their study found that while each method had individual merits, combined use provided a comprehensive view of paper interrelationships. A subsequent user study supported the effectiveness of integrating bibliographic coupling and co-citation analysis, underscoring the value of multi-faceted approaches in academic recommendation systems.

Gipp et al. (2019)

Gipp et al. [8] introduced Citation Proximity Analysis (CPA), a technique designed to find similar documents and assist researchers in literature discovery. CPA considers the proximity of citations within a document, arguing that closely spaced citations are likely thematically related. This method provided precise results than traditional co-citation analysis and was evaluated using a dataset from Scienstein.org, which contained 1.2 million publications. CPA's precision in capturing meaningful research connections highlighted its potential to enhance literature review processes.

Mustafa et al. (2021)

Mustafa et al. [55] thoroughly evaluated metadata-based features in classifying research paper topics. They highlighted the limitations of relying solely on content due to accessibility issues with many journal articles. By combining metadata elements like title, abstract, keywords, and general terms, the authors demonstrated that metadata can significantly enhance classification accuracy, achieving an F-measure score of 0.88. Their findings advocate for the broader use of metadata in document classification to improve retrieval effectiveness in academic databases.

Guo et al. (2020)

Guo et al. [56] introduced a hybrid approach that utilizes content-based and metadata-driven information to uncover connections between research papers. By leveraging metadata elements such as citation data, author information, and

publication dates, their study illustrated the importance of metadata in identifying complex interrelationships not readily apparent through traditional keyword searches alone.

In summary metadata-based techniques in academic research retrieval have evolved into sophisticated systems that integrate content analysis, citation networks, and advanced machine learning algorithms. These methods reveal complex connections between research works that go beyond simple keyword matching, aiding in the discovery of related literature and enhancing the comprehensiveness of academic research.

Researcher Sub-categories Metadata in these two major techniques.

- Co-Citation
- Bibliographic Coupling

Co-Citation

Small's Co-Citation Analysis (1973):
Small [25] introduced co-citation analysis as a method to discover related research papers. This method identifies connections between two papers based on how frequently other works cite them. Co-citation analysis helps unveil hidden relationships between documents that might not be obvious through regular citations. It is akin to finding documents that often go hand-in-hand with research, revealing connections that might not be explicit. Despite its usefulness, co-citation analysis comes with challenges. One notable challenge is its static nature.

Since it relies on past citations, co-citation analysis may not capture dynamic changes in the research landscape, particularly emerging trends or recent developments, which limits its ability to stay current with the most relevant papers. Additionally, it may overlook shifts in research focus or changes in the significance of specific works. Despite these limitations, co-citation analysis remains valuable, but researchers must be mindful of its static nature when interpreting results.

Gipp et al.'s Citation Proximity Analysis (2009)

Gipp et al. [8] introduced Citation Proximity Analysis (CPA) as an advanced form of co-citation analysis. CPA considers the proximity of citations within a document, providing a precise measure of relatedness than traditional co-citation. CPA can detect stronger thematic connections between documents by focusing on the distance between citations in the text. Despite its improved precision, CPA involves a considerable effort to process documents, and while it enhances traditional co-citation analysis, it does not entirely replace it.

Haruna et al.'s Citation-Based Recommender System (2018)

Haruna et al. [66] proposed a citation-based recommender system designed to help researchers navigate the vast amounts of scholarly information. Their approach utilizes the latent relations between citations and research papers to deliver personalized recommendations without necessitating extensive user profiles.

This innovation addresses critical limitations of existing systems, particularly the challenge of accessing full paper contents due to copyright restrictions. The authors demonstrated the effectiveness of their system through experimental results, which showed significant improvements in recommendation quality compared to baseline methods. By leveraging publicly available metadata, their work enhances accessibility and applicability across diverse research contexts, offering a practical solution to the issue of information overload in academia.

Chen et al.'s Dynamic Co-Citation Analysis (2021)

Chen et al. proposed a dynamic co-citation analysis approach to address the static nature of traditional co-citation methods. Their method can capture emerging research areas and recent developments by incorporating real-time citation data and analyzing trends. This approach improves the relevance of literature recommendations by continuously updating the co-citation network.

However, it requires significant computational resources and access to up-to-date citation databases, posing scalability and operational implementation challenges.

Shahid et al.'s In-Text Citation Frequency Analysis (2021)

Shahid et al. [67] developed a recommendation approach that leverages in-text citation frequencies to enhance the identification of relevant research papers.

Their method tracks how frequently and prominently references are mentioned within the citing paper's text, surpassing traditional techniques like bibliographic coupling and metadata analysis. Evaluated on a dataset of 1,200 documents from J.UCS, the approach demonstrated higher precision, with a 96% accuracy rate compared to the 76% achieved by content-based methods. This model delivers improved relevance but introduces challenges in extracting citations from PDFs (Portable Document Format).

Bibliographic Coupling

Kessler's Bibliographic Coupling (1963)

Kessler [6] laid the foundation for bibliographic coupling by introducing a method to identify related research papers based on their shared references. His approach posited that if two papers cite the same sources, they likely have a thematic connection. This insight was groundbreaking, providing an alternative to content-based methods, which primarily focused on analyzing the actual text of the documents. Kessler's method was relatively simple and computationally feasible, which made it appealing in an era when computational resources were limited. Unlike co-citation, which considers how often other works cite two papers, bibliographic coupling directly focuses on the overlap of references between papers. This made bibliographic coupling particularly useful for exploring the historical and thematic linkage between scientific papers, as it identified clusters of research built upon common intellectual foundations. However, while bibliographic coupling was a crucial development, it did not initially account for the dynamic evolution of citations over time, a limitation addressed by subsequent research.

Salton et al.'s Similarity Measures (1983)

Salton et al. [13] expanded on the concept of bibliographic coupling by introducing the overlap coefficient as a key metric for assessing the degree of similarity between

two papers. The overlap coefficient is calculated as the ratio of shared references to the total number of references in both papers. This coefficient provided a refined approach to measuring the extent of thematic overlap in academic literature, moving beyond the mere presence of shared citations to consider the relative size of the reference lists. This refinement was particularly useful for distinguishing between papers that might share references coincidentally and those with substantial thematic overlap. However, the overlap coefficient's accuracy could be compromised in cases where papers had very short or highly diverse reference lists, skewing the perceived relevance. Despite its limitations, the introduction of the overlap coefficient was significant, setting the stage for nuanced similarity measures in bibliographic coupling research.

Jaccard's Coefficient (1901)

While Jaccard originally introduced the Jaccard similarity coefficient citejaccard1901etude for botanical studies, it later found application in bibliographic coupling. The Jaccard coefficient offered a balanced measure by comparing the size of the intersection of shared references to the union of all unique references in both papers. This approach addressed some of the overlap coefficient's limitations by considering the commonalities and differences in the reference lists. Using the Jaccard coefficient provided a nuanced analysis, especially when dealing with papers with varying lengths of bibliographies. However, while the Jaccard coefficient improved on previous methods, it still did not fully account for the importance or thematic weight of the shared references. As a result, two papers could appear highly similar based on the number of shared citations without considering the depth or context of those citations within the papers' content.

Calado et al.'s Link-Based Measures (2003)

Calado et al. [68] explored how combining link-based and content-based methods can enhance web document classification. They evaluated measures derived from link structures like bibliographic coupling and co-citation and traditional classifiers like TF-IDF. Their findings indicated that link-based metrics improve classification, though the effectiveness varies based on the type of links analyzed. This research

highlighted the potential for hybrid approaches that leverage link structures and textual content, contributing to accurate and robust classification models.

Koseki et al.'s Coupling Strength (2007)

Koseki et al. [69] introduced the concept of coupling strength to refine bibliographic coupling analysis further. This method assigned weights to shared references based on their frequency of occurrence, assuming that frequently cited references indicate stronger thematic connections. By incorporating this weighting system, Koseki et al.'s approach allowed for a granular assessment of the significance of shared citations, thereby offering a robust indication of paper relatedness. However, this emphasis on citation frequency introduced a potential bias towards popular topics, which could overshadow niche but highly relevant research areas. For instance, fields with well-established foundational texts might dominate the analysis, potentially limiting the diversity of identified research connections. Despite this drawback, the coupling strength method marked an important step toward considering the influence of citation dynamics in bibliographic coupling.

Boyack et al.'s Evaluation of Coupling Methods (2010)

In a comprehensive study, Boyack et al. [28] compared various bibliometric techniques, including Bibliographic Coupling (BC), Co-Citation Analysis (CCA), Direct Citation (DC), and a Hybrid Approach (HYB), to understand their effectiveness in mapping biomedical research literature. Their evaluation provided insights into the strengths and weaknesses of different coupling methods. Co-citation analysis demonstrated the highest coverage, capturing 98.37% of articles. However, the hybrid approach (combining bibliographic coupling, co-citation, and direct citation) exhibited the highest textual coherence, resulting in tighter thematic clusters of related papers. Although bibliographic coupling showed strong coherence, it was less comprehensive than the hybrid method. The hybrid approach, while highly effective, came with significant computational costs. In contrast, bibliographic coupling offered a efficient analysis, striking a balance between coverage and coherence. This study underscored the importance of method selection based on the specific objectives of the bibliometric analysis.

Habib and Afzal's Enhanced Coupling (2017)

Habib and Afzal [70] proposed an enhanced bibliographic coupling technique incorporating Citation Proximity Analysis (CPA) to identify related papers accurately. CPA analyzes the proximity of in-text citations within a document, offering a deeper understanding of how closely related the referenced papers are in the context of the citing paper's arguments. They employed a density-based clustering algorithm (DBSCAN) to facilitate this analysis to group papers based on their citation proximity patterns. Their approach achieved a 55% accuracy rate, significantly outperforming content similarity approaches (20% accuracy) and traditional bibliographic coupling (45% accuracy). Despite its success, the method faced challenges related to the availability and quality of citation data, as well as the computational intensity of the clustering process. This study marked an important advancement in the field by emphasizing the contextual importance of in-text citations for bibliographic coupling.

Khan et al.'s Section-Wise In-Text Citation Score (2019)

Khan et al. [9] introduced the Section-Wise In-Text Citation Score (SwICS) technique to improve bibliographic coupling by analyzing where and how frequently papers cite each other across different sections. Recognizing that citations in sections like the methodology or results often carry thematic weight than those in the introduction or background, SwICS assesses the contextual importance of citations within these sections. SwICS provided a detailed understanding of paper relatedness by assigning section-specific weights to in-text citations. Their evaluation showed that SwICS matched user-perceived paper similarity 73% of the time, significantly outperforming older methods, which had a 45% match rate. This advancement underscored the need for a nuanced consideration of citation context in bibliographic coupling.

Habib's Citation Contextualization (2019)

Building on his earlier work, Habib [3] proposed a method that focused on where references are cited within a paper's structure (e.g., introduction, discussion) to contextualize citations for bibliographic coupling. This approach aimed to discern the importance of citations within different sections, providing a refined measure

of similarity between papers. The technique demonstrated an 8.5% improvement in relevance over content-based methods and a 2.7% improvement over traditional bibliographic coupling. However, it required careful adjustment to suit different research fields and was somewhat limited by its dependence on specific digital library datasets, highlighting the need for further research in adapting citation contextualization to diverse academic domains.

Bordons et al.'s Data Integration Challenge (2019)

Bordons et al. [71] explored the challenges of integrating diverse data sources in bibliographic coupling. They emphasized the need for careful data merging to ensure compatibility, particularly when incorporating multiple citation databases with varying coverage and formats. Their study suggested that hybrid approaches combining bibliographic coupling, content analysis, and user preferences could enhance recommendation accuracy. However, they also underscored the importance of addressing data integration challenges, such as ensuring data consistency, completeness, and quality, to realize the full potential of these advanced bibliometric methods.

Zhang et al.'s Temporal Analysis in Bibliographic Coupling (2021)

Zhang et al. [72] introduced temporal analysis into bibliographic coupling to account for the dynamic evolution of research themes over time. Traditional bibliographic coupling methods often treated the citation network as static, ignoring changes in research trends. By analyzing the temporal aspects of citation networks, Zhang et al. captured the ebb and flow of research topics, allowing for a accurate recommendation of recent and emerging research areas. This approach provided a dynamic perspective on paper relatedness, reflecting the shifting landscape of academic research. Despite its innovative nature, temporal analysis presented challenges in real-time data processing and required sophisticated algorithms to handle the complexity of evolving citation networks.

Yun's Generalization of Bibliographic Coupling and Co-citation (2022)

Yun [73] proposed novel methods for estimating bibliographic coupling (BC) and co-citation (CC) using a node split network approach. This method allows for the

efficient emulation of citation-based coupling measures without the computationally expensive direct calculations typically required for large datasets. By splitting nodes into citing and cited roles, Yun's approach utilizes Personalized PageRank (PPR) and neural embedding (EMB) techniques to capture similarities between documents, even in cases where direct citation links are absent.

The findings indicate that PPR can accurately estimate similarities by analyzing paths between nodes, thus uncovering long-range connections often overlooked in traditional measures. Furthermore, the study highlights that many links with high similarity are missing in conventional BC and CC networks, suggesting the necessity of considering long-range similarities. Yun's research not only refines the methodologies for research paper recommendation systems but also enhances our understanding of citation dynamics by addressing the limitations of existing citation-based measures.

Kanwal and Amjad (2024)

Kanwal and Amjad [74] proposed a novel research paper recommendation system named RRMF, which integrates multiple citation and collaboration network features. Their study addresses the challenge researchers face in finding relevant scholarly papers amid the increasing volume of publications. By constructing a multi-level citation network, RRMF identifies structural and semantic relationships within the citation network while extracting key authors from the collaboration network.

The authors utilized the AMiner v12 DBLP-Citation Network for experimentation, which includes over 4.8 million academic papers and 45.5 million citation relationships. The performance of RRMF was evaluated using standard information retrieval metrics, including Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG).

The results demonstrated that RRMF outperformed baseline approaches, such as the Multilevel Simultaneous Citation Network (MSCN) and Google Scholar, achieving up to 87% better recommendation accuracy. This work emphasizes the importance of combining citation analysis with collaboration networks to enhance the quality and relevance of recommendations. The findings suggest

that incorporating multiple features can significantly improve the performance of research paper recommender systems, making it a valuable contribution to information retrieval.

Conclusion

The bibliographic coupling has evolved through various stages, from shared reference counts to complex, context-aware, and AI-augmented models. Each iteration has addressed specific limitations of earlier techniques, contributing to a refined understanding of research paper relationships. While content-based and metadata-based approaches offer valuable insights, the ongoing integration of machine learning, natural language processing, and dynamic data analysis in bibliographic coupling represents a promising direction for future research in academic information retrieval.

2.6 Strengths and Limitations of Content-Based and Metadata-Based Approches

The ever-expanding volume of scientific literature necessitates sophisticated document recommendation systems to assist researchers in finding relevant works. These systems are broadly categorized into content-based and metadata-based approaches, each leveraging different data aspects to provide recommendations.

2.6.1 Advantages of Content-Based Approaches

Content-based recommendation systems offer a powerful method for suggesting relevant literature by analyzing the textual content of documents. This approach has several notable advantages:

- **Direct Relevance:** Content-based systems ensure that recommendations are directly aligned with the user's research interests or the thematic content of the query document. These systems provide highly personalized and

relevant suggestions by thoroughly analyzing the text of documents. This direct relevance is particularly beneficial for researchers seeking literature that closely matches their study area, enhancing the overall research experience.

- **Independence from Citation Data:** One of the significant strengths of content-based recommendation systems is their independence from citation networks. Unlike citation-based systems, which rely on the frequency and patterns of citations to determine the relevance of documents, content-based systems focus solely on the textual content.
  This makes them particularly useful for recommending newer publications that may not have accumulated many citations yet but are highly relevant to current research topics. By not depending on citation data, these systems can identify and highlight cutting-edge research and emerging trends early on [75].

- **Detailed Analysis:** Content-based approaches employ sophisticated natural language processing (NLP) techniques to analyze documents deeply. These techniques enable the system to identify specific themes, methodologies, and results, allowing for nuanced recommendations that go beyond superficial matching [76]. This detailed analysis ensures that the suggested literature provides a comprehensive and in-depth understanding of the topic.

- **Customizable Filtering:** Users can tailor recommendations to their precise needs by setting filters based on desired topics, research methods, or the presence of specific keywords. This level of customization enhances the system's utility and flexibility, enabling researchers to focus on literature that meets their exact requirements [77]. Customizable filtering ensures the recommendations are relevant and aligned with the user's research criteria.

2.6.2 Issues with Content-Based Approaches

Despite their strengths, content-based approaches encounter several challenges:

- **Limited Analysis:** Focusing primarily on textual content, these systems may overlook important non-textual elements such as images, tables, and the document's format, which can contain critical information for some disciplines.

- **Vocabulary Mismatch:** The effectiveness of content-based recommendations can be hindered by terminology differences across research fields. Synonyms and domain-specific jargon may cause relevant documents to be missed because the system fails to recognize them as related [78].

- **Scalability Issues:** The computational demands of processing and analyzing large volumes of text in real-time can challenge scalability and responsiveness, especially as the available literature continues to grow [13].

- **Lack of Contextual Insight:** Content-based systems might not adequately consider the broader academic impact or the context within which a paper was published, such as its reception or the reputation of its authors, potentially overlooking influential but thematically divergent works [79].

2.6.3 Advantages of Metadata-Based Approaches

In contrast, metadata-based approaches rely on structured data about documents for recommendations:

- **Precise Document Matching:** By leveraging explicit details like authors, publication dates, and keywords, metadata-based systems can quickly identify documents with a high degree of topical or authorial relevance, providing accurate recommendations with minimal processing overhead.

- **Quick Filtering and Identification:** Metadata's structured nature allows for efficient sorting and matching, enabling users to rapidly find documents that meet specific criteria, such as being from a particular publication or written by a specific author.

- **Simpler Implementation:** Unlike content-based systems that require complex NLP algorithms, metadata-based approaches can be straightforward to implement, as they primarily involve database queries and matching algorithms.
- **Easy Integration with Digital Libraries:** Metadata standards like Dublin Core facilitate the integration of these systems with digital libraries and repositories, making it easier to access and recommend documents across different platforms.

2.6.4 Issues with Metadata-Based Approaches

However, the metadata-based systems also face distinct challenges:

- **Content Nuances:** This approach may overlook the nuances of document content, such as the argument's quality or novelty, as recommendations are based on descriptive rather than substantive data.
- **Metadata Quality:** The effectiveness of metadata-based recommendations is contingent on the accuracy and completeness of the metadata. Inconsistent or sparse metadata can lead to poor recommendation quality [80].
- **Missed Content-Based Relationships:** By focusing on explicit metadata, these systems may miss deeper thematic or methodological connections between documents that are not readily apparent from metadata alone.
- **Overlooked Implicit Connections:** Important but implicit connections, such as those based on the evolution of ideas or interdisciplinary relevance, might be overlooked, limiting the breadth of recommendations.

2.6.5 Collaborative Filtering Approaches

Collaborative filtering (CF) approaches recommend documents by analyzing users' preferences and behaviors regarding items (e.g., documents).

- **Advantages:** CF methods excel at discovering user preferences and predicting items that similar users have liked, offering personalized recommendations based on user interaction patterns. This approach can uncover unexpected recommendations beyond content similarity, enhancing discovery [81].

- **Challenges:** The "cold start" problem is significant in CF systems, where new items or users with few interactions are difficult to recommend accurately. Additionally, CF methods can suffer scalability issues as the user-item matrix grows, and they may not effectively capture the content's novelty or diversity [82].

2.6.6 Graph-Based Approaches

Graph-based approaches utilize the structure of networks, where documents and other entities (e.g., users, keywords) are nodes, and their relationships are edges, to recommend documents.

- **Advantages:** These methods are adept at capturing complex relationships between items, allowing for nuanced recommendations considering the interconnectedness of documents, users, and metadata. Graph algorithms can identify influential nodes or clusters within the network, providing insights into community preferences or emerging trends [83].

- **Challenges:** Constructing and maintaining large-scale graphs can be computationally intensive, particularly as the number of documents and relationships grows. Additionally, graph-based methods may require sophisticated algorithms to navigate and interpret the network's complexity effectively [84].

2.6.7 Data Mining-Based Approaches

Data mining-based approaches employ algorithms to uncover hidden patterns and relationships in large datasets, including text corpora.

- **Advantages:** These methods are powerful in detecting latent patterns, trends, and associations within the data, facilitating the recommendation of documents based on deep content analysis and user interaction histories. Data mining can enhance the accuracy and relevance of recommendations by leveraging classification, clustering, and association rule mining [85].

- **Challenges:** The scalability and computational efficiency of data mining-based approaches can be problematic, especially with the continuous growth of digital content. Additionally, these methods may require extensive preprocessing to transform raw data into a suitable format for analysis [86].

2.6.8 Citation-Based Approaches

Citation-based approaches recommend documents by analyzing citation networks, where citations among papers indicate relationships and potential relevance.

- **Advantages:** Citation analysis can reveal the impact and significance of documents within a research field, providing recommendations based on scholarly influence and thematic connections indicated by citation patterns. These methods benefit from the structured nature of citation data, enabling the identification of seminal works and emerging research fronts [87]. Citation-based approaches can effectively highlight the most influential papers, ensuring users are exposed to high-impact research. By leveraging the citation network, these approaches can also uncover the connections between different research areas, promoting interdisciplinary discoveries and a comprehensive understanding of a topic.

- **Challenges:** Relying solely on citation data may overlook newer, less-cited papers that are nonetheless relevant. Citation-based methods can also be biased towards older, established works, potentially stifling the discovery of innovative research. Furthermore, citation motivations can vary, and not all citations indicate positive relevance [88]. The delay in citation accumulation means that cutting-edge research might not be immediately recommended,

hindering the dissemination of novel ideas. Additionally, citation-based approaches may inadvertently reinforce the dominance of well-established research groups or institutions, potentially marginalizing emerging scholars or less mainstream areas of study.

Both content-based and metadata-based approaches offer valuable tools for document recommendation but face unique challenges. A hybrid approach that combines the depth of content analysis with the efficiency and scalability of metadata-based filtering could offer a comprehensive solution, addressing the limitations inherent in each method while leveraging their strengths. For instance, integrating citation analysis with content-based filtering can provide a balanced recommendation system that values documents' influence and content relevance. Such hybrid systems can dynamically adapt to users' needs, offering personalized recommendations that evolve with the research landscape.

