

**CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD**



**Research Paper Recommendation
Using Citation Proximity
Analysis in Bibliographic
Coupling**

by

Raja Habib Ullah

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

in the
Faculty of Computing
Department of Computer Science

2018

Research Paper Recommendation Using Citation

Proximity Analysis in Bibliographic Coupling

By

Raja Habib Ullah

(PC113003)

Foreign Evaluator 1

Dr. Joel Rodrigues, Professor

University of Beira Interior, Covilha, Portugal

Foreign Evaluator 2

Dr. Denis Helic

Graz University of Technology, Austria

Supervisor Name

(Dr. Muhammad Tanvir Afzal)

Dr. Nayyer Masood

(Head, Department of Computer Science)

Dr. Muhammad Abdul Qadir

(Dean, Faculty of Computing)

**DEPARTMENT OF COMPUTER SCIENCE
CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
ISLAMABAD
2018**

Copyright © 2018 by Raja Habib Ullah

All rights reserved. No part of this thesis may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.

To my parents and family



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 thesis, entitled "**Research Paper Recommendation Using Citation Proximity Analysis in Bibliographic Coupling**" was conducted under the supervision of **Dr. Muhammad Tanvir Afzal**. No part of this thesis has been submitted anywhere else for any other degree. This thesis 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 in Philosophy in the field of **Computer Science**. The open defence of the thesis was conducted on **20 August, 2018**.

Student Name : Mr. Raja Habib Ullah
(PC113003)

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

Examination Committee :

(a) External Examiner 1: Dr. Ahmar Rashid
Associate Professor
GIKI, Topi

(b) External Examiner 2: Dr. Ehsan Ullah Munir
Associate Professor
COMSATS, Wah Campus

(c) Internal Examiner : Dr. Muhammad Arshad Islam
Assistant Professor
CUST, Islamabad

Supervisor Name : Dr. Muhammad Tanvir Afzal
Associate Professor
CUST, Islamabad

Name of HoD : Dr. Nayyer Masood
Professor
CUST, Islamabad

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

AUTHOR'S DECLARATION

I, **Mr. Raja Habib Ullah** (Registration No. PC113003), hereby state that my PhD thesis titled, '**Research Paper Recommendation Using Citation Proximity Analysis in Bibliographic Coupling**' 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.



(**Mr. Raja Habib Ullah**)

2

Dated: 20-

August, 2018

Registration No : PC113003

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled "**Research Paper Recommendation Using Citation Proximity Analysis in Bibliographic Coupling**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

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

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of 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 thesis.



(Mr. Raja Habib Ullah)

Registration No. PC113003

Dated: 20

August, 2018

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 thesis:-

1. **R. Habib**, M. T. Afzal, “Paper recommendation using citation proximity in bibliographic coupling,” *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 25, no. 4, pp. 2708–2718, 2017.
2. **R. Habib**, M. T. Afzal, “Research paper recommender using the centiles in bibliographic coupling,” submitted in *Journal of Information Science*, 2018.
3. **R. Habib**, M. T. Afzal, “Using sections in bibliographically coupled papers for paper recommendation,” submitted in *Scientometrics*, 2018.
4. **R. Habib**, M. T. Afzal, “Research Paper Recommendation Using Citations Proximity Analysis in Bibliographic Coupling,” was accepted in the PhD Symposium in *IEEE international conference Frontiers of Information Technology*, 2017.

Raja Habib Ullah

(PC113003)

Acknowledgements

First and foremost, I must thank Almighty Allah, The most merciful, Who blessed me with the opportunity, and resources to pursue my PhD, and it's all owing to His grace that I saw it through.

I dedicate my dissertation work to my PhD supervisor Dr. Muhammad Tanvir Afzal for the patient guidance, encouragement and advice he has provided throughout my time as his student. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly. Whenever I felt lost and demotivated, he pulled me out of the disbelief and showed me the way. There just aren't enough words to show my gratitude to him. I am grateful to Dr. Muhammad Abdul Qadir, Head of the Center for Distributed and Semantic Computing (CDSC), who motivated the entire group to work hard. I am also thankful to other members of CDSC whose discussion and constructive criticism maintained an environment that was conducive for research.

I take this opportunity to sincerely acknowledge the Higher Education Commission (HEC), Pakistan, for providing financial assistance in the form of Indigenous Scholarship which helped me to complete my work comfortably.

In the end, I would like to thank my parents, brothers and sister, albeit I understand any amount of gratitude shown to them is woefully inadequate. My father's unconditional support is largely the reason that this PhD is completed. No words are sufficient to describe my mother's contribution to my life. I owe every bit of my existence to her. This thesis is dedicated to her. And last but not the least, I would like to thank my wife for her support, encouragement, and quiet patience. May Allah reward you for your sacrifice