TABLE 2.1: Strengths and Limitations of Scientific Paper Recommendation Techniques

Approach	Methodology	Strengths	Limitations
Citation Based	Uses bibliographic methods such as co-citation and bibliographic coupling to establish relationships among papers.	Leverages peer-reviewed literature [89]. Provides insights into intellectual structures [90].	Subject to delays in emerging field detection [91]. Prone to self-citation manipulation [92].
Content Based	Analyzes the text content of papers using NLP, extracting topics, abstracts, and keywords for recommendations.	Improves content relevance and precision [93]. Unveils nuanced thematic relationships beyond citations.	High computational cost for processing text data [94].
User Profile Based	Generates recommendations by analyzing users' interactions, such as searches and downloads, to match papers with individual preferences.	Highly personalized recommendations [95]. Enhances user engagement with adaptive suggestions.	Raises privacy concerns due to tracking user behavior [95]. Requires frequent updates to reflect changing user interests.

Continued on next page

Table 2.1 – *Continued from previous page*

Approach	Methodology	Strengths	Limitations
Collaborative Filtering	Uses behavior patterns of multiple users to suggest papers, assuming similar users will have shared interests.	Learns from community preferences for dynamic adaptation [96].	Struggles with cold-start issues when data is sparse [79]. Can prioritize popular over niche topics.
Data Mining Based	Applies clustering, classification, and association rule mining to uncover patterns within bibliographic datasets.	Discovers latent themes and new research trends [97]. Scalable to handle large datasets [94].	Requires extensive preprocessing [97]. Models can be complex to interpret.
Metadata Based	Utilizes metadata fields such as author, venue, publication year, and keywords to recommend relevant papers.	Fast retrieval due to indexed metadata. Effective in filtering papers by specific attributes.	Limited by the quality and completeness of metadata. Struggles to capture the full content relevance.
Hybrid Methods	Combines content-based and collaborative filtering techniques to enhance recommendations.	Balances precision and diversity in recommendations [76].	Hybrid models are computationally intensive and complex to implement.

2.6.9 Dataset for Related Papers

In facilitating experimental evaluations, attention is directed toward two meticulously chosen datasets. The first, Dataset - 1[46], is a manually annotated dataset tailored for experimentation. Concurrently, Dataset - 2[3], as utilized in the bibliographical approach presented by Habib and Afzal [3], aligns seamlessly with the requisites of our research, furnishing a solid foundation for result comparison.

Manually Annotated Dataset (Dataset - 1)

To ensure the robustness and accuracy of our experiments, acquiring a meticulously annotated dataset is paramount. Kanakia et al. [46] present an invaluable resource in the Microsoft Knowledge Graph dataset, manually annotated through a rigorous user study. This dataset is a foundational component for our research, offering a comprehensive collection of 2400 recommendation pairs. The annotation process involved the active participation of 40 individuals, providing diverse and nuanced perspectives. The meticulous grading by participants adds depth to the dataset, making it a reliable source for evaluating section weights in alignment with our research objectives. The dataset is openly accessible on GitHub at `https://github.com/akanakia/microsoft-academic-paper-recommender-user-study`, ensuring transparency and facilitating reproducibility in our experiments. By leveraging this manually annotated dataset, the research aims to contribute to a thorough comparative analysis, particularly in juxtaposition with the methodology introduced by Habib et al. [3].

Previous Research Dataset - (Dataset - 2)

The dataset by Habib and Afzal [3] employed in the study was utilized to enhance the bibliographical approach. This dataset, referred to as "Dataset-2," possesses the characteristics required for our research and offers a basis for result comparison. A robust and extensive dataset was crucial to evaluating proposed methodologies comprehensively.Dataset-2, characterized by its breadth and depth, encompasses a

collection of 5,000 papers interconnected through bibliographic coupling, spanning various academic fields. To ensure a rich and varied data pool for our evaluation, the research meticulously crafted a set of 17 queries. These queries were designed to cover a broad spectrum of research interests, including but not limited to social network analysis, information retrieval techniques, Bayesian networks, feature selection methods, collaborative and recommendation systems, content-based filtering, software testing methods like black box testing, automatic content generation, regression testing, query processing strategies, advancements in sensor and wireless networks, opinion mining and subjectivity analysis, the dynamics of online marketing, and graph theory applications. This strategic selection of topics allowed for a comprehensive assessment of our system's capabilities across diverse research domains.

2.7 Discussion

This literature review highlights the intricate and evolving landscape of research paper discovery techniques, emphasizing the interplay between content-based, metadata-based, and hybrid approaches. Content-based strategies, such as keyphrase extraction and semantic analysis, provide personalized and nuanced recommendations by analyzing textual features, including titles, abstracts, and keywords. However, these techniques encounter limitations in scalability and domain-specific vocabulary mismatches, which hinder effective information retrieval across interdisciplinary fields [51, 78]. The need to refine these methods further is evident, particularly given the rapid growth of academic publications that demand sophisticated, scalable solutions [76].

Metadata-based approaches, such as co-citation analysis and bibliographic coupling, offer structured, efficient filtering through citation networks, author metadata, and publication dates. These methods excel at identifying influential papers within established research areas, but they face challenges in recognizing emerging research topics and handling sparse citation data for recent publications [87, 90]. Notably, bibliographic coupling—first introduced by Kessler [90]—has evolved significantly,

with modern advancements incorporating section-based weighting and in-text citation analysis to enhance the precision of related paper discovery [9].

The review also underscores the importance of hybrid approaches that integrate content-based and metadata-driven techniques. For example, combining bibliographic coupling with semantic filtering demonstrates improved recommendation accuracy by leveraging shared references and thematic similarity [3]. These hybrid models overcome individual limitations by aligning content with citation patterns, offering a dynamic and adaptable framework for identifying related research. The use of real-time citation networks, such as those proposed by Chen et al. [98], further highlights the potential of dynamic approaches to keep pace with the evolving research landscape.

In addition, collaborative filtering and graph-based techniques extend the scope of recommendation systems beyond traditional content and metadata analysis. By utilizing user behavior patterns and network structures, these methods can uncover implicit relationships and recommend interdisciplinary works that may not be evident through conventional approaches [81, 84]. However, these techniques face the cold-start problem, especially for new users and items, and require sophisticated algorithms to process large datasets efficiently [82].

Recent research has also introduced advanced metrics, such as the Section-Wise In-Text Citation Score (SwICS), to refine bibliographic coupling by assigning dynamic section weights based on their relevance [9]. This approach highlights the growing importance of contextual information in citation networks, underscoring the need for nuanced analysis methods. Additionally, integrating machine learning and natural language processing techniques in recommendation systems offers new avenues for improving recommendation quality and relevance [47, 49].

In conclusion, this review reveals that the most effective solutions for research paper discovery lie in hybrid approaches combining content-based, metadata-based, and collaborative filtering techniques. Future research should focus on developing adaptive models that dynamically integrate multiple data sources and account for evolving research trends. Furthermore, improving dataset quality and ensuring

comprehensive validation across different academic disciplines will be essential for advancing the state of the art in research paper recommendations. This multifaceted approach will enable researchers to navigate the expanding academic literature effectively, promoting interdisciplinary discoveries and fostering scientific innovation.

2.8 Summary

The literature review chapter is crucial and challenging, as it aims to discover related research papers amidst the rapid growth of online publications. The review explores methods for identifying relevant sections within research papers, investigates techniques for finding related papers, examines the use of bibliographic information, explores new approaches to enhance bibliographic coupling techniques, and analyzes existing datasets for the most suitable collection of associated papers.

A literature survey focused on methods like bibliographic coupling, co-citation analysis, and text-based approaches. This review identified the strengths and weaknesses of different techniques for finding related papers and selected the most notable techniques based on their effectiveness and relevance. This included carefully analyzing seminal works and adapting the latest techniques to review influential papers.

The chapter categorizes the hundreds of paper recommendation approaches identified by Beel et al. into content-based strategies, metadata-based approaches, collaborative filtering-based approaches, user profile-based approaches, and data mining approaches. Each category encompasses various methods to enhance the discovery of scholarly articles, with unique perspectives and methods contributing to the richness of the research paper recommendation landscape. Content-based strategies are discussed using keyword-based, concept-based, graph-based, and hybrid approaches. Each technique has its methodology, strengths, and limitations, ranging from precise relevance and independence from citation data to challenges

in accurately capturing and interpreting scientific texts and scalability issues. Collaborative filtering is highlighted as a technique that suggests items by examining users' preferences with similar tastes. It is applicable across various domains, such as e-commerce, movies, and research paper recommendations. Despite its utility, challenges such as the 'cold start' problem and the necessity for comprehensive data processing are noted.User profile-based approaches leverage digital library access logs and user profiles to recommend related scientific papers, focusing on personalizing recommendations based on individual research needs and interests. However, privacy concerns and the need for continuous updating to accurately reflect user interests are mentioned as limitations.

Data mining techniques are presented as revolutionary methods for generating scientific paper recommendations. They offer personalized and relevant suggestions based on user preferences and behaviors. Challenges include the need for significant preprocessing and the complexity of models.

Techniques based on metadata are discussed for their role in uncovering connections between research papers. These techniques utilize titles, summaries, and references to facilitate the discovery of related works. The chapter concludes by exploring techniques for extracting sections, mapping them to the IMRaD logical structure, and reviewing various approaches to identifying related research papers. Recent studies demonstrate promising results when using section-wise bibliographic coupling, underscoring the critical role that sections of research papers play in improving the accuracy of bibliographic coupling.

In summary, this literature review chapter comprehensively analyzes existing methods for discovering related research papers and evaluating their methodologies, strengths, and limitations.

Chapter 3

Extraction of Section-Headings and Mapping to IMRaD Structure

3.1 Overview

The chapter explains the process of extracting section headings from research articles and mapping them to the widely recognized IMRaD structure, which stands for Introduction, Methods, Results, and Discussion. Introducing an intelligent system designed to accurately detect section boundaries, categorize them correctly, and map them to the appropriate IMRaD structure. Finally, the results will be discussed and compared with the existing research.

The answer to this question needs the answers to the following three questions:

RQ1: How can a method be devised with improved accuracy to map the sections of research papers to the IMRaD structure, considering the variations of these sections?

RQ1a-1: What potential features can be used to increase the accuracy of section mapping onto the IMRaD structure?

RQ1a-2: How can an intelligent algorithm be devised to identify the section, its sub-sections, and its boundaries?

RQ1a-3: How the wrongly identified subsections in the content file can be auto-corrected.

The proposed technique expands the research carried out by Shahid et al. [2]. Upon reviewing the literature, it was observed that their approach lacked the use of certain features that could enhance the accuracy of identifying and mapping sections to the IMRaD structure, thereby improving precision and recall of Habib and Afzal [3]. This research has significant implications for citation indexes and digital libraries.

3.2 Introduction

Research articles are typically organized into well-defined sections to present scientific work logically. One of the most widely used organizational structures is IMRaD, which divides research papers into four major sections: **Introduction**, **Methods**, **Results**, and **Discussion**. The IMRaD structure provides a clear framework for communicating scientific findings, enabling readers to easily understand the problem, methodology, outcomes, and conclusions.

Despite the widespread adoption of IMRaD, research papers often exhibit substantial variability in their structure and formatting. Different journals, conferences, and disciplines may introduce variations in naming conventions and the organization of sections. Such inconsistencies present significant challenges for automated systems attempting to map sections accurately to the IMRaD structure. Additionally, academic papers may contain subsections, figures, tables, and citations distributed unevenly across sections, further complicating the task.

This work proposes a robust system to extract and map section headings from research articles into the IMRaD framework, addressing the challenges posed by structural variability. The system identifies primary and secondary sections with high precision by leveraging advanced techniques, including XML parsing, regular

expressions, and natural language processing. Furthermore, the system incorporates additional discriminative features, such as the frequency of figures, tables, and citations, to enhance the mapping process. These features provide critical insights into the logical flow of documents and ensure accurate section classification.

The proposed methodology integrates four key modules: the Schema Generation Engine (SGE), the Data Extraction Engine (DEE), the Data Mapping Engine (DME), and the Mapping View Engine (MVE). Each module plays a vital role, from parsing the structure of research articles and mapping sections to IMRaD to visualizing the results effectively. The process begins with converting research papers into XML format, followed by the systematic extraction and classification of sections using a relational database schema for structured storage.

This research builds upon prior studies but overcomes limitations in existing approaches that often fail to recognize the hierarchical relationship between headings and subheadings. For instance, earlier work did not account for the nuances of subsections, leading to misclassification and false mappings. Our system addresses these issues by employing regular expressions to distinguish between main and sub-sections, preserving the logical integrity of the research document.

The proposed system's effectiveness is validated through rigorous performance evaluations using a benchmark dataset, comparing it against state-of-the-art techniques. The results demonstrate significant improvements in precision, recall, and F-measure, establishing the proposed system as a reliable solution for section extraction and IMRaD mapping. These advancements offer practical applications for citation indexing, digital libraries, and information retrieval systems, facilitating efficient navigation and understanding of scientific literature.

3.2.1 Features for Section Extraction and IMRaD Mapping

At the onset of this chapter, the exploration reveals potential features that could enhance the accuracy of section mapping.

- Subheadings Mapping
- Object Counts
- In-Text Citations Frequency

Subheadings Mapping

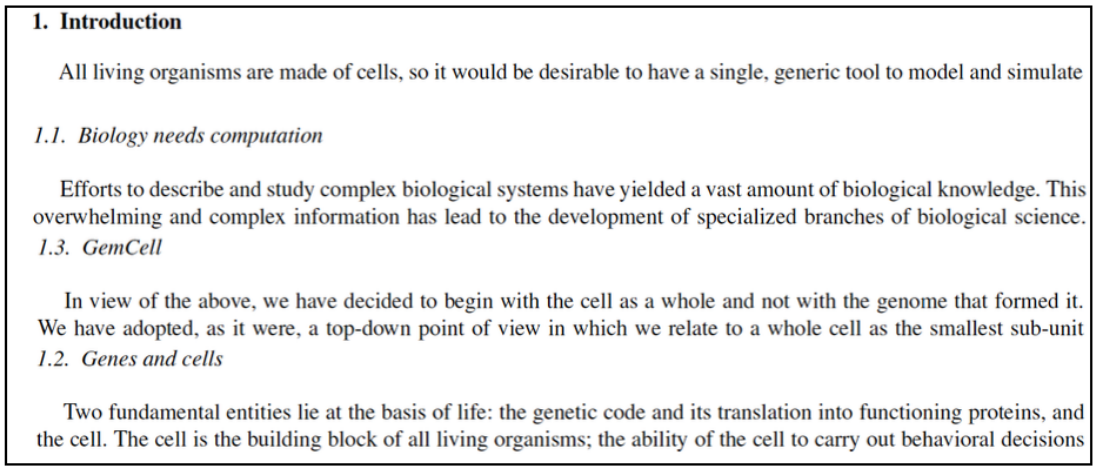
Headings in research papers organize the content into major sections, such as Introduction, Methods, Results, and Discussion. Subheadings are smaller titles within these sections that further divide the content into specific topics or themes, aiding in navigation and comprehension of the paper's structure and content. Therefore, correct subheading identification plays an important role in identifying the heading.

In the PDF file depicted in Figure 3.1, the Introduction section of a research paper titled "1. Introduction" encompasses three subsections: "1.1. Biology needs computation," "1.2. Genes and cells," and "1.3. GemCell." In the corresponding XML file shown in Figure 3.3, the main heading (i.e., 1. Introduction) is represented with the `<h1>` tag, while all subheadings are marked with the `<h2>` tag. The content within the brackets of the heading tag is considered the name of an independent logical section. Notably, the existing approach [2] fails to differentiate between a paper's main section and subsections, treating them all as independent headings. This deficiency leads to a significant issue, which is explained below.

Ding et al. [1] have previously utilized the heading `<h1>` and `<h2>`. However, in Figure 3.2, there is a sub-section named "1.3 Related Work." As per the IMRaD structure, this section would be considered independent and mapped to the "Related Work" section. Unfortunately, the section does not belong to the literature review section of IMRaD. Such challenges result in false mapping, compromising the performance of Information Retrieval (IR) systems. A closer examination revealed that most headings start with a bullet number, like "1..." for main headings and "1.1.." "1.2.." "1.3.." and "1.4.." for subheadings. The inability to

differentiate the logical structure through XML can be addressed by employing regular expressions to distinguish sections based on these patterns intelligently.

In summary, the examples discussed above illustrate that the existing research in logical section mapping struggles to map subsections intelligently. Treating all subsections as independent sections and explicitly mapping them to the IMRaD structure can negatively impact the overall precision of IR systems. Our study overcomes these challenges by implicitly using regular expressions to map subsections to their respective main sections. As mentioned earlier, our proposed study incorporates potential features such as In-Text citation count, Figure count, and Table count to determine the association of a section with specific logical sections of IMRaD. The rationale for leveraging these parameters is provided below.



1. Introduction

All living organisms are made of cells, so it would be desirable to have a single, generic tool to model and simulate

1.1. Biology needs computation

Efforts to describe and study complex biological systems have yielded a vast amount of biological knowledge. This overwhelming and complex information has lead to the development of specialized branches of biological science.

1.3. GemCell

In view of the above, we have decided to begin with the cell as a whole and not with the genome that formed it. We have adopted, as it were, a top-down point of view in which we relate to a whole cell as the smallest sub-unit

1.2. Genes and cells

Two fundamental entities lie at the basis of life: the genetic code and its translation into functioning proteins, and the cell. The cell is the building block of all living organisms; the ability of the cell to carry out behavioral decisions

FIGURE 3.1: Subheading example in the form of PDF file.

```
<section class="DoCO:Section">
<h1 class="DoCO:SectionTitle" id="5" page="1" column="1">
1 Introduction 1.1 Skyline Queries </h1>

<h1 class="DoCO:SectionTitle" id="13" confidence="possible"
page="2" column="1"> 1.2 Online Skyline Computation </h1>

<h1 class="DoCO:SectionTitle" id="18" confidence="possible"
page="2" column="2"> 1.3 Related Work </h1>
```

FIGURE 3.2: Subheading Example-2

Objects Count

Scientific articles commonly convey results and findings using figures, tables, algorithms, and graphs, and their frequency often correlates with specific IMRaD logical sections. For instance, sections such as "Methodology" and "Results" are typically characterized by a higher occurrence of figures and tables, reflecting their emphasis on experimental setups and data visualization. However, contemporary approaches, such as those proposed by Shahid et al. [2], overlook the inclusion of "Figure count" and "Table count" as parameters for section identification.

```
<section class="deo:Introduction">
<h1 class="DoCO:SectionTitle" id="9" page="2" column="1">
1. Introduction </h1>

<h2 class="DoCO:SectionTitle" id="20" page="3" column="1">
1.1. Biology needs computation </h2>

<h2 class="DoCO:SectionTitle" id="40" page="3" column="1">
1.2. Genes and cells</h2>

<h2 class="DoCO:SectionTitle" id="44" page="3" column="1">
1.3. GemCell </h2>
```

FIGURE 3.3: Subheading Example-1

In this study, we address this limitation by incorporating these parameters, hypothesizing that the frequency of figures and tables can serve as reliable indicators of their association with specific IMRaD sections. Figures and tables are identified as distinct objects within XML-encoded documents, enabling systematic extraction and analysis. This integration enhances the accuracy of section mapping by leveraging the inherent relationship between object frequency and the logical structure of scientific articles.

```
<region class="DoCO:FigureBox" id="F1">
<caption class="deo:Caption" id="9" page="1" column="2">
Figure 1: Skyline of hotels in Nassau (Bahamas) </caption>
```

FIGURE 3.4: Figures in research publications.

```
<region class="DoCO:TableBox" id="Tx100">
<content>
    <table class="DoCO:Table" number="1" page="9">
    <thead class="table"/>
        <tbody>
            <tr class="table">
                <td class="table"></td>
                <td class="table"> 100,000 Points</td>
                <td class="table"> 1,000,000 Points</td>
            </tr>
        </tbody>
    </table>
</content>
```

FIGURE 3.5: Tables in research publications.

In-Text Citations Frequency

Similar to the assumptions made for "Figure Count" and "Table Count," the "In-Text Citation count" can be crucial in determining association with a specific logical section. Ding et al. [1] have used the frequency of in-text citations in all logical sections. Our manual investigation found varying in-text citation counts among sections. For example, the "Literature Review" section contains more in-text citations than others. Our approach maps the logical section with the highest in-text citations to the Literature Review section. Citations are also represented as objects in XML files.

Contemporary studies have overlooked the importance of potential features like "In-Text Citation count," "Figure count," 3.6, and "Table count." 3.5 These parameters, as explained above, can contribute to section identification. This study aims to address these deficiencies to improve system performance significantly.

3.3 System Architecture

This section details the proposed methodology, working in four modules to map logical sections of research articles to the IMRaD structure.

The modules include:

- Schema Generation Engine (SGE)
- Data Extraction Engine (DEE)
- Data Mapping Engine (DME)

The PDF file dataset is collected from CiteSeer and converted into XML using the PDFX [99] tool. SGE generates the schema of the XML files, which are maintained in PostgreSQL to parse and insert the XML data. DEE extracts headings, subheadings, and other objects like citations, figures, and tables. Mapping SQL Engine (MSE) maps the extracted headings and subheadings to IMRaD using a devised algorithm. The last module, Mapping View Engine (MVE), visualizes the resulting mapping using XPath/XQuery expressions. The mapped sections are evaluated using a benchmark dataset containing section annotations formed with the help of a user study. The overall structure of the proposed methodology is shown in Figure 3.7. The proposed algorithm is formally represented below, and detailed explanations of all implemented modules are delineated in the following sections.

```
<xref ref-type="bibr" rid="R1" id="21" class="deo:Reference">
1 </xref>,
<xref ref-type="bibr" rid="R2" hidden="1" id="22" class =
"deo:Reference" > 2
</xref>,
<xref ref-type="bibr" rid="R3" hidden="1" id="23" class =
"deo:Reference" > 3
</xref>,
```

FIGURE 3.6: Example of a figure caption.

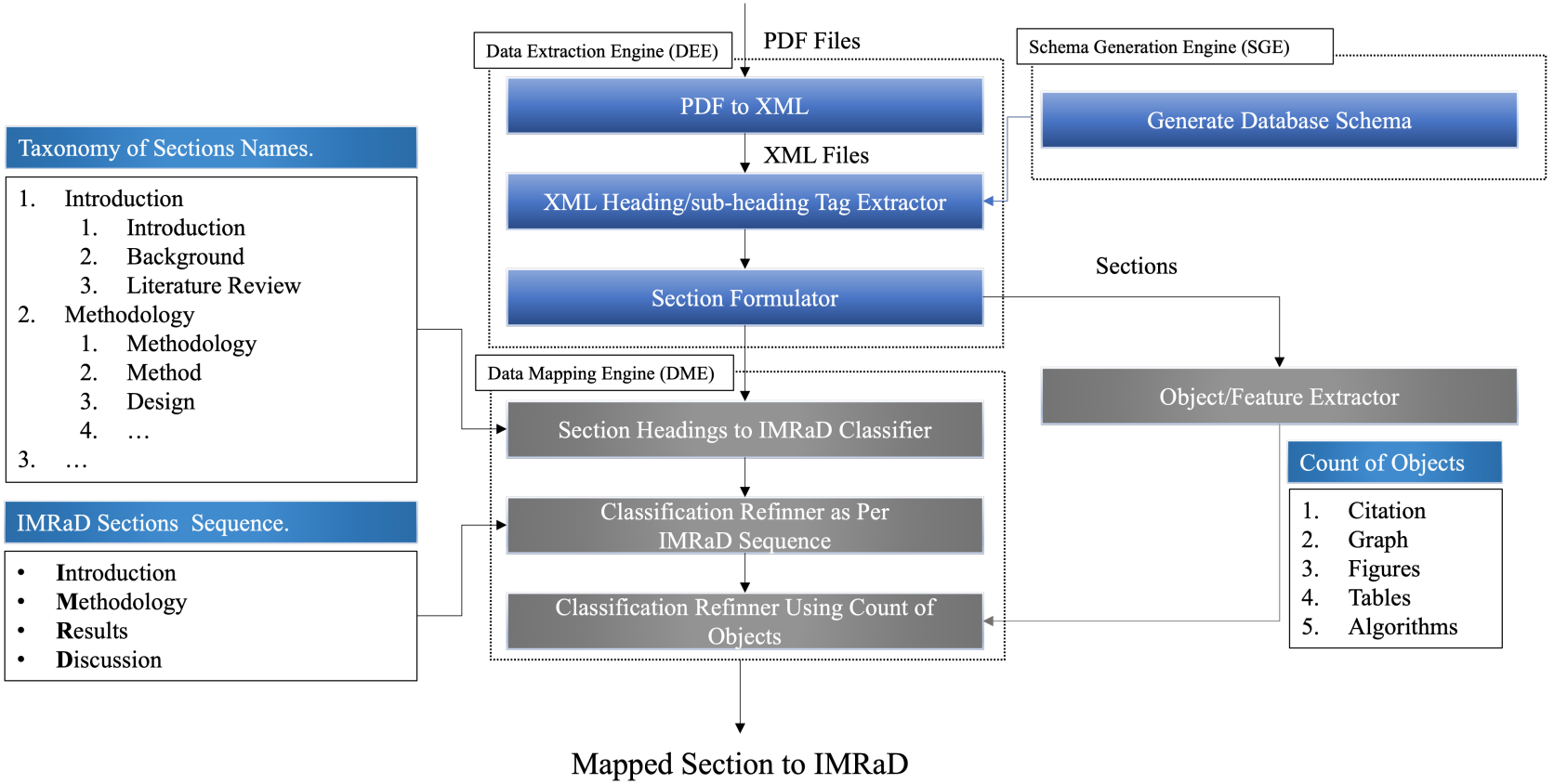


FIGURE 3.7: System Architecture

3.3.1 Schema Generation Engine (SGE)

The Schema Generation Engine (SGE) is an integral component in facilitating the transition of research papers from their native PDF format to a structured XML representation. This conversion is pivotal for the systematic processing and querying of the dataset. The primary function of the SGE is to architect and implement a database schema that accurately reflects the complex hierarchical and relational aspects of research publications. Utilizing PostgreSQL, known for its extensive XML data manipulation capabilities, the SGE maps the elaborate structure of research papers into a relational database schema. This schema allows for organized XML data storage and simplifies data manipulation and retrieval processes, significantly enhancing data handling efficiency.

The **purpose** of the **SGE** is to define the structure for storing essential elements of research documents, including metadata, sections, figures, tables, and citations. This relational database schema ensures that extracted information is stored systematically for efficient retrieval and manipulation, enhancing the precision of downstream processing tasks.

Schema Design and Table Definitions

The following tables represent the core structure of the schema used to store and manage research documents and their components.

```
CREATE TABLE documents (
    document_id SERIAL PRIMARY KEY,
    title VARCHAR(255) NOT NULL,
    author VARCHAR(255),
    publication_date DATE,
    abstract TEXT,
    keywords TEXT,
    file_path VARCHAR(255)
);
```

FIGURE 3.8: Table Contain the Paper Information

The table in figure 3.8 holds metadata for each document's title, author, and publication date. The primary key `document_id` ensures unique identification.

```
CREATE TABLE sections (
    section_id SERIAL PRIMARY KEY,
    document_id INT NOT NULL,
    section_title VARCHAR(255),
    section_level INT,
    imrad_type VARCHAR(50),
    start_page INT,
    end_page INT,
    FOREIGN KEY (document_id)
      REFERENCES documents(document_id) ON DELETE CASCADE
);
```

FIGURE 3.9: Table Contains the Sections

The table in figure 3.9 stores information about each section within a document. It uses a foreign key to link each section to its corresponding document, ensuring referential integrity.

```
CREATE TABLE section_mappings (
    mapping_id SERIAL PRIMARY KEY,
    imrad_type VARCHAR(50) NOT NULL,
    confidence_score FLOAT
);
```

FIGURE 3.10: Table Contains Section Mapping

The table in figure 3.10 records the mapping of sections to IMRaD categories, along with a confidence score for the classification.

The table in figure 3.11 allows many-to-many relationships between mappings and sections.

The table in figure 3.12 captures metadata for figures, linking them to their sections and documents.

The table in figure 3.13 stores information about tables within documents, ensuring each entry is linked to its corresponding section and document.

```
CREATE TABLE section_mapping_sections (
    mapping_id INT NOT NULL,
    section_id INT NOT NULL,
    PRIMARY KEY (mapping_id, section_id),
    FOREIGN KEY (mapping_id)
        REFERENCES section_mappings(mapping_id) ON DELETE CASCADE,
    FOREIGN KEY (section_id)
        REFERENCES sections(section_id) ON DELETE CASCADE
);
```

FIGURE 3.11: Table Contain Section Mapping Relations

```
CREATE TABLE figures (
    figure_id SERIAL PRIMARY KEY,
    document_id INT NOT NULL,
    section_id INT,
    caption TEXT,
    page_number INT,
    FOREIGN KEY (document_id)
      REFERENCES documents(document_id) ON DELETE CASCADE,
    FOREIGN KEY (section_id)
      REFERENCES sections(section_id) ON DELETE CASCADE
);
```

FIGURE 3.12: Table Contains the Figures Information

```
CREATE TABLE tables (
    table_id SERIAL PRIMARY KEY,
    document_id INT NOT NULL,
    section_id INT,
    caption TEXT,
    page_number INT,
    FOREIGN KEY (document_id)
      REFERENCES documents(document_id) ON DELETE CASCADE,
    FOREIGN KEY (section_id)
      REFERENCES sections(section_id) ON DELETE CASCADE
);
```

FIGURE 3.13: Table Contains the Table Information

```
CREATE TABLE citations (
    citation_id SERIAL PRIMARY KEY,
    document_id INT NOT NULL,
    section_id INT,
    citation_text TEXT,
    page_number INT,
    FOREIGN KEY (document_id)
      REFERENCES documents(document_id) ON DELETE CASCADE,
    FOREIGN KEY (section_id)
      REFERENCES sections(section_id) ON DELETE CASCADE
);
```

FIGURE 3.14: Table Contain the Citaion Information

The table in figure 3.14 captures references within the document, linking them to relevant sections.

```
CREATE TABLE algorithms (
    algorithm_id SERIAL PRIMARY KEY,
    document_id INT NOT NULL,
    section_id INT,
    algorithm_name VARCHAR(255),
    description TEXT,
    page_number INT,
    FOREIGN KEY (document_id)
      REFERENCES documents(document_id) ON DELETE CASCADE,
    FOREIGN KEY (section_id)
      REFERENCES sections(section_id) ON DELETE CASCADE
);
```

FIGURE 3.15: Table Contains the Algorithms Information

The table in figure 3.15 stores metadata for algorithms, linking them to their respective sections and documents.

The table in figure 3.16 stores metadata for graphs, linking them to their respective sections and documents.

```sql
CREATE TABLE graphs (
    graph_id SERIAL PRIMARY KEY,
    document_id INT NOT NULL,
    section_id INT,
    graph_title VARCHAR(255),
    description TEXT,
    page_number INT,
    FOREIGN KEY (document_id)
      REFERENCES documents(document_id) ON DELETE CASCADE,
    FOREIGN KEY (section_id)
      REFERENCES sections(section_id) ON DELETE CASCADE
);
```

FIGURE 3.16: Table Contains the Graphs Informations

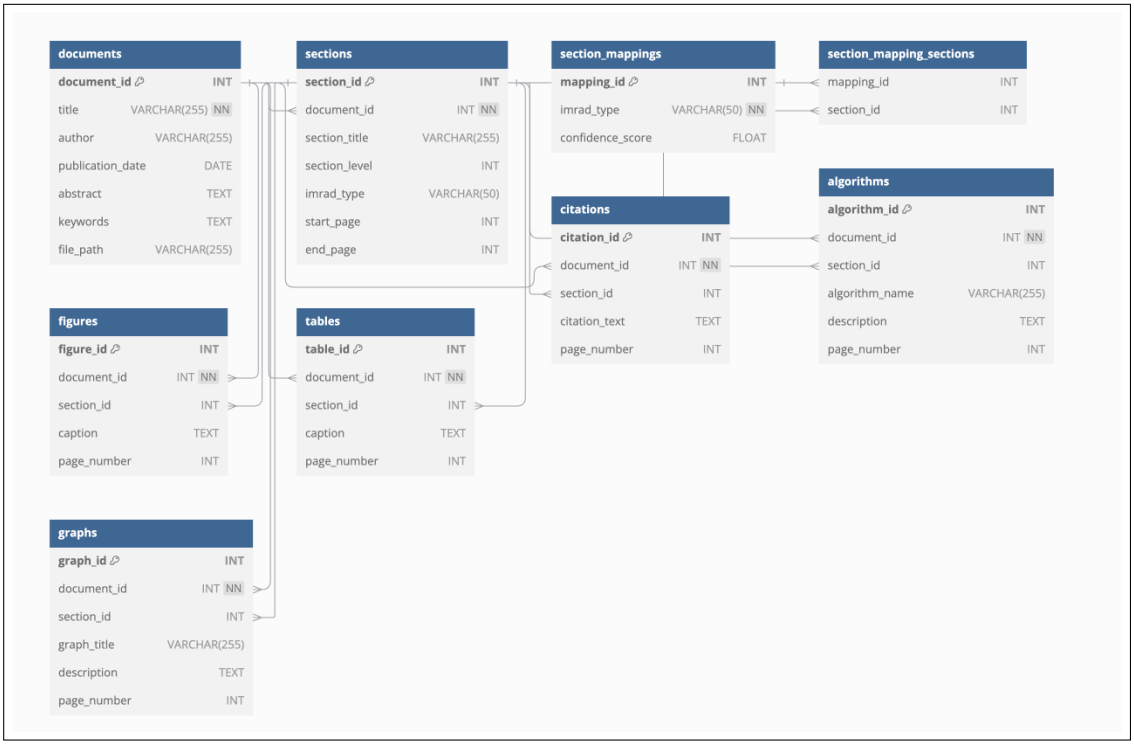


FIGURE 3.17: ERD of Database Schema

3.4 Data Extraction Engine (DEE)

The Data Extraction Engine (DEE) is critical in pre-processing the XML-formatted data. It meticulously filters and normalizes the data to remove "noise," which includes non-standard formatting, extraneous metadata, or inconsistent heading usage, thereby ensuring the data's integrity. The DEE extracts all pertinent XML tags through sophisticated parsing techniques, enabling the precise determination of the document's structure, including main headings and subheadings. This approach departs from previous methodologies by accurately preserving the document's intended logical structure. In addition to structural tags, the DEE identifies and extracts essential elements such as citations, figures, and tables, enriching the dataset with crucial metadata for further analysis.

The **purpose** of **DEE** is to ensure that only relevant content is stored in the database by filtering noise, such as footers and non-standard metadata. It identifies XML tags representing key elements and inserts them into appropriate database tables for further analysis.

3.5 Data Mapping Engine (DME)

Upon data extraction, the Data Mapping Engine (DME) is tasked with interpreting and aligning the extracted information with the IMRaD format, a standardized structure for scientific articles encompassing Introduction, Methods, Results, and Discussion sections. Employing a blend of XQuery, XPath, and SQL queries, the DME maps each document's sections to the corresponding IMRaD components based on established structural rules observed in the XML data. This engine ensures the accurate classification of subheadings within their respective main sections, eliminating the common misclassification issues observed in previous systems. By acknowledging the hierarchical relationship between headings and subheadings, the DME guarantees that the document's logical flow is maintained, significantly improving the precision of bibliographic analyses.

The **purpose** of **DME** is to enhance the logical flow of documents by eliminating misclassification issues common in previous systems. It accurately identifies subheadings and aligns them with their parent sections using SQL and XPath queries.

3.5.1 Subheadings Mappings

The structural organization of documents plays a pivotal role in bibliographic analysis, particularly within the context of scientific research publications. These documents are typically delineated into various sections, utilizing headings and subheadings to outline the hierarchical organization of content. A critical aspect of analyzing these documents involves accurately segregating and mapping these headings and subheadings to standardized sections such as those defined in the IMRaD (Introduction, Methods, Results, and Discussion) structure.

Historically, certain methodologies have approached the mapping of document structures by treating headings and subheadings as independent, standalone sections without due consideration for their hierarchical relationship. For example, in scenarios where an `<h1>` tag encapsulates multiple subheadings denoted by `<h2>` and `<h3>` tags, these methodologies might erroneously consider each tag as demarcating a distinct section. This approach can lead to inaccuracies, such as misclassifying IMRaD headings, whereby subheadings are mistakenly mapped as primary section headings.

To address this issue, the proposed approach introduces a nuanced method of distinction between main headings and subheadings utilizing regular expressions. This methodology diverges from previous practices by refraining from directly mapping subheadings to the IMRaD structure. Instead, subheadings are associated with the IMRaD section corresponding to their parent heading, thereby preserving the inherent document hierarchy. This strategy addresses two critical oversights in the scientific community's approach to document structure analysis:

1. **XML Tag Utilization:** Subheadings are distinguishable in the XML representation of documents through the use of specific tags (`<h2>`, `<h3>`, etc.). This tagging system facilitates the straightforward identification of hierarchical relationships between headings and subheadings.

2. **Regular Expression Analysis:** In instances where document formatting does not explicitly distinguish between headings and subheadings or where such distinctions are not adequately represented in XML tags, regular expressions are employed. These expressions are designed to discern the structural hierarchy based on patterns, such as numbering or bullet points, that imply a hierarchical organization.

Consideration of XML files in the dataset has revealed instances where document structures deviate from standard tagging conventions, necessitating alternative approaches for identifying headings and subheadings. For example, the `"region class=\DoCO:TextChunk""` element in XML can serve as an indicator of section demarcations in the absence of explicit header tags **??**. This observation underscores the necessity of a flexible, adaptive approach to document structure analysis capable of accommodating various formatting styles and conventions.

In summary, the proposed approach enhances the accuracy of section mapping within bibliographic analysis by judiciously distinguishing between main headings and subheadings. This distinction is achieved through explicit XML tagging and the strategic application of regular expressions, thereby addressing previously overlooked complexities in document structure interpretation.

3.5.2 Sections Sequences

Adopting the IMRaD (Introduction, Methods, Results, and Discussion) structure in academic writing represents a cornerstone in disseminating scientific knowledge. This standardized format facilitates the organization of research findings and enhances the readability and comprehension of scholarly documents. For Information Retrieval (IR) systems, the IMRaD structure provides a predictable framework

for indexing and querying research papers, thereby significantly improving the efficiency and accuracy of literature searches.

Despite the widespread endorsement of the IMRaD format within the scholarly community, deviations in adherence to its sequential order have been observed. A detailed analysis of PDF files from the dataset revealed instances where research papers diverged from the conventional sequence of the IMRaD structure. Such variations pose significant challenges for IR systems, which rely on the expected order of sections to identify and categorize content within documents accurately.

The implications of these deviations are twofold. Firstly, the non-standard arrangement of sections can lead to misinterpretation of content, whereby IR systems may incorrectly assign relevance to search queries based on misplaced sections. Secondly, the variability in document structure necessitates the development of sophisticated algorithms capable of recognizing and adapting to non-conventional arrangements of sections. Alternative methodologies have been explored to address these challenges. These include implementing advanced parsing algorithms that employ natural language processing (NLP) techniques to understand the context and content of sections, regardless of their order. Analyzing linguistic cues and thematic continuity, these algorithms can accurately identify the intended IMRaD sections even when presented in an unconventional sequence. In summary, while the IMRaD structure remains a foundational element in scientific communication, the observed deviations in section sequencing necessitate reevaluating IR systems' strategies for document analysis. Through the adoption of NLP and machine learning techniques, it is possible to enhance the robustness of IR systems, ensuring accurate section identification and categorization across a wide range of document structures. This approach's evolution mitigates the challenges posed by non-standard arrangements and paves the way for adaptive and intelligent literature retrieval systems.

3.5.3 Sections Known Names

Academic research papers are meticulously organized into sections that guide the reader through the research journey, from initial inquiry to conclusions. This

organization not only facilitates a systematic approach to presenting research but also aids in the comprehension and analysis of the study. Recognized universally across scholarly domains, these sections serve distinct purposes, reflecting the multifaceted nature of research dissemination. The following detailed overview provides insights into each section's role within the academic manuscript:

- **Abstract**: A succinct summary of the research study, including its aims, methodology, key findings, and conclusions, designed to provide a quick overview for the reader.
- **Keywords**: Selected terms that encapsulate the core themes and subjects of the paper, facilitating searchability and indexing in databases.
- **Introduction**: Introduces the research topic, outlines the research problem, and states the study's objectives, setting the stage for the subsequent investigation.
- **Background or Theoretical Framework**: Provides a detailed context for the study, reviewing relevant literature and establishing a theoretical foundation for the research.
- **Literature Review / History / Related Work**: A comprehensive examination of existing research related to the study's focus, highlighting gaps the current research aims to fill.
- **Methodology / Methods**: Describes the research design, data collection methods, and analytical techniques employed, ensuring reproducibility and transparency.
- **Experimental Design / Experimental Setup**: Details the experimental framework, including variables, controls, and the experimental environment, critical for empirical studies.
- **Data Collection / Data Acquisition**: Elucidates the processes and tools used to gather research data, underscoring the study's empirical basis.
- **Data Analysis / Data Processing**: Explains the methodologies applied to analyze the collected data, including statistical tests, qualitative analysis techniques, and data interpretation methods.

- **Results / Findings / Observations**: Presents the outcomes of the data analysis, objectively reporting the study's findings without interpretation.
- **Discussion**: Interprets the results, discusses their implications in the context of existing literature, and may also integrate findings directly.
- **Conclusions and Future Directions**: Summarizes the study's key takeaways, its contributions to the field, and potential avenues for future research.
- **Limitations**: Acknowledges the study's constraints and potential biases, lending credibility to the research.
- **Recommendations**: Offers suggestions for practical applications of the research findings or proposes areas for further study.
- **Acknowledgments**: Credits individuals, organizations, or funding bodies that contributed to the research but were not directly involved in its execution.
- **Appendices / Supplementary Materials**: Contains additional data and materials that support the paper's content but are not essential to its primary narrative.
- **Ethical Considerations**: Addresses ethical issues related to the research, including consent and data privacy, where applicable.
- **Conflict of Interest Statement**: Discloses any potential conflicts of interest that could influence the research outcomes or interpretations.
- **Funding Information**: Lists the sources of financial support for the research, acknowledging the role of funding bodies in facilitating the study.
- **References / Bibliography**: Compiles all works cited in the paper, enabling readers to explore the research context and background further.

While generally consistent across disciplines, the nomenclature for these sections can exhibit considerable variation, reflecting the diverse styles and preferences within the scientific community. This variability presents a significant challenge for Information Retrieval (IR) systems tasked with parsing and categorizing the content of research articles. The deviation from standardized section names, such as the substitution of "Evaluation" or "Analysis" for the traditional "Results/Discussion" section, necessitates a flexible and adaptive approach to document analysis.

Recognizing the importance of accurately mapping document sections regardless of the specific terminology employed, this study embarked on a comprehensive review of synonyms used to denote the various sections of research papers. Drawing upon a dataset of 1200 research articles, an exhaustive list of synonyms was compiled for each commonly recognized section. This list serves as a critical resource for the proposed methodology, enabling matching section names to their logical equivalents within the IMRaD structure, even in instances where non-standard terminology is used.

The process of synonym extraction and matching represents a significant advancement in the field of document analysis. It offers a solution to one of the intractable problems faced by IR systems: the accurate identification and categorization of document sections in the presence of varied nomenclature. By leveraging this comprehensive synonym database, the proposed methodology demonstrates a heightened sensitivity to the nuances of academic language, ensuring that the logical structure of research papers is preserved and accurately reflected in IR systems.

In summary, identifying and utilizing section name synonyms play a pivotal role in enhancing the precision of IR systems. Through a meticulous analysis of a vast corpus of research articles, this study has not only highlighted the diversity of terminology within academic writing but also provided a robust framework for navigating this variability. The result is a adaptable and intelligent approach to document analysis, capable of accommodating the rich tapestry of expressions and terms employed by the scholarly community.

Introduction

- Introduction
- Background or Theoretical Framework
- Literature Review

Methods

- Methodology / Methods
- Experimental Design / Experimental Setup

- Data Collection / Data Acquisition
- Data Analysis / Data Processing
- Architecture / Architectural Design

Results

- Results
- Findings
- Observations

Discussion

- Discussion (may include findings from the Results section)
- Limitations
- Recommendations
- Conclusions and Future Work

Note that some sections may not fit neatly into IMRaD and are thus categorized separately. Following is the list of those sections.

- Abstract
- Keywords
- Acknowledgments
- Appendices / Supplementary Materials
- Ethical Considerations
- Conflict of Interest Statement
- Funding Information
- References / Bibliography

The detailed exploration and mapping elucidate the structured approach to academic writing, facilitating clarity, coherence, and comprehensiveness in research dissemination. Recognizing and aligning the varied section names with the IMRaD

format underscores the adaptability and universality of this structure in scholarly communication. This alignment not only aids authors in organizing their manuscripts effectively but also assists readers and researchers in navigating the document efficiently, enhancing the accessibility and impact of scientific findings.

In summary, the meticulous categorization and mapping of section names to the IMRaD structure, supplemented by an understanding of additional sections beyond this format, provide a foundational framework for academic writing. This framework supports the effective presentation of research, ensuring that each element of the manuscript contributes meaningfully to the overarching narrative of scientific discovery. This structured approach renders academic papers comprehensive and navigable, maximizing their utility and relevance within the scholarly community.

3.5.4 Citation Count as a Metric for Section Identification

Citations are a critical aspect of scientific communication, providing a way to acknowledge previous work and establish the context for current research. Different sections of a paper exhibit varying citation densities. For example, sections like *Literature Review* or *Related Work* are often citation-rich, reflecting their role in summarizing existing studies. In contrast, sections like *Results* or *Discussion* may contain fewer citations, as they focus on presenting new findings. In this study, XML-formatted documents are analyzed to extract citation elements, such as `<Xref>` and `<ref>`, using attributes like `ref-type='bibr'`. The citation count for each section is calculated and used as a metric to identify and categorize the sections accurately.

3.5.5 Object For Section Identification

3.5.5.1 Figures as a Metric for Section Identification

Figures provide visual summaries of data and are critical for conveying experimental results or methodologies. Sections such as *Results* and *Discussion* typically have

a higher density of figures, reflecting their focus on presenting and interpreting findings. In XML-formatted documents, figures are represented with tags like `<FigureBox>`. By analyzing the occurrence and distribution of these elements, the system can effectively identify and classify sections that rely heavily on visual data representation.

3.5.5.2 Algorithms as a Metric for Section Identification

Algorithms are often included in papers to describe computational methods or processes. These elements are most commonly found in the *Methodology* section, where they outline the steps taken to achieve the results.

Using XML tags such as `<Algorithm>`, this study identifies and counts algorithms within each section. The frequency and placement of algorithms are used to refine section classification and confirm the logical structure of the document.

3.5.5.3 Graphs as a Metric for Section Identification

Graphs are essential for illustrating trends and relationships in data, making them a frequent feature of the *Results* and *Discussion* sections. Their occurrence provides valuable insight into the purpose of a section.

By detecting XML elements like `<Graph>`, this study analyzes the distribution of graphs to enhance the accuracy of section identification and mapping to logical categories.

3.5.5.4 Tables as a Metric for Section Identification

Tables are widely used to organize and summarize data in a concise format. They are particularly prevalent in the *Methodology* and *Results* sections, where they present experimental setups, datasets, or outcomes.

XML tags such as `<TableBox>` are used to locate and count tables within a document. This information aids in identifying sections that rely on tabular data, contributing to the overall accuracy of section classification.

Algorithm 1 IMRaD Section Mapping

Input: XML File: *xmlFile*
Output: IMRaD Section Mapping: *imradMapping*

Initialize *imradMapping* ← {} ▷ Empty mapping dictionary
Initialize *imradSequence* ← {Introduction, Methodology, Results, Discussion}
Load XML structure from *xmlFile* into *sectionsList*
Analyze *sectionsList* to extract section titles and order
▷ Step 1: Match section sequence to IMRaD
for all section ∈ *sectionsList* **do**
    Match *section.title* to closest label in *imradSequence*
    Add mapping to *imradMapping*
end for ▷ Step 2: Refine mapping using IMRaD sequence
for all mappedSection ∈ *imradMapping* **do**
    if *mappedSection* is out of logical order based on *imradSequence* **then**
        Reorder *mappedSection* to align with IMRaD flow
    end if
end for ▷ Step 3: Refine using count objects
Initialize statistics ← {}
for all section ∈ *sectionsList* **do**
    Compute counts: *citations*, *figures*, *graphs*, *algorithms*, *tables*
    Update statistics for each section
end for ▷ Step 4: RaMapp the IM D
for all mappedSection ∈ *imradMapping* **do**
    if counts suggest a different IMRaD label **then**
        Update *imradMapping* based on refined statistics
    end if
end for
Return *imradMapping* ▷ Return the IMRaD mappings

3.6 IMRaD Section Mapping

The proposed algorithm, detailed in Algorithm 1, outlines a systematic approach for mapping sections of a document to the IMRaD (Introduction, Methodology, Results, and Discussion) structure. The algorithm takes an XML file as input and produces a refined mapping of sections to the IMRaD categories as output.

Input and Output

- **Input:** An XML file (*xmlFile*) containing the structure of a document, including section titles and content.
- **Output:** A mapping (*imradMapping*) that assigns each section in the document to one of the IMRaD categories.

Steps of the Algorithm 1

The algorithm is divided into three primary steps, as explained below:

Step 1: Match Section Sequence to IMRaD

The algorithm starts by analyzing the section titles from the XML file. Each section title is compared with the IMRaD categories (*Introduction, Methodology, Results, Discussion*) to identify the closest match. This step creates an initial mapping of sections to IMRaD labels.

- Section titles are matched using heuristic or textual similarity measures.
- The initial mapping is stored in a dictionary (*imradMapping*).

Step 2: Refine Mapping Using IMRaD Sequence

Once the initial mapping is created, the algorithm refines it by ensuring that the sequence of sections adheres to the logical flow of the IMRaD structure:

1. Checks if any section is mapped out of order, e.g., a *Results* section appearing before *Methodology*.
2. Reorders the mapping to align with the standard IMRaD sequence.

Step 3: Refine Using Counts of Citations, Figures, Graphs, Algorithms, and Tables

The final refinement step involves leveraging statistical data within each section to validate and improve the mapping:

- For each section, counts of citations, figures, graphs, algorithms, and tables are computed.
- Based on these counts, sections are re-evaluated to ensure that their mapping aligns with their content. For example:
  - Sections with a high number of figures and graphs are likely to belong to the *Results* category.
  - Sections with a high number of citations may correspond to *Introduction* or *Related Work*.

Final Output

After the refinements, the algorithm returns the completed *imradMapping*, which provides a reliable assignment of each document section to one of the IMRaD categories.

3.7 Execution and Processing Time

The performance of the proposed approach was evaluated on an **AWS medium instance environment**. Each experiment was repeated ten times to ensure consistency, and the **median values** were reported for comparison. Table 3.1 presents a detailed comparison of the execution time between the proposed approach and three state-of-the-art techniques. The experiments in this research were conducted on a medium Amazon Web Services (AWS) compute instance. The hardware specifications of the instance used are as follows:

- **Instance Type:** AWS EC2 m5.xlarge
- **vCPUs:** 4
- **Memory:** 16 GiB
- **Storage:** EBS-optimized instance with 100 GB SSD
- **Networking:** Up to 10 Gbps
- **GPU:** Not applicable (CPU-based computation)

The method by Ding et al. [1] was the fastest, completing the task in only 10 seconds due to its reliance on **dictionary-based section identification**, which is less computationally demanding. Shahid et al. [2] required 14 seconds, as it combines **templates with dictionary terms**, increasing processing complexity slightly. Habib et al. [3] took 21 seconds since it involves **citation counting and text parsing**, which requires additional computational effort.

The proposed approach offers improved accuracy while taking the most time at 34 seconds. This additional processing time results from the **integration of multiple features**, such as figures, tables, and citation frequency, making it comprehensive than other methods. However, since the proposed method is executed offline, **preprocessing and database maintenance** ensure that query response times remain fast, with results precompiled and ready for users.

TABLE 3.1: Processing Time Comparison of Techniques

Techniques	**Time (Seconds)**
Ding et al. [1]	10
Shahid et al. [2]	14
Habib et al. [3]	21
Proposed Approach	34

The comparison shown in Table 3.1 highlights the trade-off between **speed and accuracy**. Ding et al.'s dictionary-based method is the fastest but lacks the

depth required for complex analyses. Shahid et al.'s approach balances speed and complexity, while Habib et al.'s method provides better accuracy through text parsing and citation analysis. The proposed method achieves the **highest accuracy** but with a longer processing time. Nevertheless, since the method operates offline, users benefit from precompiled results, ensuring **fast query responses**. This offline approach minimizes the impact of the longer processing time on user experience and significantly improves precision for identifying related research papers.

3.8 Results

Once all the modules of the proposed methodology have been implemented, the next steps involve evaluating data using the benchmark data set.

TABLE 3.2: Comparison of Precision of Combined Sections

Approach	**Precision**
Ding	0.87
Shahid	0.81
Habib	0.89
Proposed	0.97

TABLE 3.3: Comparison of Recall of Combined Sections

Approach	**Recall**
Ding	0.81
Shahid	0.81
Habib	0.89
Proposed	0.97

The outcomes are evaluated in two phases: (1) identified a one-layer hierarchy of sub-headings and then mapped the sections to the IMRaD (2) Afterwards, determined the collective contribution of all the parameters in a hybrid manner by combining the parameters "citation count," "figures count" and "tables count."

TABLE 3.4: Comparison of F-Measure of Combined Sections

Approach	**F-Measure**
Ding	0.81
Shahid	0.81
Habib	0.89
Proposed	0.97

3.8.1 Precision and Recall Calculations for IMRaD Sections

This section explains how precision and recall are computed for different sections of research papers (Introduction, Methodology, Results, and Discussion).

Each term and symbol is carefully explained to ensure clarity.

Equation 3.1: Precision and Recall for Each Section

$$\begin{aligned} P_x &= \frac{TP_x}{TP_x + FP_x} \\ R_x &= \frac{TP_x}{TP_x + FN_x} \\ \Rightarrow x_0 &= \text{Introduction} \\ \Rightarrow x_1 &= \text{Methodology} \\ \Rightarrow x_2 &= \text{Results} \\ \Rightarrow x_3 &= \text{Discussion} \end{aligned} \tag{3.1}$$

Explanation of Equation 3.1

Equation 3.1 defines the precision and recall for each section x, where x is an index referring to one of the four sections of a research paper.

Precision (P_x) Measures the proportion of correctly identified instances of section x to all instances classified as section x.

$$P_x = \frac{TP_x}{TP_x + FP_x}$$

Recall (R_x) Measures the proportion of correctly identified instances of section x to all actual instances of section x.

$$R_x = \frac{TP_x}{TP_x + FN_x}$$

Notation Used

- TP_x: True Positives – The number of correctly identified section x instances.
- FP_x: False Positives – The number of instances incorrectly classified as section x.
- FN_x: False Negatives – The number of actual instances of section x that were missed.
- x_0: Introduction section.
- x_1: Methodology section.
- x_2: Results section.
- x_3: Discussion section.

Equation 3.2: Average Precision and Recall Across All Sections

$$\begin{aligned} P_{\text{avg}} &= \frac{\sum_{x=0}^{3} P_x}{4} \\ R_{\text{avg}} &= \frac{\sum_{x=0}^{3} R_x}{4} \end{aligned} \tag{3.2}$$

Explanation of Equation 3.2

Equation 3.2 calculates the system's overall performance by computing the average precision and recall across the four IMRaD sections (Introduction, Methodology, Results, and Discussion).

Average Precision (P_{avg}) The arithmetic mean of precision values for all sections. It measures the system's overall accuracy in identifying the correct sections.

$$P_{\text{avg}} = \frac{\sum_{x=0}^{3} P_x}{4}$$

Average Recall (R_{avg}) The arithmetic mean of recall values for all sections. It measures the system's ability to correctly identify each section's instances.

$$R_{\text{avg}} = \frac{\sum_{x=0}^{3} R_x}{4}$$

Notation Used

- P_{avg}: Average precision across all sections.
- R_{avg}: Average recall across all sections.
- P_x: Precision for section x.
- R_x: Recall for section x.

Detailed Breakdown of the Calculations

The precision (P_x) and recall (R_x) for each section x are computed using the true positive, false positive, and false negative counts. The index x is used to denote the specific section being evaluated:

$$x_0 = \text{Introduction}$$
$$x_1 = \text{Methodology}$$
$$x_2 = \text{Results}$$
$$x_3 = \text{Discussion}$$

How the Metrics Are Averaged The average precision and recall are computed by summing 3.3 the individual precision and recall values for all four sections and dividing by 4, reflecting each section's equal contribution to the overall performance evaluation.

$$\begin{aligned} P_{avg} &= \sum_{x=0}^{3} P_x/4 \Rightarrow \ P_{avg} = Average\ Precision \\ R_{avg} &= \sum_{x=0}^{3} R_x/4 \Rightarrow \ R_{avg} = Average\ Recall \end{aligned} \tag{3.3}$$

3.9 Evaluation

The comprehensive evaluation of the proposed system for mapping research paper sections to the IMRaD structure is crucial for assessing its effectiveness and advancements over existing methodologies. This section outlines the evaluation framework, detailing the dataset, metrics, comparative analysis, and statistical significance of the findings.

3.9.1 Comparative Analysis

System performance was compared with three renowned methodologies: Ding et al. [1], Shahid et al. [2], and Habib and Afzal [3], based on the metrics mentioned

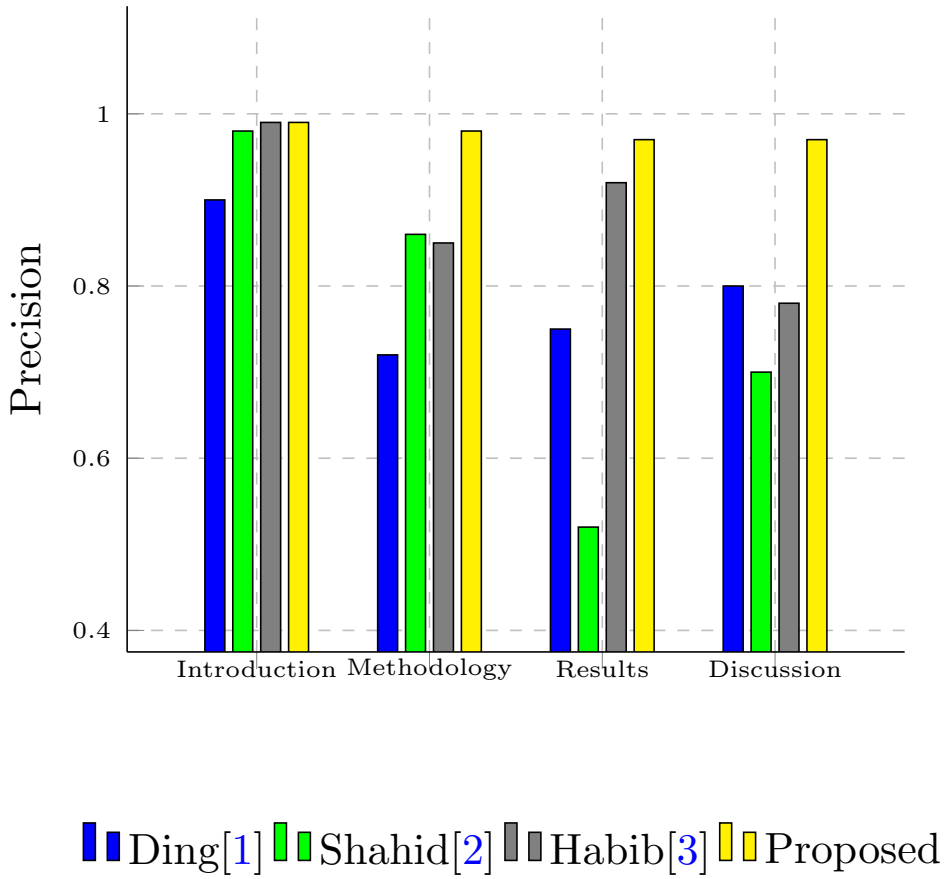


FIGURE 3.18: Section-Wise comparison of Precision

above, under identical conditions on the same dataset.

This analysis indicated that the proposed system consistently outperformed the comparative group across all metrics, achieving a precision, recall, and F-measure of 0.97. This suggests significant improvements in accurately identifying and mapping sections to the IMRaD structure.

The proposed system achieved a precision rate of 97%, a notable improvement over the closest competing system, which recorded a precision of 89%. This enhancement in precision is attributable to the refined algorithm's ability to accurately identify and categorize sections into the IMRaD structure, underscoring the system's sophisticated analytical capabilities.

In conclusions the detailed analysis of results substantiates the proposed system's superiority in accurately mapping research paper sections to the IMRaD structure. These advancements open new avenues for future research, particularly in optimizing

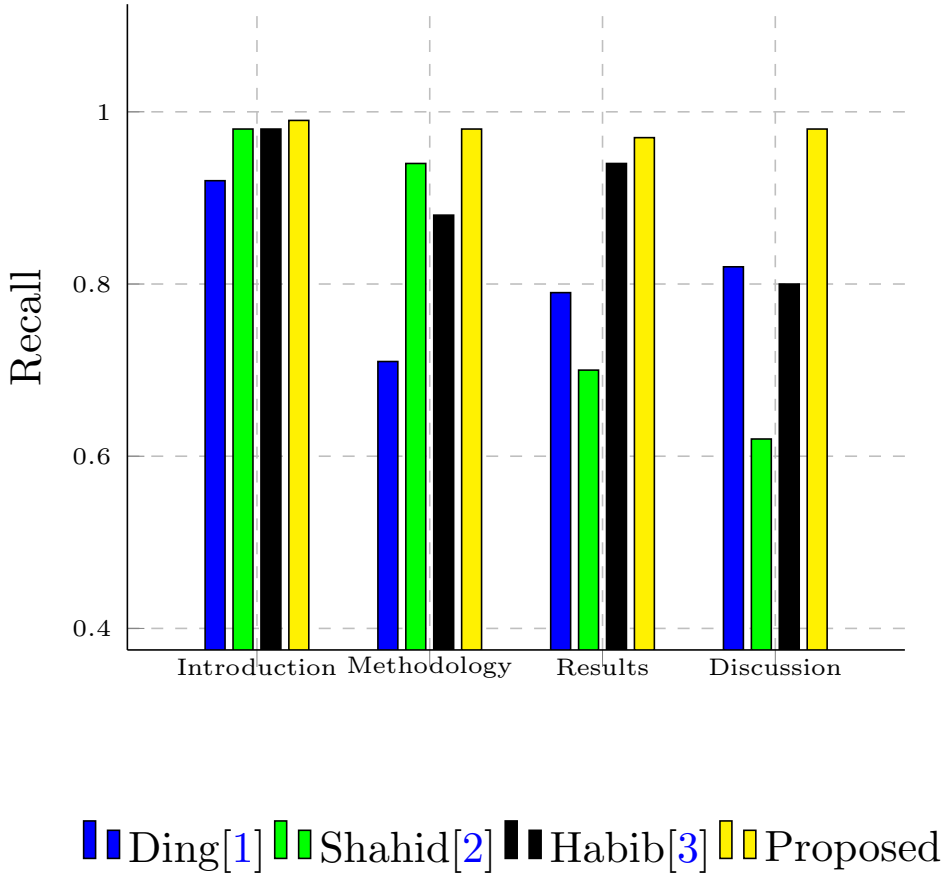


FIGURE 3.19: Section-Wise comparison of Recall

algorithmic efficiency and exploring additional discriminative features that could further enhance precision and recall. Future studies might also explore the system's scalability to accommodate larger datasets and its adaptability to different scientific disciplines.

3.10 Discussion

The performance evaluation and results of the proposed system reveal a significant improvement over existing methodologies in mapping section headings to the IMRaD structure. The enhancements in precision, recall, and F-measure demonstrate the effectiveness of incorporating features such as subheadings mapping, figures and tables count, and in-text citation frequency. These findings highlight the potential of these features to address some of the core challenges identified in previous research, such as Shahid et al. [2], which primarily relied on template

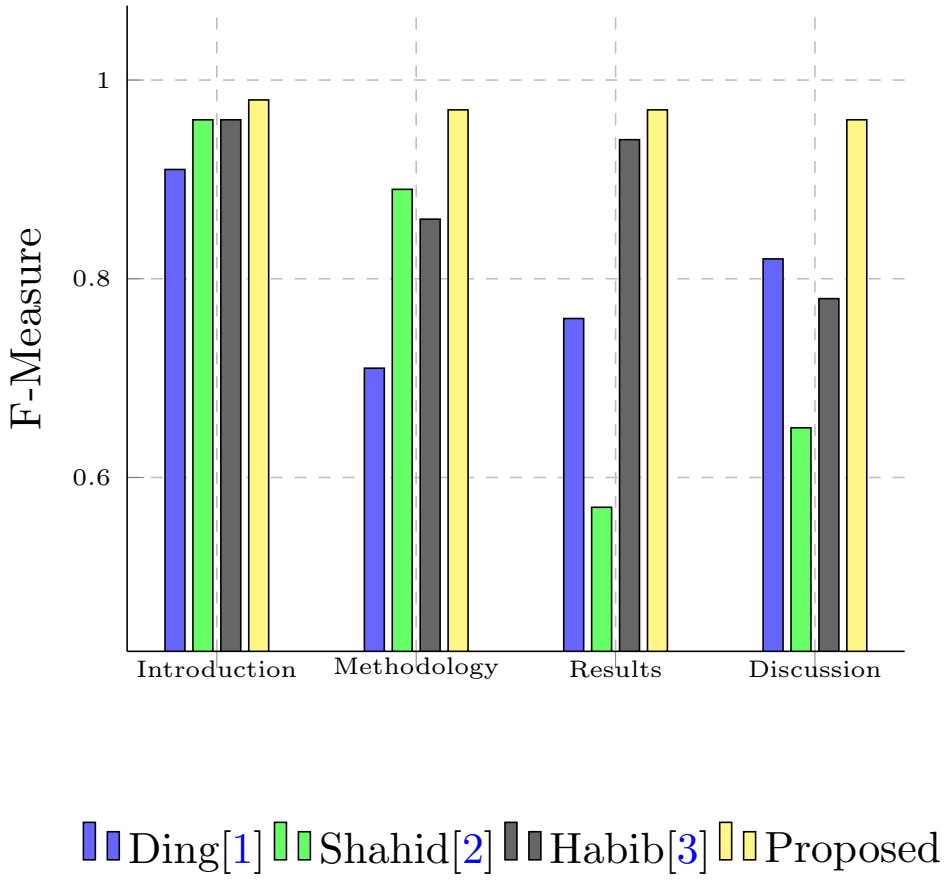


FIGURE 3.20: Section-Wise comparison of F-Measure

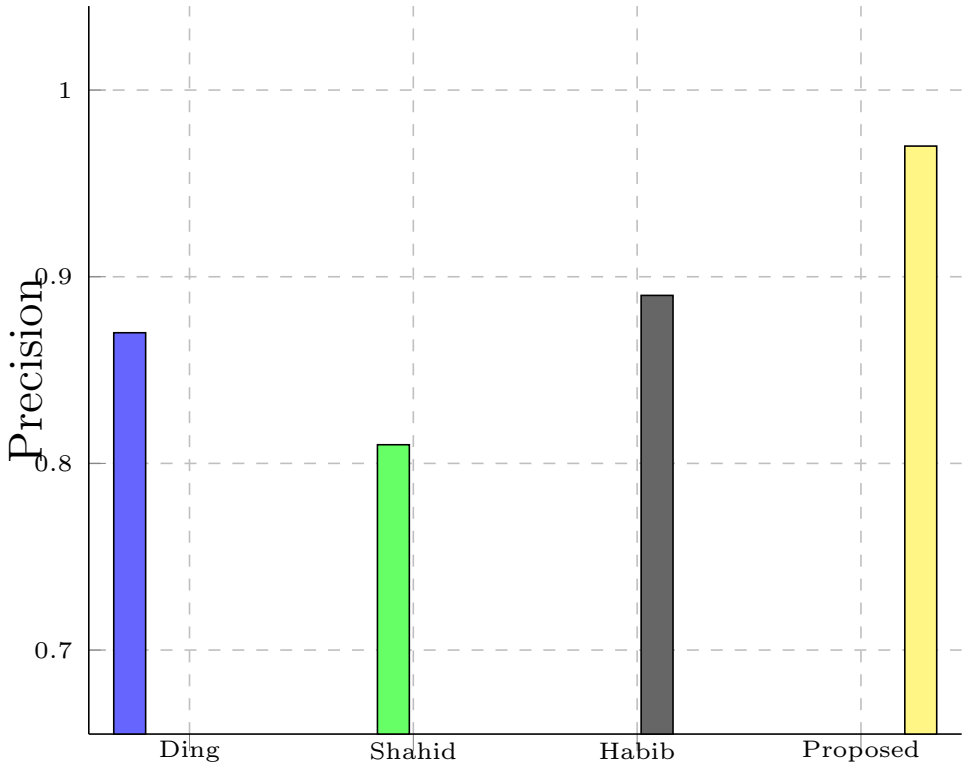


FIGURE 3.21: Comparison of Precision of combined sections

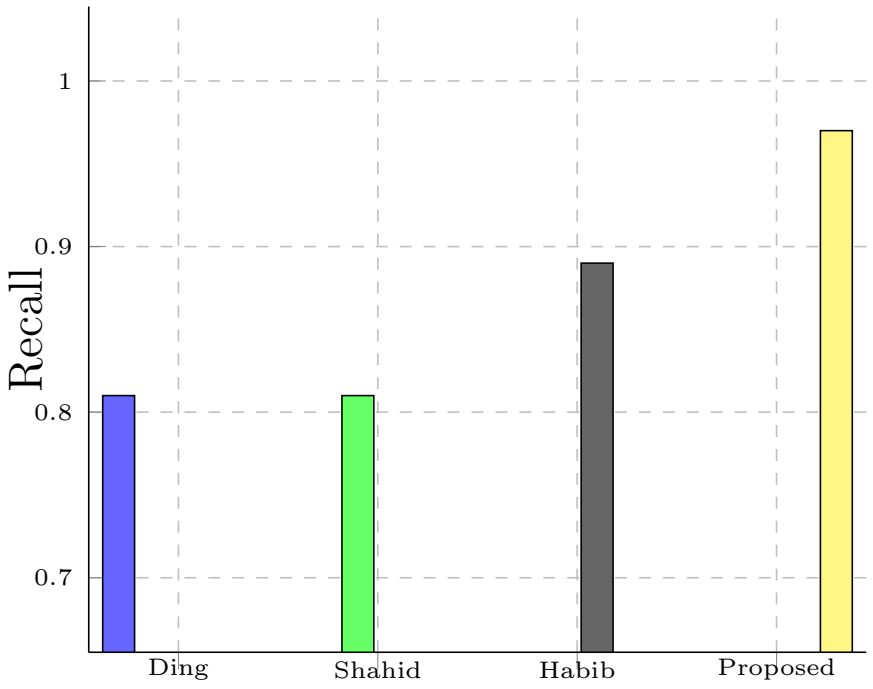


FIGURE 3.22: Comparison of Recall of combined sections

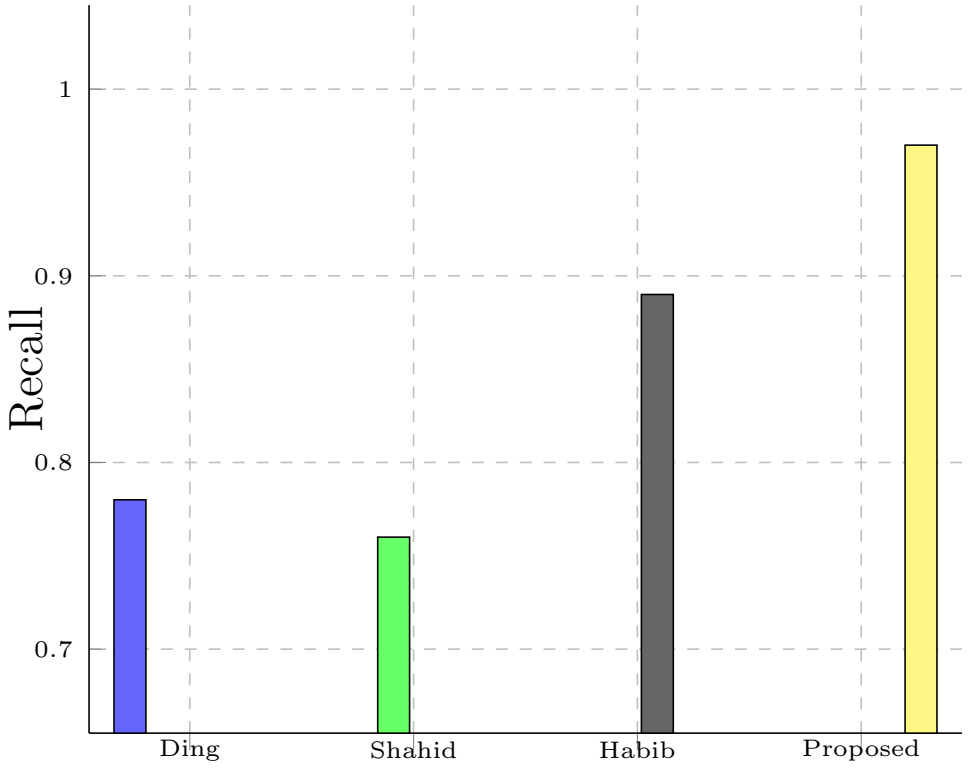


FIGURE 3.23: Comparison of F-Measure of combined sections

matching and dictionary terms. Ding et al. [1], which did not fully account for the hierarchical relationships between sections and subsections. The proposed methodology's use of regular expressions for implicit mapping of subsections to their respective main sections has also positively impacted the overall accuracy of the system. Unlike previous studies, which treated subsections as independent entities, this approach acknowledges the logical structure of academic documents. It ensures that subheadings contribute to the correct identification of primary IMRaD sections. This finding supports the assertion by Habib and Afzal [3] that improving the handling of document structures can enhance the performance of information retrieval systems. Moreover, introducing new discriminative features, such as the count of figures, tables, and in-text citations, has proven beneficial in distinguishing between sections that are often challenging to classify. For instance, sections like "Results" or "Discussion" generally contain more figures and tables, while the "Literature Review" tends to have a higher frequency of citations. By incorporating these features, the system can accurately identify previously misclassified sections, as noted by Ding et al. [1]. This addresses the gap in prior research where content-specific features were ignored or underutilized.

However, while the proposed system's offline processing addresses the computational complexity concerns, it introduces a trade-off regarding processing time. Although the average processing time of 34 seconds per document is longer than other methods, this is mitigated by the fact that this process is conducted offline and results are precompiled. This implies that the system is suitable for applications where the quality of mapping and classification takes precedence over real-time performance, such as in citation indexing and digital library management. The statistical significance of the improvements in precision, recall, and F-measure further confirms the robustness of the proposed methodology. Despite the advancements, there remain areas for further exploration. For example, future work could investigate the system's scalability for larger datasets and its adaptability to different academic domains. Additionally, integrating natural language processing (NLP) techniques could further refine section identification, especially in documents that deviate from conventional IMRaD structures. In conclusion,

this discussion highlights the proposed system's contributions to addressing the deficiencies of previous methodologies and outlines potential avenues for enhancing document section mapping in academic literature. By integrating novel features and acknowledging the inherent hierarchical structure of research papers, this work lays a foundation for accurate and reliable bibliographic analyses.

Summary

This chapter introduces a comprehensive system designed to accurately identify section headings within research papers and map them to the IMRaD structure. It addresses research questions aimed at enhancing the accuracy of this mapping process. This involves determining potential features to increase mapping accuracy, devising an intelligent section and boundary identification algorithm, and auto-correcting wrongly identified subsections. The methodology builds upon the research of Shahid et al. and Habib and Afzal, identifying gaps in their approaches, such as the need for features that could improve the precision and recall of section mapping. The system has been developed to overcome these challenges by integrating novel features such as subheading mapping, figures and tables count, and in-text citation frequency. These features are crucial for accurately identifying and classifying sections using the IMRaD format.

The chapter delves into the specifics of these features, explaining how subheadings play a vital role in the structure of research papers and how the frequency of figures, tables, and in-text citations can indicate the association with specific IMRaD sections. It highlights the importance of distinguishing between main headings and subheadings using regular expressions to prevent false mapping and improve information retrieval system performance. The methodology section outlines the proposed system's four main modules: Schema Generation Engine (SGE), Data Extraction Engine (DEE), Data Mapping Engine (DME), and Mapping View Engine (MVE). Each module is specific in converting PDF files to XML format, extracting relevant data, mapping extracted data to IMRaD sections, and visualizing the mapping results. The system architecture is designed to handle

extensive datasets, with the data stored in a relational database format for easy manipulation. Performance evaluation compares the proposed system with state-of-the-art techniques, demonstrating its precision, recall, and F-measure superiority. The evaluation is conducted in two phases, focusing on the hierarchy of sub-headings and the hybrid contribution of features like citation, figure, and table counts to the mapping process.

The results section compares the proposed system's performance with existing techniques. It showcases the significant improvements achieved by incorporating novel features and a sophisticated mapping algorithm. The proposed system accurately identifies and maps sections to the IMRaD structure, outperforming previous approaches. This chapter presented a system for accurately extracting section headings from research papers and mapping them to the IMRaD structure. By integrating unique features and a comprehensive methodology, the system significantly improves the precision and recall of section mapping, offering promising implications for citation indexes and digital libraries. The system's performance, validated through rigorous evaluation against existing techniques, establishes a new benchmark for future research in this domain.

Chapter 4

Section's Ranking and Weights Adjustment to Discover Bibliographically Coupled Papers

4.1 Overview

This chapter presents a system for improving the identification of bibliographically coupled research papers by dynamically adjusting section weights. The main research question is assigning the weights to the IMRaD structure and finding the related papers. To achieve this, a deep learning neural network is used to fine-tune the weights assigned to different sections of research articles.

The literature review highlights the availability of two datasets, Dataset-1 and Dataset-2. The dataset referred to as Dataset-2, used in prior studies, provides a benchmark for comparison. However, since this dataset was created using statistical methods like Jensen-Shannon Divergence (JSD) and is not manually annotated, it requires further validation. The Dataset-1 consists of manually annotated data, ensuring precise labeling for evaluation. This chapter also covers the validation process for the first dataset to ensure its reliability. Finally, the chapter presents

the outcomes of the proposed technique on both datasets and offers a detailed discussion of the results.

The chapter is Phase 3 (Planning the Research Study) of Research Methodology. The specific focus of this chapter is to answer the following research questions:

- **RQ2**: How can sections' weights be tuned to maximize the correlation between Bibliographic Coupling strength and paper relatedness?

4.2 Introduction

In today's digital age, researchers are increasingly challenged by the sheer volume of academic publications. Identifying related research papers has become critical for conducting comprehensive literature reviews, building on existing knowledge, and avoiding duplication of efforts. Bibliographic coupling, which connects papers through shared references, offers a promising solution. However, traditional bibliographic coupling approaches treat all research papers equally, disregarding the distinct roles that sections such as Introduction, Methodology, Results, and Discussion (IMRaD) play in academic discourse.

Citations in different sections reflect varying types of relationships between papers. For instance, a citation in the introduction often reflects conceptual or theoretical influence, whereas one in the results section might indicate alignment in experimental outcomes. Assigning uniform weights to all sections in bibliographic coupling can obscure these nuances, reducing the effectiveness of the coupling process. To address this gap, this research proposes a system that dynamically adjusts the weights assigned to different sections, improving the discovery of related research papers.

This study builds on the work of Habib and Afzal [3], who demonstrated the value of section-specific weights but relied on static, heuristic assignments. This research seeks to refine the weight adjustment process by employing a deep learning-based

approach, optimizing the weights through backpropagation. Additionally, Jensen-Shannon Divergence (JSD) is utilized to validate the dataset and ensure accurate identification of related papers. The outcome is a nuanced bibliographic coupling mechanism that better reflects the conceptual connections between research papers, enabling researchers to discover relevant works precisely.

4.3 Methodology

This section outlines the methodology for enhancing bibliographic coupling by assigning dynamic section weights. The approach leverages machine learning and statistical techniques to identify related research papers accurately. Specifically, a Deep Neural Network (DNN) model is employed to fine-tune section weights.

4.3.1 Data Collection

The datasets used for this research include manually annotated and statistically generated datasets.

- **Dataset-1:** A manually curated dataset providing ground truth for evaluating section-wise bibliographic coupling. Domain experts have annotated this dataset to ensure reliable clustering of related papers.
- **Dataset-2:** This dataset was generated using JSD-based clustering by Habib and Afzal [3]. Although it captures thematic similarities, its automated creation introduces potential biases, necessitating further validation.

The collected datasets include metadata such as titles, authors, abstracts, publication dates, and full-text sections structured according to the IMRaD format.

Dataset-1 is a manually annotated collection that serves as the benchmark or "ground truth" for our experiments. This dataset has been carefully curated by domain experts who manually assessed and annotated relationships between

papers based on their thematic and conceptual connections. Due to this manual curation process, Dataset-1 is inherently reliable and accurately reflects the expected clustering of research papers. Therefore, it does not require further validation.

On the other hand, **Dataset-2** is derived using an automated approach. Specifically, Habib and Afzal [3] utilized **Jensen-Shannon Divergence (JSD)** to generate this dataset by quantifying textual similarity between papers. Unlike Dataset-1, Dataset-2 is not manually annotated; it relies entirely on JSD's statistical analysis to identify and cluster related documents. Given this automatic generation process, the clusters produced by JSD in Dataset-2 need to be validated to ensure they accurately capture the thematic and conceptual relationships in the research papers. This validation is crucial because JSD effectively clusters documents based on their semantic content.

4.3.2 Dataset Validation Methodology

To validate the accuracy and reliability of Dataset-2, employed a step-by-step procedure that involves comparing the clusters generated by JSD in Dataset-2 with the ground truth provided by Dataset-1. Precision, recall, and F1-score metrics are used to evaluate the clustering performance and identify areas for improvement.

4.3.2.1 Step-by-Step Validation Process

1. **Cluster Extraction** Extract clusters from both Dataset-1 (manually annotated) and Dataset-2 (JSD-generated). Each cluster in Dataset-2 contains papers grouped based on their JSD scores, representing thematic similarity.

2. **Pairwise Comparison** For each cluster in Dataset-2, identify all possible pairs of papers. Check if these pairs exist in the corresponding clusters of Dataset-1 to determine if they match the ground truth.

3. **Precision Calculation** Calculate precision as the ratio of correctly identified paper pairs in Dataset-2 (also in Dataset-1) to the total number of paper pairs generated by JSD in Dataset-2. Precision indicates how many of the

identified pairs in Dataset-2 are genuinely related, based on the manual annotations in Dataset-1.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

4. **Recall Calculation:** Compute recall as the ratio of correctly identified pairs in Dataset-2 to the total number of related pairs in Dataset-1. This measures how JSD captures all the relevant paper relationships in the ground truth.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

5. **F1-score Calculation:** Determine the F1-score, which is the harmonic mean of precision and recall. The F1-score provides a balanced assessment, considering the clustering process's accuracy (precision) and completeness (recall).

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

6. **Analyze Discrepancies** If discrepancies arise between Dataset-2 and Dataset-1 (e.g., pairs that exist in Dataset-1 but not in Dataset-2), a in-depth analysis is conducted to understand the nature of these discrepancies. This may involve examining specific clusters to identify patterns in JSD's clustering performance, such as whether some thematic regions are under or overrepresented.

4.3.2.2 Preprocessing

Text preprocessing is crucial to normalizing the data and making it suitable for analysis. The preprocessing steps include:

- **Tokenization** This step involves splitting the text into individual words or phrases, ensuring each word or phrase is treated as an individual unit of meaning. For example, the sentence "The quick brown fox" would be tokenized into ["The", "quick", "brown", "fox"].

- **Stop-word Removal** Common words that do not contribute significant meaning to the text (e.g., "the," "and," "in") are removed. This step reduces noise and focuses on meaningful content. For instance, from the tokenized list ["The", "quick", "brown", "fox"], the word "The" would be removed.

- **Lemmatization** This involves reducing words to their base or root form (e.g., "running" to "run") to ensure that variations of a word are treated as a single item. This step is performed using libraries such as NLTK or spaCy in Python.

4.3.2.3 Vectorization

After preprocessing, the textual data is converted into numerical representations using the Term Frequency-Inverse Document Frequency (TF-IDF) method. TF-IDF vectorization involves the following steps:

- **Term Frequency (TF):** This measures the frequency of a word in a document. The term frequency $TF(t, d)$ is calculated as:

$$TF(t, d) = \frac{f_t}{N_d}$$

  where f_t is the number of times term t appears in document d, and N_d is the total number of terms in document d.

- **Inverse Document Frequency (IDF):** This measures how important a word is to a document in a corpus. The inverse document frequency $IDF(t)$ is calculated as:

$$IDF(t) = \log\left(\frac{N}{n_t}\right)$$

  where N is the total number of documents, and n_t is the number of documents containing the term t.

- **TF-IDF Calculation:** The TF-IDF value is computed by multiplying the term frequency and inverse document frequency:

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

This results in a vector representation of each document, which is used for similarity calculations.

4.3.2.4 Jensen-Shannon Divergence Calculation

Jensen-Shannon Divergence (JSD) measures the similarity between the TF-IDF vectors of two documents. The steps to compute JSD are as follows:

1. **Compute the Average Distribution M:** Given two probability distributions P and Q, the average distribution M is calculated as:

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

2. **Calculate Kullback-Leibler Divergence (KLD):** The KLD between P and M, and Q and M, is computed as:

$$KLD(P||M) = \sum_i P(i) \log \frac{P(i)}{M(i)}$$

$$KLD(Q||M) = \sum_i Q(i) \log \frac{Q(i)}{M(i)}$$

3. **Compute JSD:** Finally, the JSD is calculated as:

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

4.3.2.5 Algorithm in Pseudocode for JSD

The following pseudocode outlines the steps for calculating JSD between two documents:

Algorithm 2 Jensen-Shannon Divergence (JSD) Calculation

Require: Two documents, *Doc1* and *Doc2*
Ensure: Jensen-Shannon Divergence (*JSD*)

Initialization:
Initialize empty vectors P and Q for the TF-IDF values of *Doc1* and *Doc2*.
Set *threshold* for JSD similarity (optional, for further decision-making).

Step 1: Preprocessing
Tokenize *Doc1* and *Doc2* into words.
Remove stop-words and apply lemmatization to each token in both documents.

Step 2: Vectorization
Convert the preprocessed *Doc1* and *Doc2* into TF-IDF vectors, resulting in probability distributions P and Q.
Normalize P and Q so that the sum of their elements equals 1.

Step 3: Compute Average Distribution
Calculate the average distribution:

$$M \leftarrow \frac{1}{2}(P + Q)$$

Step 4: Calculate Kullback-Leibler Divergence (KLD)
Compute KLD for P relative to M:

$$KLD(P||M) \leftarrow \sum_i P(i) \log \frac{P(i)}{M(i)}$$

Compute KLD for Q relative to M:

$$KLD(Q||M) \leftarrow \sum_i Q(i) \log \frac{Q(i)}{M(i)}$$

Step 5: Calculate Jensen-Shannon Divergence (JSD)
Compute the JSD:

$$JSD(P||Q) \leftarrow \frac{1}{2}KLD(P||M) + \frac{1}{2}KLD(Q||M)$$

Step 6: Decision (Optional)
if $JSD(P||Q)$ ¡ *threshold* **then**
    Output: Documents are considered similar.
else
    Output: Documents are not considered similar.
end if

Step 7: Return the JSD Value
return $JSD(P||Q)$ as the measure of similarity between *Doc1* and *Doc2*.

Explanation of the Algorithm 2

The algorithm calculates the Jensen-Shannon Divergence (JSD) between two documents to measure their similarity. It follows a systematic process broken down into seven steps:

Initialization The algorithm begins by initializing empty vectors P and Q to store the TF-IDF values for the two documents. Additionally, an optional similarity threshold can be set to decide the similarity of the documents at a later stage.

Step 1: Preprocessing In this step, the documents undergo preprocessing to prepare them for analysis. Tokenization breaks the documents into individual words, after which stopwords are removed. Lemmatization reduces words to base forms, making the text consistent for further analysis.

Step 2: Vectorization The preprocessed documents are then converted into TF-IDF vectors, representing each document as a probability distribution over its vocabulary. These vectors are normalized so that the sum of their elements equals 1, ensuring they can be treated as probability distributions.

Step 3: Compute Average Distribution Here, the algorithm calculates the average distribution M by taking the midpoint of the two probability distributions P and Q. This distribution M serves as a reference point for computing the divergence of each document from this average.

Step 4: Calculate Kullback-Leibler Divergence (KLD) The algorithm computes the Kullback-Leibler Divergence for both P relative to M and Q relative to M. The KLD quantifies the difference between the original distributions (P and Q) and the average distribution (M).

Step 5: Calculate Jensen-Shannon Divergence (JSD) Using the KLD values calculated in Step 4, the algorithm computes the JSD. This step averages the two KLD values to obtain a symmetric and bounded measure of divergence, ensuring that $JSD(P||Q) = JSD(Q||P)$. The resulting JSD value ranges between 0 (identical distributions) and 1 (completely different distributions).

Step 6: Decision (Optional) The algorithm can decide whether the documents are similar if a threshold is set during initialization. If the JSD is less than the threshold, the documents are considered similar; otherwise, they are not. This decision-making step is useful in applications with a clear cutoff to classify documents as similar or dissimilar.

Step 7: Return the JSD Value Finally, the algorithm returns the computed JSD value, providing a numerical representation of the similarity between the two documents. Lower JSD values indicate higher similarity, making this algorithm valuable for document clustering, information retrieval, and academic paper recommendation systems.

This detailed algorithm effectively combines natural language processing (NLP) techniques and statistical measures to quantify document similarity, offering a robust tool for text analysis in various fields.

4.3.3 Results and Discussion of Validation of Dataset-2

The validation of Dataset-2, generated through Jensen-Shannon Divergence (JSD), aimed to confirm its reliability by comparing it against a gold-standard reference: Dataset-1. As Dataset-1 is a manually annotated collection, it provides high-quality ground truth for assessing the performance of the JSD-based clustering method used for Dataset-2.

To validate the structure of Dataset-2 indirectly, applied the same JSD-based clustering technique on Dataset-1. This allowed us to measure the correlation

between the bibliographic coupling relationships identified by JSD and the manually annotated relationships within Dataset-1. There is strong correlation approximately 0.90 between JSD and Mannualy annoyed dataset, affirming that the JSD-based clustering aligns well with human expert annotations.

Given these results, it can confidently conclude that the JSD-based structure of Dataset-2 is reliable for bibliographic coupling analysis. Since the clustering approach aligns closely with the manually curated relationships in Dataset-1, Dataset-2 can be utilized for further bibliometric research and related studies with high confidence. This validation confirms the quality of Dataset-2 and highlights the efficacy of JSD as a clustering method for bibliographic coupling tasks, consistent with prior findings from Habib and Afzal [3].

Given the consistency across datasets, researchers can leverage Dataset-2 for bibliometric tasks such as citation analysis, thematic clustering, or academic recommendation systems. The minimal variance between the two datasets ensures that Dataset-2 can be reliably used in studies where manually annotated datasets are not feasible, thereby extending the applicability of JSD-based clustering for large-scale bibliometric analyses. The discussion could include practical implications of using Jensen-Shannon Divergence (JSD) in real-world scenarios, highlighting its impact on research workflows and information retrieval systems. JSD's superior performance enhances the relevance of search results by providing a nuanced similarity measure, which directly improves the ranking and recommendation quality. For information retrieval systems, this ensures that users receive highly relevant documents, thereby enhancing user satisfaction and engagement.

Additionally, JSD is crucial in content-based recommendation systems, particularly in digital libraries and academic search engines, where precise recommendations are essential. By accurately identifying subtle similarities between research papers, JSD improves the efficiency of literature reviews and supports trend detection and knowledge discovery in research workflows. As a result, integrating JSD into information retrieval systems can streamline the research process, reduce cognitive load, and promote targeted exploration of scholarly content.

4.4 System Architecture to Find Dynamic Weights of Sections

This section details the methodology proposed to assign weights to the citations appearing in logical sections of two bibliographically coupled papers. Contemplation of IMRaD while assessing the potential of a citation in a particular context has already been proven useful by various studies [3]. Habib and Afzal [3] presented a section-wise weight assignment strategy to determine relatedness between two bibliographically coupled papers and achieved 5% improved results compared to those yielded without weight assignment. As discussed earlier, Habib and Afza [3] have assigned weights to the sections in a heuristic manner. On the contrary, the proposed method incorporates a novel deep learning neural network (ANN) based approach that utilizes a backpropagation algorithm to discover appropriate weights.

4.4.1 DNN: Deep Neural Network with Backpropagation

Several techniques are used to form decision-making models, such as machine learning (ML), Artificial Neural Networks (ANN), deep learning, and Backpropagation. The neural network mimics the functioning of the human brain, following the concepts of neurons to train the data to make decisions. The neural networks are considered a subset of machine learning, functioning similarly but varying in capabilities. Unlike ML models that require human guidance for improvement, deep learning models, such as Artificial Neural Networks (ANN), can independently sense the accuracy of predictions and proceed accordingly. Deep learning becomes essential when an extensive training process is required, achieved by increasing the number of hidden layers in the ANN to form a deep neural network.

The prime purpose of the proposed study is to discover the section-wise weights for the citations of bibliographically coupled papers. The ANN uses certain weights and biases to train the model in this context. The deep neural network was tuned

to exploit its weight assignment strategy for discovering section-wise weights of citations.

The two datasets, dataset-1 and dataset-2, were partitioned into training and testing datasets with a 70%-30% ratio. Initially, all papers were transformed into XML, followed by the identification of the sections and the computation of the bibliographic coupling strength of each paper with the entire dataset. This bibliographic coupling strength was further reduced to determine the section-wise bibliographic coupling strength, which served as input to the neural network.

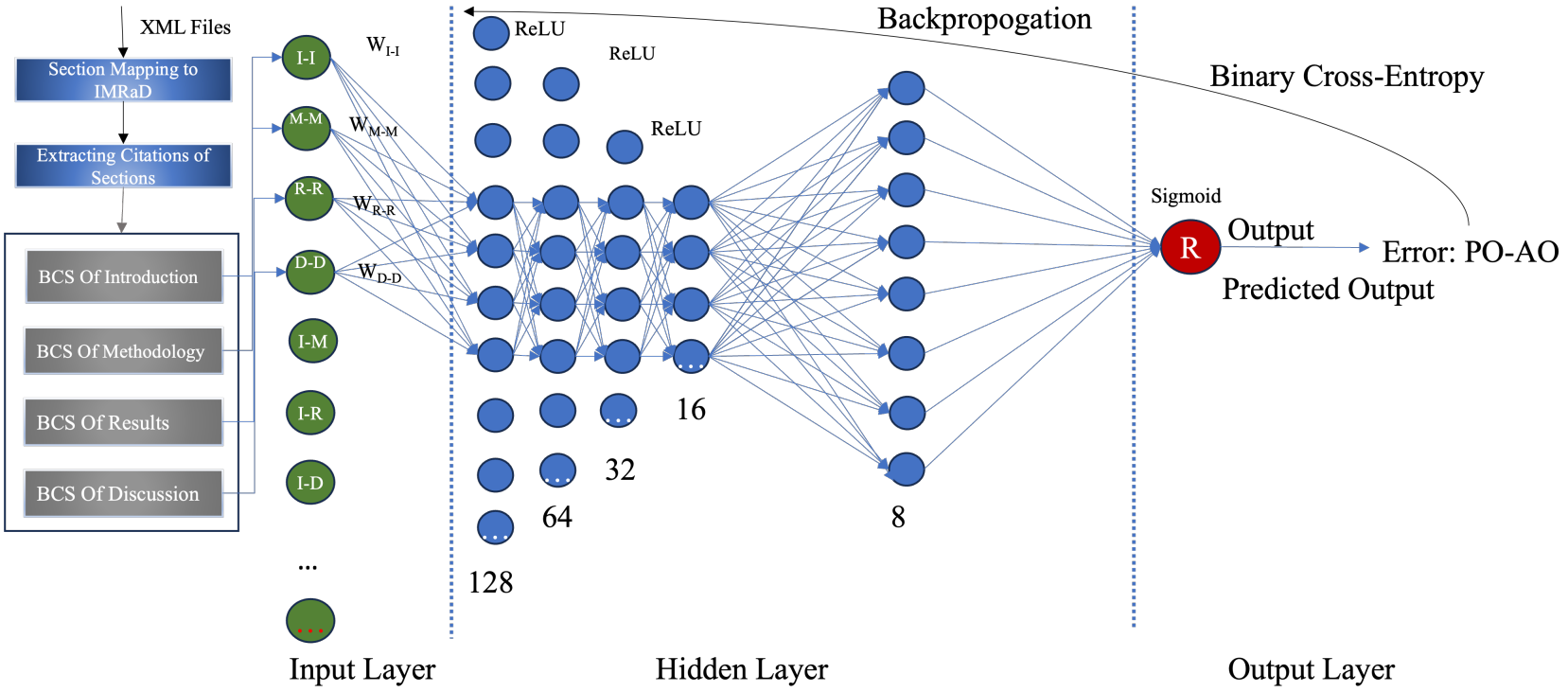


FIGURE 4.1: ANN with Backpropagation for Section Weight Tuning

Model Design and Training

Model Inputs

The proposed model leverages section-wise bibliographic coupling strengths (BCS) as input features. For each pair of papers A and B, the bibliographic coupling strengths for both same-section and cross-section pairs are considered.

Same-Section Coupling:

$$S_w \cdot II \quad \text{(Introduction-Introduction)}$$
$$S_w \cdot MM \quad \text{(Methodology-Methodology)}$$
$$S_w \cdot RR \quad \text{(Results-Results)}$$
$$S_w \cdot DD \quad \text{(Discussion-Discussion)}$$

Cross-Section Coupling:

$$S_w \cdot IM \quad \text{(Introduction-Methodology)}$$
$$S_w \cdot IR \quad \text{(Introduction-Results)}$$
$$S_w \cdot ID \quad \text{(Introduction-Discussion)}$$
$$S_w \cdot MR \quad \text{(Methodology-Results)}$$
$$S_w \cdot MD \quad \text{(Methodology-Discussion)}$$
$$S_w \cdot RD \quad \text{(Results-Discussion)}$$

Neural Network Architecture

The Deep Neural Network (DNN) with backpropagation is designed to process the section-wise BCS input features and predict the relatedness between the two papers. The architecture comprises the following layers:

Input Layer: The input layer accepts section-wise bibliographic coupling strengths and is normalized to ensure efficient processing.

Hidden Layers: The DNN includes multiple hidden layers to capture complex interactions:

- **Layer 1:** 128 neurons with ReLU activation.
- **Layer 2:** 64 neurons with ReLU activation.
- **Layer 3:** 32 neurons with ReLU activation.
- **Layer 4:** 16 neurons with ReLU activation.
- **Layer 5:** 8 neurons with ReLU activation.

These layers progressively condense the input features into meaningful representations.

Output Layer: The output layer consists of a single neuron with a sigmoid activation function. This outputs a probability score between 0 and 1, representing the likelihood that the two papers are related.

Training and Backpropagation

The training process involves the following steps:

1. **Forward Propagation:** Input data is passed through the network, and predictions are generated.

2. **Loss Calculation:** The Binary Cross-Entropy loss function is used to compute the error between predicted and actual values.

3. **Backpropagation:** Gradients of the loss function with respect to weights are computed using the chain rule.

4. **Weight Updates:** Weights are updated using the Adam optimizer to minimize the loss.

Backpropagation Update Rule

The weights are updated iteratively using the following rule:

$$w^{(i)} \leftarrow w^{(i)} - \alpha \frac{\partial L}{\partial w^{(i)}}$$

where:

- $w^{(i)}$: Weight at layer i.
- α: Learning rate.
- L: Loss function.
- $\frac{\partial L}{\partial w^{(i)}}$: Gradient of the loss with respect to $w^{(i)}$.

Iterative Weight Adjustment Algorithm

The algorithm for refining section weights during training is presented in Algorithm 3.

In conclusion the proposed DNN effectively learns the relationships between section-wise bibliographic coupling strengths of papers.

By leveraging backpropagation, the model dynamically adjusts section weights to optimize the relatedness prediction, enhancing the accuracy of bibliographic coupling analysis.

Algorithm 3 Refining Section Weights for Enhanced Bibliographic Coupling Analysis

Require: Section-wise bibliographic coupling strengths for each paper pair (A, B)
Ensure: Refined section weights $w_{\text{Intro}}, w_{\text{Meth}}, w_{\text{Res}}, w_{\text{Disc}}$

```
Initialize weights: w_Intro, w_Meth, w_Res, w_Disc ← Equal values
iter ← 0, max_iter ← Maximum allowed iterations
while iter < max_iter do
    for all paper pairs (A, B) do
        Compute predicted relatedness: r̂ ← f(w, A, B)
        Calculate actual relatedness: r
        Compute error: err ← |r̂ − r|
        if err > Threshold then
            Update weights: w ← w − α ∂L/∂w
        end if
    end for
    iter ← iter + 1
end while
```

4.5 Results and Evaluations

To elucidate the outcomes and analytical insights derived from applying the proposed methodology. The evaluation leveraged two distinct datasets: "Dataset-1," a manually annotated dataset for verifying the technique, and "Dataset-2," sourced from Habib and Afzal [3], comprising 5000 academic papers. The application of the methodology on "Dataset-2" yielded a correlation coefficient of 0.84, surpassing the 0.77 coefficient reported by Habib and Afzal [3]. Similarly, "Dataset-1" demonstrated correlation coefficients of 0.83 and 0.85, indicating the efficacy of the proposed approach in capturing the relational dynamics within the data.

Significantly, the methodology employs a deep neural network, fine-tuned over several epochs, to ascertain the optimal weights for section mapping to the IMRaD structure. Unlike the arbitrary weight assignments by Habib and Afzal [3], the proposed system systematically derives weights, normalizing them within the 0 to 1 range for coherence and comparability. Tables 4.2 and 4.3 delineate the weights discerned for the datasets under study, showcasing minor variations across datasets but maintaining correlation coefficients within the 0.83 to 0.85 range. This consistency underscores the methodological robustness and adaptability of the proposed system across diverse datasets.

The analysis also explores section-specific weights and their implications for bibliographic coupling in academic literature. When two papers are bibliographically coupled, common citations within identical sections (e.g., Introduction, Methodology, Results, Discussion) suggest thematic and methodological congruence. This scenario is addressed through the weights presented in Table 4.2. Conversely, cross-sectional citation patterns, where common citations span different sections across the coupled papers, necessitate a separate weighting schema, as illustrated in Tables 4.2 and 4.3 for "Dataset-1" and "Dataset-2," respectively.

The slight variation in weights across datasets, without significant deviation, suggests the proposed methodology's scalability and applicability across varied research themes. Figure 4.4 and 4.5, alongside comparative analysis tables 4.4 and 4.5, affirm the methodological merit and the consistent performance of the system. This evidence collectively advocates for the proposed approach as a viable and scalable solution for enhancing bibliographic coupling analysis and section mapping in academic research. The methodological application reveals a systematic and intelligent approach to determining section-specific weights within bibliographic coupling. The derived insights validate the proposed methodology's effectiveness and highlight its potential for broad-scale implementation in academic research analysis. Through strategic weight allocation and deep neural network optimization, the system significantly improves correlation scores, offering a refined tool for scholarly article evaluation and classification.

4.6 Complete Example: JSD and Spearman's Correlation

4.6.1 Step 1: Data Preparation

The normalized weights and common references for Papers A and B are in table 4.1:

TABLE 4.1: IMRaD Sections Normalized Weights

Section	Normalized Weight	Paper A	Paper B
Introduction	1.00	4	3
Methodology	1.82	6	5
Results	2.73	5	4
Discussion	3.55	2	3

4.6.2 Step 2: Weighted Common References

The weighted common references are calculated as:

Weighted Common References = Normalized Weight * Common Reference Count

For Paper A

$$\begin{aligned} \text{Introduction:} &\quad 1.00 \times 4 = 4.00 \\ \text{Methodology:} &\quad 1.82 \times 6 = 10.92 \\ \text{Results:} &\quad 2.73 \times 5 = 13.65 \\ \text{Discussion:} &\quad 3.55 \times 2 = 7.10 \end{aligned}$$

For Paper B

$$\begin{aligned} \text{Introduction:} &\quad 1.00 \times 3 = 3.00 \\ \text{Methodology:} &\quad 1.82 \times 5 = 9.10 \\ \text{Results:} &\quad 2.73 \times 4 = 10.92 \\ \text{Discussion:} &\quad 3.55 \times 3 = 10.65 \end{aligned}$$

4.6.3 Step 3: Probabilities

Convert the weighted references into probabilities by dividing each section's weight by the total weighted references.

Total Weighted References for Paper A: $4.00 + 10.92 + 13.65 + 7.10 = 35.67$

Total Weighted References for Paper B: $3.00 + 9.10 + 10.92 + 10.65 = 33.67$

$$P_A = \left[\frac{4.00}{35.67}, \frac{10.92}{35.67}, \frac{13.65}{35.67}, \frac{7.10}{35.67}\right] = [0.112, 0.306, 0.383, 0.199]$$

$$P_B = \left[\frac{3.00}{33.67}, \frac{9.10}{33.67}, \frac{10.92}{33.67}, \frac{10.65}{33.67}\right] = [0.089, 0.270, 0.324, 0.317]$$

4.6.4 Step 4: Jensen-Shannon Divergence (JSD)

Calculate the mean distribution:

$$M = \frac{P_A + P_B}{2} = \left[\frac{0.112 + 0.089}{2}, \frac{0.306 + 0.270}{2}, \frac{0.383 + 0.324}{2}, \frac{0.199 + 0.317}{2}\right]$$

$$M = [0.1005, 0.288, 0.3535, 0.258]$$

Compute the Kullback-Leibler divergence for P_A and P_B:

$$\mathrm{KL}(P_A||M) = \sum P_A \log \frac{P_A}{M}$$

$$\mathrm{KL}(P_A||M) = 0.112 \log \frac{0.112}{0.1005} + 0.306 \log \frac{0.306}{0.288} + 0.383 \log \frac{0.383}{0.3535} + 0.199 \log \frac{0.199}{0.258}$$

$$\mathrm{KL}(P_A||M) = 0.012 + 0.018 + 0.030 - 0.054 = 0.006$$

$$\mathrm{KL}(P_B||M) = \sum P_B \log \frac{P_B}{M}$$

$$\mathrm{KL}(P_B||M) = 0.089 \log \frac{0.089}{0.1005} + 0.270 \log \frac{0.270}{0.288} + 0.324 \log \frac{0.324}{0.3535} + 0.317 \log \frac{0.317}{0.258}$$

$$\mathrm{KL}(P_B||M) = -0.011 - 0.017 - 0.027 + 0.065 = 0.010$$

The JSD is:

$$\text{JSD}(P_A||P_B) = \frac{1}{2}\text{KL}(P_A||M) + \frac{1}{2}\text{KL}(P_B||M)$$

$$\text{JSD}(P_A||P_B) = \frac{1}{2}(0.006 + 0.010) = 0.008$$

4.6.5 Step 5: Spearman's Rank Correlation

The JSD measures the similarity between the distributions of common references as 0.008, indicating high similarity. The Spearman's correlation coefficient is 1, indicating perfect agreement in rankings.

TABLE 4.2: The section's Ranking and Weights - (Dataset-1)

IMRaD Sections	**Weights**
Introduction	0.10
Methodology	0.23
Results	0.31
Discussion	0.36

TABLE 4.3: The section's Ranking and Weights - (Dataset-2)

IMRaD Sections	**Weights**
Introduction	0.11
Methodology	0.20
Results	0.30
Discussion	0.39

TABLE 4.4: Cross Sections Weights Paper A and Paper B - (Dataset-1)

Sections of A/B	**I**	**M**	**R**	**D**
Introduction	0.10	0.18	0.09	0.03
Methodology	0.10	0.23	0.09	0.07
Results	0.05	0.07	0.31	0.08
Discussion	0.20	0.10	0.40	0.36

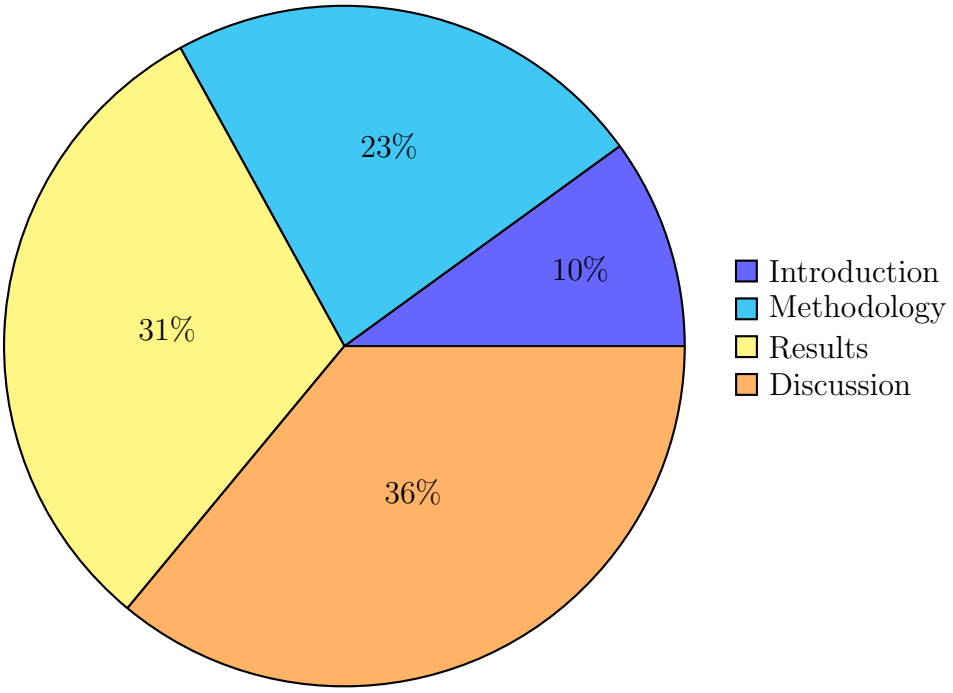


FIGURE 4.2: IMRaD Sections Weights (Dataset-1)

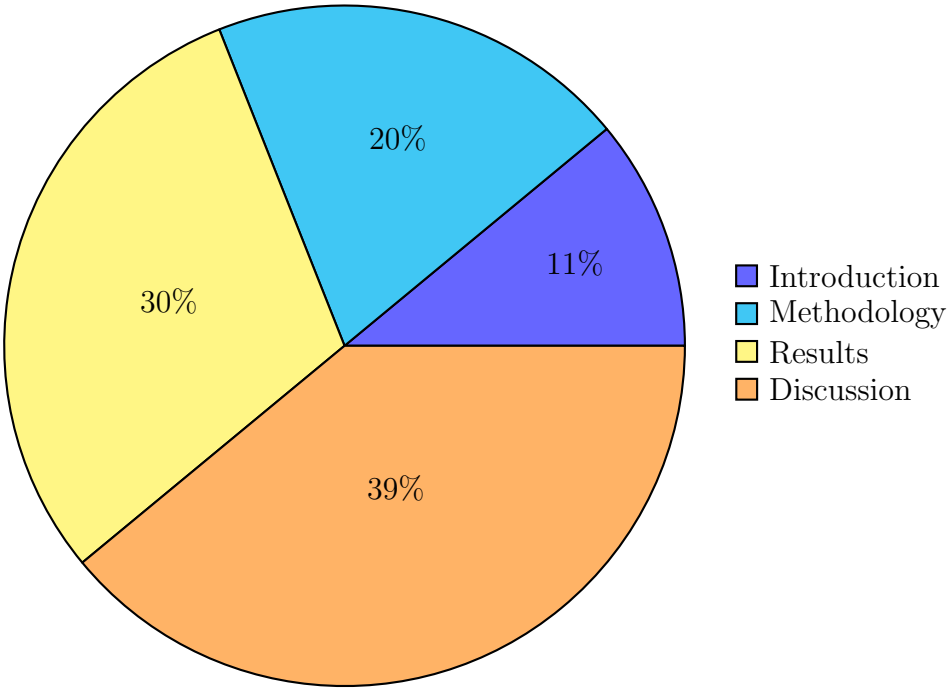


FIGURE 4.3: IMRaD Sections Weights (Dataset-2)

TABLE 4.5: Cross Sections Weights Paper A and Paper B - (Dataset-2)

Sections of A/B	**I**	**M**	**R**	**D**
Introduction	0.11	0.19	0.03	0.01
Methodology	0.10	0.23	0.09	0.07
Results	0.05	0.09	0.30	0.02
Discussion	0.11	0.12	0.14	0.39

TABLE 4.6: Correlation (%) for Dataset-1

Method	**Correlation (%)**
Proposed	86
Habib	75
BC	72
CBF	71

TABLE 4.7: Correlation (%) for Dataset-2

Method	**Correlation (%)**
Proposed	84
Habib	77
BC	71
CBF	73

4.7 Conclusion

The study aimed to improve the process of identifying relevant research papers. An Artificial Neural Network (ANN) with a backpropagation algorithm was utilized to assign weights to the different sections of scholarly articles, such as Introduction, Methods, Results, and Discussion (IMRaD).

The findings revealed insights into the importance of different sections within research papers. In Dataset-1, the Discussion section was assigned the highest weight, followed by Results, Methods, and Introduction. This reflects the common understanding that the Discussion and Results sections often contain crucial insights. Dataset 2 showed a similar pattern, reinforcing the consistency of the approach across different datasets. Comparatively, the accuracy of the proposed method demonstrated a significant improvement over traditional bibliographic coupling

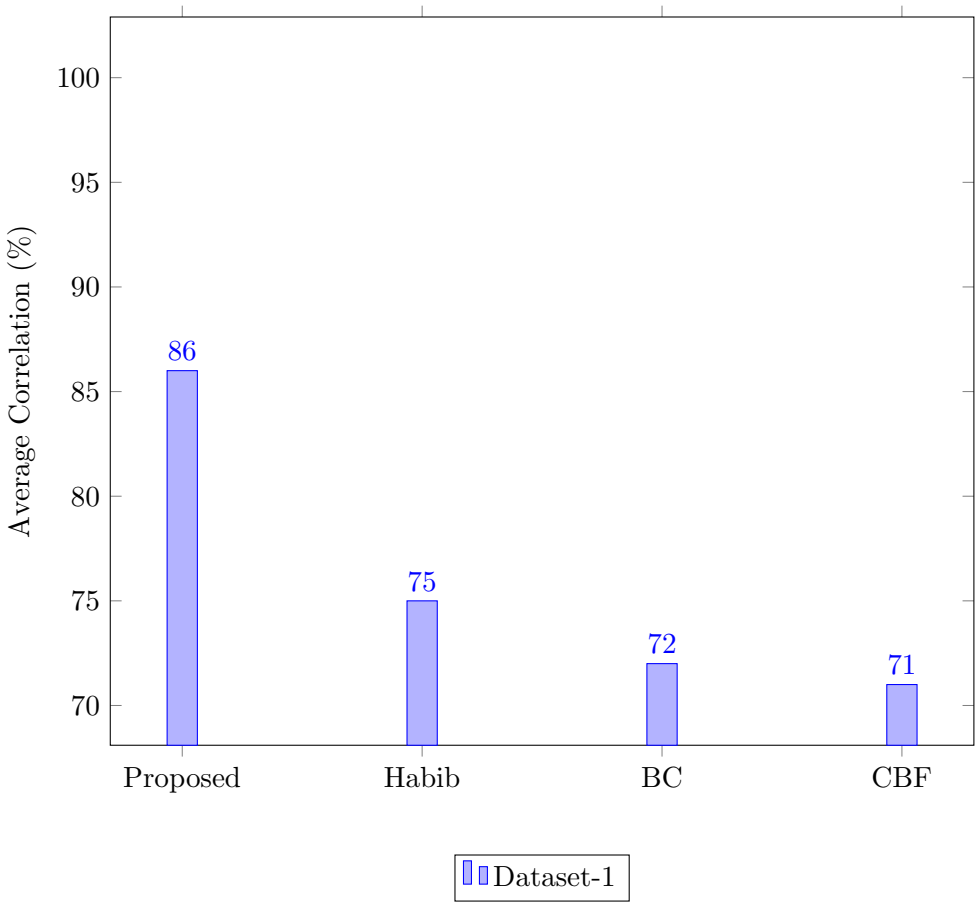


FIGURE 4.4: Results: Comparison of correlation for dataset-1

methods. Specifically, the correlation scores obtained from analysis—0.83 to 0.85 for Dataset-1 and 0.84 for Dataset-2—markedly surpass the benchmark set by Habib and Afzal, which was previously reported at 0.77. This enhancement in accuracy validates the efficacy of incorporating dynamic section weights and highlights the potential of the proposed methodology to revolutionize the identification of relevant research papers.

- Employing an Artificial Neural Network (ANN) with a backpropagation algorithm to assign weights to different sections of scholarly articles, such as Introduction, Methods, Results, and Discussion (IMRaD), the research demonstrates significant improvements over traditional methods.

- The study introduces a dynamically assigning weights to sections to enhance the accuracy of bibliographic coupling in identifying relevant research papers.

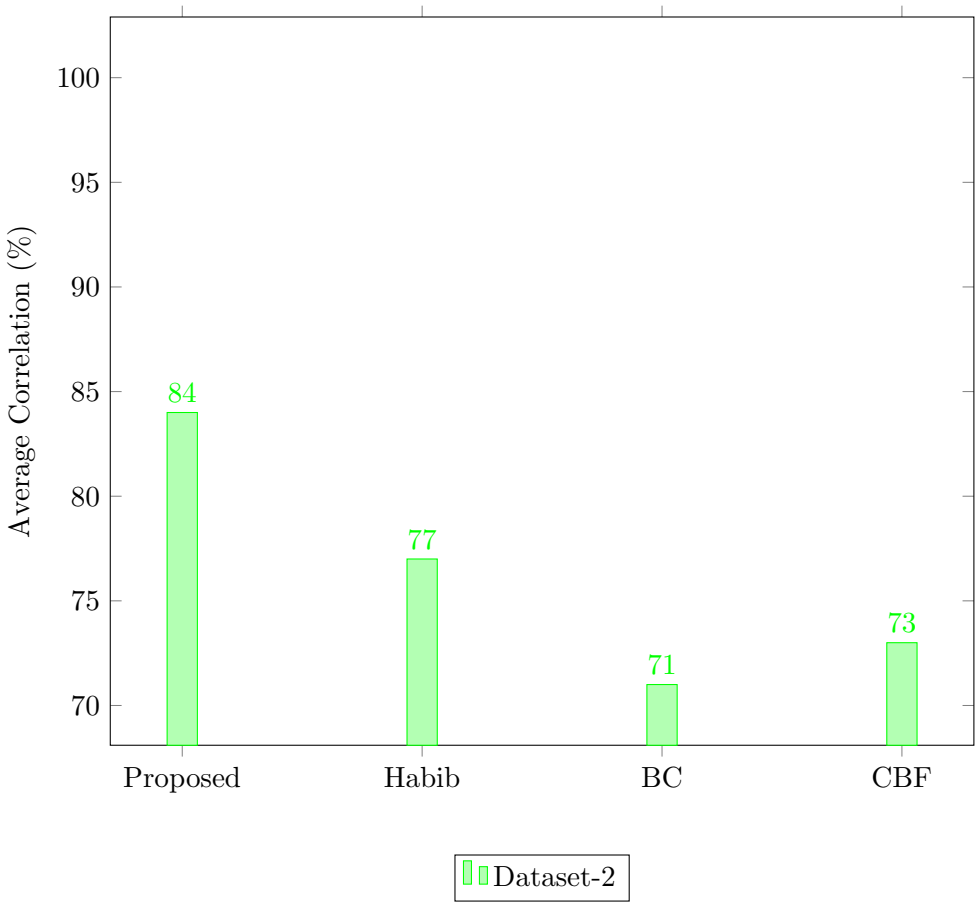


FIGURE 4.5: Results: Comparison of correlation for dataset-2

- The findings reveal that certain sections, like the Discussion, hold importance, reflecting the inherent value of scholarly contributions.
- Moreover, the methodology's adaptability and ability to integrate with existing bibliographic coupling approaches offer robust tools for researchers across various domains, promising effective and efficient exploration of scholarly literature

4.7.1 Limitations and Challenges

Despite the positive outcomes, this study encountered several limitations and challenges that warrant discussion:

- **Dataset Limitations** Dataset-2, which relies on automated clustering using JSD, lacks manual annotation, introducing potential biases in the generated

clusters. Although validation was conducted by comparing with Dataset-1, further refinement and validation of Dataset-2 are required to ensure its robustness.

- **Scalability Issues** While the proposed deep learning model performed well on the available datasets, its computational requirements may increase significantly with larger datasets.
  Training deep neural networks demands substantial hardware resources, including high-performance GPUs, which may limit the system's applicability in resource-constrained environments.

- **Section-Level Granularity** The proposed model assumes uniform importance for each section across all research papers. However, the significance of specific sections may vary depending on the research domain. Future work could explore context-aware models that dynamically adjust section weights based on the type or field of the research paper.

- **Generalizability to Other Formats** The methodology is based on the IMRaD structure, which is prevalent but not universally used across all disciplines. Papers in humanities or interdisciplinary research often deviate from the IMRaD format, posing challenges to this model's generalizability.

- **JSD Assumptions and Limitations** Although JSD effectively captures thematic similarities, it assumes that the text distributions are well-represented by TF-IDF vectors. In cases where the vocabulary distribution is sparse or inconsistent, JSD may not perform optimally.

  Future work could explore alternative similarity measures or hybrid approaches.

- **Impact of Hyperparameter Tuning:** The neural network's performance heavily depends on hyperparameters such as learning rate, activation functions, and the number of hidden layers.
  While the model was fine-tuned for the current datasets, hyperparameter optimization for different datasets and domains remains challenging.

4.8 Summary

This chapter introduces a novel system designed to enhance the discovery of bibliographically coupled papers by dynamically adjusting section weights. The research directly addresses the question of maximizing the correlation between bibliographic coupling strength and paper relatedness, leveraging both deep learning neural networks and Jensen-Shannon Divergence (JSD) for validation and clustering.

The chapter first validates two datasets: Dataset-1, which serves as a manually annotated "ground truth," and Dataset-2, generated using JSD. The validation process emphasizes the necessity of confirming Dataset-2's accuracy due to its automated generation. This establishes the foundation for using JSD as a reliable method for identifying thematic relationships in academic papers.

JSD's role in this system is crucial. By measuring the similarity between the probability distributions of textual content, JSD accurately clusters related papers. The chapter provides a detailed algorithm for calculating JSD, illustrating how it can efficiently uncover relationships based on word distributions. For example, in academic research databases, this clustering enables precise literature reviews by filtering relevant studies based on thematic content, ultimately streamlining researchers' workflows. Additionally, JSD's application in citation management software helps identify papers that may not be immediately apparent as related, offering researchers a broader view of the literature landscape.

The chapter then outlines the system architecture of the proposed method, which employs a deep neural network (DNN) with backpropagation to assign and refine section weights. Unlike previous heuristic methods, the neural network systematically learns the importance of sections such as Introduction, Methodology, Results, and Discussion. For instance, in a real-world application, this approach could be used to prioritize the discussion sections of research papers when suggesting references for a review article, as these sections often contain crucial insights. Through extensive evaluation, the system's performance demonstrates a high correlation between the predicted paper relatedness and bibliographic coupling strength, with accuracy scores surpassing traditional bibliographic coupling methods. The

study's results indicate that using dynamic section weights significantly enhances the identification of related research papers. This improvement has direct practical implications: academic search engines can utilize the weighted sections to provide targeted search results, improving the accuracy of paper recommendations for researchers. Similarly, in personalized recommendation systems for academic libraries, these weights can guide the system to suggest literature aligned with specific research interests.

4.8.1 Applications of the Proposed Method

While the summary mentions potential implications, this section provides specific examples of real-world scenarios where the proposed methodology can be applied:

- **Academic Search Engines** Search platforms such as Google Scholar or IEEE Xplore can leverage section-weighted bibliographic coupling to provide accurate search results.
  For instance, if the system detects a stronger coupling in the Discussion section, it can prioritize papers with similar conclusions, aiding researchers in quickly finding relevant studies.

- **Citation Recommendation Tools** Tools like Zotero and EndNote can utilize the dynamic weight model to suggest citations accurately based on the content section. For example, if a user is working on the Methodology section of their paper, the system can recommend research papers with similar methodologies, improving citation relevance.

- **Systematic Reviews and Meta-Analyses** The methodology can assist researchers in conducting systematic reviews by automatically identifying papers with similar methodologies or results. This ensures the inclusion of relevant studies and speeds up the literature review process.

- **Academic Libraries and Institutional Repositories** University libraries can use the proposed method to create thematic clusters of publications from

their repositories. For example, clustering dissertations or theses by weighted sections allows for better organization, helping students find relevant research efficiently.

- **Interdisciplinary Research Tools** Since interdisciplinary research often involves connecting papers from different fields, the dynamic section-weighting system can prioritize meaningful connections across disciplines. For example, it can link papers based on shared methodologies or similar results, even if they belong to different domains.

- **Research Grant Applications and Reports** Research institutions could use the system to identify trends in specific research areas. For example, clustering previous papers by weighted sections can reveal the most impactful methods or findings, supporting data-driven decisions for grant applications or policy recommendations.

- **Personalized Recommendation Systems** Personalized recommendation engines, embedded within academic platforms or learning management systems, could benefit from section-wise weight adjustment.

  For example, the system could suggest relevant papers for PhD students or early-career researchers based on their ongoing work, such as finding studies with complementary results or methodologies.

In conclusion, the chapter presents a comprehensive approach that combines the strengths of JSD and deep learning for bibliographic coupling analysis. By incorporating section weight adjustment into the clustering process, the system offers an innovative tool that not only improves the accuracy of related paper identification but also has the potential to revolutionize academic literature exploration, making research accessible and interconnected for scholars.

Chapter 5

Overall Conclusion and Future Work

5.1 Overview

This chapter concludes the dissertation's key findings. It emphasizes accurately mapping IMRaD sections (Introduction, Methods, Results, and Discussion) and later focuses on identifying and weighting research paper sections to enhance bibliographic coupling. The chapter also discusses the results and compares them with previous research. The paper addresses the issues found during the research process and opens opportunities for future research to refine methods for finding related papers.

5.2 Conclusion

In this dissertation, it has been argued that logical sections of research papers are crucial in finding related papers through various information retrieval techniques, such as content-based, bibliographic coupling, and co-citation-based approaches. Accurately identifying and mapping these sections to the IMRaD (Introduction, Methods, Results, and Discussion) structure is a foundational step in bibliometric

analysis. Additionally, since each section holds specific importance, assigning appropriate weights is critical to enhancing the effectiveness of related paper identification methods. This research primarily focused on accurately identifying these sections and assigning appropriate weights, thus improving the precision of the bibliographic coupling approach.

This research addressed two major gaps identified through a critical review of the literature: (1) the low accuracy reported in mapping section headings to the IMRaD structure and (2) the arbitrary assignment of weights to sections in bibliographically coupled papers. These issues formed the basis for the following research questions:

- **RQ1**: How can a method be devised with improved accuracy to map the sections of research papers to the IMRaD structure, considering the variations of these sections?
- **RQ2**: How can sections' weights be tuned to maximize the correlation between Bibliographic Coupling strength and paper relatedness?

5.2.1 RQ1: Section Mapping to IMRaD

The first research question was addressed by developing a comprehensive methodology for accurately mapping the logical sections of research papers to the IMRaD structure. The methodology employed advanced techniques to achieve this goal, including PDF-to-XML conversion, feature extraction, and machine learning algorithms. The key elements of the approach were:

- Data Conversion: Research papers were converted from PDF to XML format to enable structured data processing. This conversion was crucial as it provided a standardized way to extract textual content and metadata.
- Feature Extraction The methodology focused on extracting potential features such as subheadings mapping, figures and tables count, and in-text citation frequency. These features were critical in identifying the unique characteristics

of different sections and mapping them to the corresponding logical sections in the IMRaD structure.

5.2.1.1 Results Comparison

To quantify the impact of the proposed methodology, a comparative analysis with existing approaches was conducted using precision, recall, and F-measure. Table 5.1 presents the results, highlighting the significant improvements achieved by the proposed method.

TABLE 5.1: Comparison of Methodologies for Section Mapping Accuracy

Methodology	**Precision**	**Recall**	**F-Measure**
Ding et al. [1]	0.87	0.81	0.81
Shahid et al. [2]	0.81	0.81	0.81
Habib et al. [3]	0.89	0.89	0.89
Proposed Method	**0.97**	**0.97**	**0.97**

Table 5.1 indicates that the proposed methodology significantly outperforms existing techniques for all three metrics. The precision and recall rates of 0.97 demonstrate the method's high accuracy and ability to effectively map sections in the IMRaD structure. This advancement addresses the earlier limitations in the literature, where false positives and missed section boundaries often led to incorrect mapping. By incorporating subheading mapping, figures and table counts, and citation frequency as features, the proposed method improved the precision by 8% and recall by 8% compared to the best existing approach.

5.2.1.2 Issues in PDF-to-XML Conversion and Annotation

Despite advancements in automated tools, several technical challenges arise during the PDF-to-XML conversion process, requiring manual corrections and interventions. This section outlines some key challenges encountered, especially when using tools like PDFX.

PDFX-Related Issues

PDFX and similar PDF-to-XML converters encounter various difficulties due to the inherently visual nature of PDF documents. Designed for presentation, PDFs often make extracting structured data into XML complex and prone to errors [100].

1. **Loss of Hierarchical Structure:** Nested headings, such as section titles and subsections, often become flattened during conversion, resulting in an inaccurate XML representation. This loss of hierarchy complicates downstream processing, such as document parsing.

2. **Handling Complex Layouts:** Converters struggle with layouts involving multiple columns, tables, or embedded images. Mismatches and broken sections are common in XML output, which affects the readability and usability of converted documents [100].

3. **Text Formatting Errors:** Fonts, special characters, and text styles are frequently misrepresented or lost during conversion. These errors require manual adjustments to accurately restore the original document's intent and structure.

4. **Graphical Data Extraction:** Extracting data from charts and images is challenging, often requiring manual intervention to align figures with their corresponding sections in XML. This limits the automation potential of tools like PDFX [100].

5. **Manual Corrections and Annotations:** Semi-structured PDFs demand manual corrections, particularly for complex academic documents. The limitations of PDFX highlight the need for careful validation to ensure reliable conversion and annotation [100].

In conclusion the enhanced accuracy achieved by the proposed method is evident when compared to the performance of existing techniques (Table 5.1). While Ding et al. [1] and Shahid et al. [2] relied on dictionary terms and templates for mapping, they struggled to capture the nuanced differences between sections, often resulting in misclassifications. The proposed method's integration of multiple discriminative

features, including the frequency of figures, tables, and in-text citations, allowed for a comprehensive analysis of section characteristics, significantly improving mapping accuracy.

5.2.2 RQ2: Optimizing Section Weights for Bibliographic Coupling

The second part of the study focused on addressing the problem of arbitrary weight assignments in bibliographically coupled papers. The research introduced an Artificial Neural Network (ANN) with a backpropagation algorithm to optimize these weights systematically. By learning from a large corpus of annotated data, the ANN could fine-tune the importance of different sections about paper relatedness. Table 5.2 and Table 5.3 summarize the accuracy improvements achieved through this approach.

TABLE 5.2: Correlation Scores for Different Recommendation Approaches (Dataset-1)

Approach	**Correlation (%)**
Proposed	86
Habib et al. [3]	75
Bibliographic Coupling (BC)	72
Content-Based Filtering (CBF)	71

TABLE 5.3: Correlation Scores for Different Recommendation Approaches (Dataset-2)

Approach	**Correlation (%)**
Proposed	84
Habib et al. [3]	77
Bibliographic Coupling (BC)	71
Content-Based Filtering (CBF)	73

The proposed approach achieved an correlation of 86% with Dataset-1, showing a significant improvement over the method by Habib et al. [3], which achieved 75%, and a 15% improvement over traditional Bibliographic Coupling (72%). Similarly, the method outperformed Content-Based Filtering (CBF), which achieved 71%. With Dataset-2, the proposed method maintained a high correlation score of

84%, outperforming Habib et al. [3] by 7%, who achieved 77%. The results also surpassed traditional Bibliographic Coupling (71%) and Content-Based Filtering (CBF) (73%).

These results highlight the effectiveness of the proposed machine-learning-based approach in dynamically adjusting section weights, maximizing the correlation between bibliographic coupling strength and paper relatedness across different datasets. The superior performance demonstrates the model's ability to better capture nuanced patterns in section-wise coupling, setting a new benchmark in academic paper recommendation systems.

5.3 Discussion

The results achieved in this study represent a significant advancement over existing methodologies in bibliometric analysis. The research addressed the long-standing issues of section misclassification and arbitrary weight assignments by combining structured data conversion, discriminative feature extraction, and machine learning.

One key contribution of the study is its focus on subheading mapping and integrating additional features, such as figures, tables, and citation counts, to improve the accuracy of section mapping.

This approach goes beyond the traditional reliance on static dictionary terms or templates, allowing for a nuanced understanding of the document's structure. Incorporating machine learning models further refined this process, enabling the system to learn complex patterns in section characteristics.

Despite these successes, the study also identified limitations. Although essential, the PDF-to-XML conversion process introduced potential errors due to the PDF files' semi-structured nature. This issue underscores the need for sophisticated document parsing algorithms to handle complex formats. Furthermore, the approach's dependency on bibliographic coupling limits its applicability to research papers

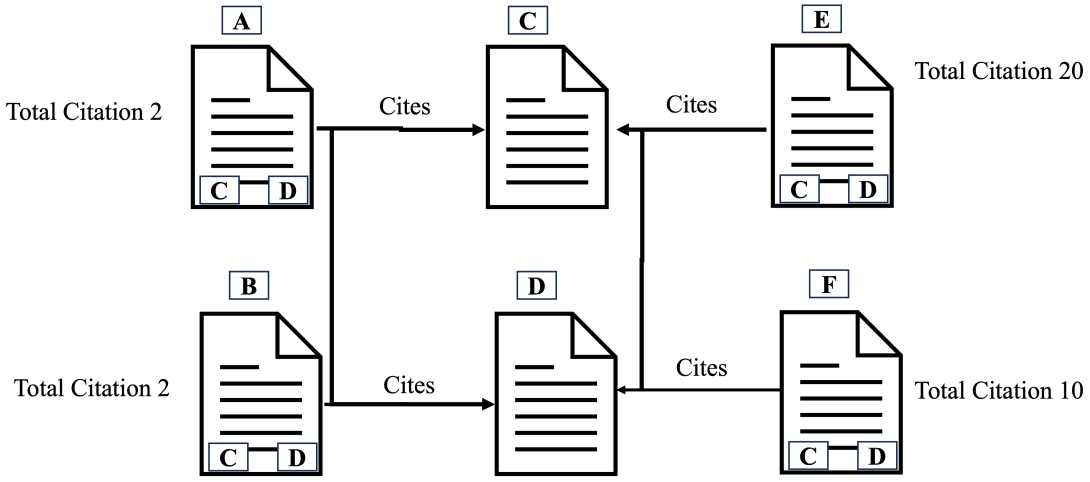


FIGURE 5.1: Influence of Citation Context on Bibliographic Coupling

with citation links. Future work should explore integrating co-citation analysis and content-based methods to broaden the methodology's scope.

5.4 Future Work

Building on the findings of this research, several avenues for future exploration are proposed:

- **Advanced Document Parsing:** Develop enhanced algorithms for accurately converting PDF files to XML, focusing on resolving issues related to complex document structures.
- **Dynamic Weight Adjustment:** Investigate using reinforcement learning models to adjust section weights in bibliographic coupling to adapt to changing research trends.
- **Cross-Disciplinary Validation:** Extend the methodology to other academic disciplines to assess its generalizability and adapt it for fields with different publication norms.

- **Integration with Co-Citation Analysis:** Explore the combination of bibliographic coupling and co-citation analysis to enhance the identification of related papers, potentially developing a hybrid IR model.

- **Influence of Citation Context on Bibliographic Coupling:** The relationship between two bibliographically coupled papers may be influenced by their bibliographic coupling strength and the total number of citations each paper has. For example:

  - **Case 1:** Paper E has 20 citations, Paper F has 10 citations, and their bibliographic coupling strength is 2 as both cites paper C and D as mentioned in Figure 5.1.

  - **Case 2:** Paper A has 2 citations, Paper B has 2 citations, and their bibliographic coupling strength is also 2 as both cites paper C and paper D as mentioned in figure 5.1.

  While the bibliographic coupling strength is the same, the context differs significantly. It can be checked in the future whether Case 2, with a higher proportion of overlapping references relative to total citations, indicates a stronger relationship compared to Case 1, where the overlap is a smaller fraction of the total references. This analysis can help refine the interpretation of bibliographic coupling strength in the context of citation distributions.

5.5 Summary

The chapter "Overall Conclusion and Future Work" encapsulates the culmination of the dissertation's findings and proposes avenues for future research. The dissertation's central argument revolves around the significance of logical section identification in research papers and the criticality of assigning appropriate weights to these sections to enhance the effectiveness of related paper identification methods, particularly in bibliographic coupling. Here's a detailed summary of the chapter: The dissertation begins by highlighting the importance of accurately identifying and mapping logical sections of research papers to the IMRaD structure. It emphasizes

the need for assigning weights to these sections to improve the efficacy of related paper identification methods, especially in bibliographic coupling.

A critical literature analysis identifies two primary research gaps: low accuracy in mapping section headings to the IMRaD structure and assigning arbitrary weights for tuning section weights as outlined in papers. The research questions (RQ1 and RQ2) are formulated based on these identified gaps. They focus on maximizing the correlation between bibliographic coupling strength and paper relatedness and collecting/preparing an annotated dataset for evaluation.

The dissertation presents a novel method for improving the accuracy of mapping logical sections onto the IMRaD structure. This method incorporates in-text citation, figure, and table count, resulting in significantly improved precision and recall compared to contemporary approaches. Additionally, a comprehensive methodology for tuning section weights is outlined to maximize the correlation between bibliographic coupling strength and paper relatedness. This methodology utilizes an Artificial Neural Network with a backpropagation algorithm, demonstrating substantial accuracy improvement in weight assignment.

Despite the significant contributions, the proposed approach has limitations, including dependency on PDF-to-XML conversion tools and applicability restricted to bibliographically coupled papers.

The chapter concludes with a comprehensive discussion of future research directions, focusing on enhancing conversion algorithms for section identification, optimizing section weights using ANN, broadening the scope of application across disciplines, extending the approach to co-cited papers, and leveraging big data and AI for bibliometric analysis.

Each future research direction is elaborated with specific strategies and potential advancements, aiming to address the identified limitations and push the boundaries of bibliometric analysis.

Bibliography

[1] C. G. Ying Ding, Xiaozhong Liu and B. Cronin, "The distribution of references across texts: Some implications for citation analysis," *Journal of Informetrics*, vol. 7, no. 3, pp. 583–592, July 2013.

[2] A. Shahid and M. T. Afzal, "Section-wise indexing and retrieval of research articles," *Cluster Computing*, vol. 21, no. 1, pp. 481–492, 2018. [Online]. Available: https://doi.org/10.1007/s10586-017-0914-4

[3] R. Habib and M. T. Afzal, "Sections-based bibliographic coupling for research paper recommendation," *Scientometrics*, vol. 119, no. 2, p. 643–656, May 2019. [Online]. Available: https://doi.org/10.1007/s11192-019-03053-8

[4] M. Khabsa and C. L. Giles, "The number of scholarly documents on the public web," *PLOS ONE*, vol. 9, no. 5, pp. 1–6, 05 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0093949

[5] E. Garfield, "Citation indexes for science: A new dimension in documentation through association of ideas," *Science*, vol. 122, no. 3159, pp. 108–111, 1955.

[6] M. M. Kessler, "Bibliographic coupling between scientific papers," *American Documentation*, vol. 14, pp. 10–25, 1963.

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

[8] B. Gipp and J. Beel, "Citation proximity analysis: A new method for improving the efficiency of citation-based recommendations," *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 387–390, 2009.

[9] M. W. Khan, A. Nawaz, M. Yousuf, and M. A. Javed, "Section-wise citation-similarity and its impact on citation based information retrieval," in *Proceedings of the International Conference on Soft Computing Systems and Information Security (ICSSI)*, 2019, pp. 250–261.

[10] 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.

[11] R. Habib and M. Afzal, "Paper recommendation using citation proximity in bibliographic coupling," *TURKISH JOURNAL OF ELECTRICAL ENGINEERING and COMPUTER SCIENCES*, vol. 25, pp. 2708–2718, 01 2017.

[12] A. Khan, A. Shahid, M. Afzal, F. Nazar, F. Alotaibi, and K. Alyoubi, "Swics: Section-wise in-text citation score," *IEEE Access*, vol. PP, pp. 1–1, 09 2019.

[13] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1983.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019, pp. 4171–4186.

[16] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[17] K. Hong, H. Jeon, and C. Jeon, "User-profile-based personalized research paper recommendation system," in *2012 8th International Conference on*

Computing and Networking Technology (INC, ICCIS and ICMIC). IEEE, 2012, pp. 134–138.

[18] M. Magara, S. Ojo, and T. Zuva, "Towards a serendipitous research paper recommender system using bisociative information networks (bisonets)," in *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*. IEEE, 2018, pp. 1–6.

[19] C. Nascimento, A. H. Laender, A. S. da Silva, and M. A. Gonçalves, "A source independent framework for research paper recommendation," in *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, ser. JCDL '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 297–306. [Online]. Available: https://doi.org/10.1145/1998076.1998132

[20] Q. He, D. Kifer, J. Pei, P. Mitra, and C. Giles, "Citation recommendation without author supervision," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. ACM, 2011, pp. 755–764.

[21] M. S. Pera and Y.-K. Ng, "A personalized recommendation system on scholarly publications," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ser. CIKM '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 2133–2136. [Online]. Available: https://doi.org/10.1145/2063576.2063908

[22] K. Sugiyama and M.-Y. Kan, "A comprehensive evaluation of scholarly paper recommendation using potential citation papers," *International Journal on Digital Libraries*, vol. 16, no. 2, pp. 91–109, June 2015. [Online]. Available: https://doi.org/10.1007/s00799-014-0122-2

[23] K. Neethukrishnan and K. Swaraj, "Ontology based research paper recommendation using personal ontology similarity method," in *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. IEEE, 2017, pp. 1–4.

[24] X. Liu, Y. Yu, C. Guo, Y. Sun, and L. Gao, "Full-text based context-rich heterogeneous network mining approach for citation recommendation," in *IEEE/ACM Joint Conference on Digital Libraries*. IEEE, 2014, pp. 361–370.

[25] 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, pp. 265–269, 1973.

[26] J. Son and S. Kim, "Academic paper recommender system using multilevel simultaneous citation networks," *Decision Support Systems*, vol. 105, 10 2017.

[27] 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.

[28] 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.

[29] 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.

[30] X. Y. Zhen Wang, "An improved co-citation recommendation method based on graph neural network for cold-start problem in academic paper recommendation," *Applied Intelligence*, vol. 51, no. 10, pp. 7002–7013, 2021.

[31] L. Egghe and R. Rousseau, "Co-citation, bibliographic coupling and a characterization of the h-index," *Journal of the American Society for Information Science and Technology*, vol. 53, no. 12, pp. 1105–1112, 2002.

[32] W. Glänzel and H. Czerwon, "A new methodological approach to bibliographic coupling and its application to the national, regional, and institutional level," *Scientometrics*, vol. 37, no. 2, pp. 195–221, 1996.

[33] E. Gündoğan and M. Kaya, "A novel hybrid paper recommendation system using deep learning," *Scientometrics*, vol. 127, pp. 3837–3855, 2022.

[34] R. Kumar, *Research Methodology: A Step-by-Step Guide for Beginners*, 3rd ed. SAGE Publications Ltd, 2011.

[35] S. L. Joeran Beel, Bela Gipp and C. Breitinger, "Research paper recommender systems: A literature survey," *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, 2013.

[36] N. P. Felice Ferrara and C. Tasso, "A keyphrase-based paper recommender system," in *Digital Libraries and Archives*, C. M. Maristella Agosti, Floriana Esposito and N. Orio, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 14–25.

[37] K. Sarkar, M. Nasipuri, and S. Ghose, "A new approach to keyphrase extraction using neural networks," *International Journal of Computer Science Issues (IJCSI)*, vol. 7, no. 2, March 2010, preprint available on arXiv. [Online]. Available: https://arxiv.org/abs/1004.3274

[38] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405959524000651

[39] L. Ajallouda, F. Z. Fagroud, Z. Ahmed, 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. [Online]. Available: https://doi.org/10.11591/eei.v12i1.4130

[40] Q. Liu, W. Ke, X. Yuan, Y. Yang, H. Zhao, and P. Wang, "Adaptiveuke: Towards adaptive unsupervised keyphrase extraction with gated topic modeling," *Expert Systems with Applications*, vol. 228, p. 123926, 2024.

[41] 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, pp. 1–11, 05 2021.

[42] H. I. Pohan, H. L. H. S. Warnars, B. Soewito, and F. Gaol, "Recommender system using transformer model: A systematic literature review," in *2022 1st International Conference on Information System & Information Technology (ICISIT)*, July 2022, pp. 1–7.

[43] K. Hong, H. Jeon, and C. Jeon, "Advanced personalized research paper recommendation system based on expanded userprofile through semantic analysis," *International Journal of Digital Content Technology and its Applications*, vol. 7, no. 15, pp. 67–76, 2013.

[44] Z. Huang, W. Chung *et al.*, "A hybrid recommendation system for book purchasing," in *Proceedings of the International Conference on Electronic Commerce (ICEC)*, 2002, pp. 139–146.

[45] Q. Guo *et al.*, "Improving academic paper discovery using heterogeneous information networks," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015, pp. 123–130.

[46] A. Kanakia, Z. Shen, D. Eide, and K. Wang, "A scalable hybrid research paper recommender system for microsoft academic," in *Proceedings of the 24th ACM International Conference on Knowledge Discovery and Data Mining*. ACM, 2019, pp. 1–8.

[47] A. M. Nair, O. Benny, and J. George, "Content-based scientific article recommendation system using deep learning technique," in *Inventive Systems and Control, Lecture Notes in Networks and Systems 204*, 2021, pp. 965–977. [Online]. Available: https://doi.org/10.1007/978-981-16-1395-1_70

[48] S.-A. Teh, S.-C. Haw, and H. A. Santoso, "Hybrid-based research article recommender system," *International Journal of Membrane Science and Technology*, vol. 10, no. 2, pp. 1587–1606, September 2023.

[49] C. Li, I. Ishak, H. Ibrahim, M. Zolkepli, and F. Sidi, "Deep learning-based recommendation system: Systematic review and classification," *IEEE Access*, vol. 11, pp. 113 790–113 811, 2023.

[50] A. Singhal, "Modern information retrieval: A brief overview," in *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2001, pp. 35–43.

[51] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2006.

[52] G. W. Morgan and J. A. Sager, *Elementary Statistics for Computer Applications*, 4th ed. Prentice-Hall, 1995.

[53] R. W. Hamming, "Error detecting and error correcting codes," *Bell System Technical Journal*, vol. 29, pp. 147–160, 1950.

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

[55] 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. PP, no. 99, pp. 1–1, 2021.

[56] W. Guo *et al.*, "A hybrid model for paper discovery using metadata and content analysis," *International Journal of Digital Libraries*, vol. 21, pp. 15–30, 2020.

[57] S. Bethard and D. Jurafsky, "Learning to rank scientific papers based on citation patterns," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 29–38.

[58] K. Shahid *et al.*, "In-text citations for related paper identification in scientific articles," *Journal of Digital Libraries*, vol. 25, pp. 54–63, 2009.

[59] M. Nassiri *et al.*, "Normalized similarity index for paper recommendation based on citation networks," in *Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR)*, 2010, pp. 213–220.

[60] M. Krapivin *et al.*, "Focused pagerank for paper recommendations in academic networks," *Journal of Digital Information Management*, vol. 9, pp. 44–52, 2011.

[61] M. Gori and A. Pucci, "Random walks on citation graphs for efficient paper ranking," *Journal of Machine Learning Research*, vol. 13, pp. 1103–1127, 2012.

[62] K. El-Arini and C. Guestrin, "Beyond keyword search: Discovering relevant scientific papers using influence flow," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2013, pp. 1484–1492.

[63] T. Strohman *et al.*, "A system for drafting and retrieving related research papers," *Journal of Digital Libraries*, vol. 9, no. 2, pp. 137–148, 2015.

[64] P. Reyhani *et al.*, "Simcc: A citation contribution-based paper similarity metric," in *Proceedings of the 2017 ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 2017, pp. 501–510.

[65] E. Bichteler *et al.*, "Enhancing research paper recommendations through bibliographic coupling and co-citation analysis," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2018, pp. 411–420.

[66] K. Haruna, M. A. Ismail, A. B. Baffa, V. Chang, S. Wibawa, and T. Herawan, "A citation-based recommender system for scholarly paper recommendation," in *Computational Science and Its Applications – ICCSA 2018*. Springer, 2018, pp. 514–525.

[67] A. Shahid, M. T. Afzal, A. Alharbi, H. Aljuaid, and S. Alotaibi, "In-text citation's frequencies-based recommendations of relevant research papers," *PeerJ Computer Science*, vol. 7, p. e524, 2021.

[68] 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.

[69] K. W. Takahiro Koseki and M. Yoshikawa, "A novel coupling strength measure based on shared references for improved citation analysis," *Journal of Information Science*, vol. 33, no. 6, pp. 765–775, 2007.

[70] R. Habib and M. T. Afzal, "Paper recommendation using citation proximity in bibliographic coupling," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 25, pp. 2708–2718, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:2402723

[71] F. G. M. Bordons and B. Fernandez, "Using bibliometric coupling to analyze collaboration patterns in research fields: A case study of nanotechnology," *Scientometrics*, vol. 120, pp. 785–801, 2019.

[72] Q. L. Wei Zhang and B. Zhao, "Temporal bibliographic coupling analysis for detecting emerging trends in research," *Journal of Scientometric Research*, vol. 10, no. 4, pp. 432–448, 2021.

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

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

[75] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," *The Adaptive Web*, vol. 4321, pp. 325–341, 2007.

[76] M. Balabanović and Y. Shoham, "Fab: Content-based, collaborative recommendation," in *Communications of the ACM*, vol. 40, no. 3, 1997, pp. 66–72.

[77] K. Lang, "Newsweeder: Learning to filter netnews," in *Proceedings of the Twelfth International Conference on Machine Learning (ICML)*, Tahoe City, CA, USA, 1995, pp. 331–339.

[78] G. W. Furnas, T. K. Landauer, R. L. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Communications of the ACM*, vol. 30, no. 11, pp. 964–971, 1987.

[79] S. McNee, J. A. Konstan, and J. Riedl, "Being accurate is not enough: How accuracy metrics have hurt recommender systems," in *Proceedings of the*

SIGCHI Conference on Human Factors in Computing Systems (CHI), 2002, pp. 109–116.

[80] L. Giles, K. Bollacker, and S. Lawrence, "Citeseer: An automatic citation indexing system," in *Proceedings of the Third ACM Conference on Digital Libraries*, 1998, pp. 89–98.

[81] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens: An open architecture for collaborative filtering of netnews," in *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work (CSCW)*, 1997, pp. 175–186.

[82] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems," in *The Adaptive Web*, 2007, pp. 291–324.

[83] L. Getoor and B. Taskar, *Introduction to Statistical Relational Learning*. MIT Press, 2007. [Online]. Available: https://mitpress.mit.edu/9780262538688

[84] C. Aggarwal and H. Wang, *Managing and Mining Graph Data*. Springer, January 2010, vol. 40.

[85] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.

[86] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," in *Proceedings of AAAI*, 1996, pp. 37–54.

[87] E. Garfield, "Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies," *Science*, vol. 178, no. 4060, pp. 471–479, 1972.

[88] L. Bornmann and H.-D. Daniel, "What do citation counts measure? a review of studies on citing behavior," *Journal of Documentation*, vol. 64, no. 1, pp. 45–80, 2008.

[89] K. W. Boyack, H. Small, and R. Klavans, "Improving the accuracy of co-citation clustering using full text," in *Journal of the American Society for Information Science and Technology*, vol. 64, no. 9, 2013, pp. 1759–1767.

[90] M. M. Kessler, "Bibliographic coupling between scientific papers," *American Documentation*, vol. 14, pp. 10–25, 1963.

[91] 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, pp. 265–269, 1973.

[92] J. H. Fowler and D. W. Aksnes, "Does self-citation pay?" *Scientometrics*, vol. 72, no. 3, pp. 427–437, 2007.

[93] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems," in *The Adaptive Web: Methods and Strategies of Web Personalization*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2007, vol. 4321, pp. 291–324.

[94] J. Beel *et al.*, "Introducing mr. dlib: Recommendations-as-a-service (raas) for academia," in *Proceedings of the ACM Conference on Recommender Systems*, 2013, pp. 331–332.

[95] K. Sugiyama *et al.*, "Exploiting linkage between research articles for efficient recommendation in digital libraries," *Journal of Information Science*, vol. 39, no. 2, pp. 232–242, 2013.

[96] D. Goldberg *et al.*, "Using collaborative filtering to weave an information tapestry," in *Proceedings of the ACM Conference on Communications of the ACM*, 1992, pp. 61–70.

[97] T. Theeramunkong *et al.*, "Discovery of research paper groups by clustering based on citation," in *Proceedings of the International Conference on Computational Science and Its Applications*, 2007, pp. 523–532.

[98] X. Chen *et al.*, "Dynamic co-citation analysis for evolving research paper recommendations," in *Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 2021, pp. 521–530.

[99] A. Constantin *et al.*, "Evaluating the efficiency of citation networks for recommendation systems," *Journal of Informetrics*, vol. 7, no. 4, pp. 458–472, 2013.

[100] A. Constantin, S. Pettifer, and A. Voronkov, "Pdfx: Fully-automated pdf-to-xml conversion of scientific literature," in *Proceedings of the ACM Symposium on Document Engineering*, 2013, pp. 177–180.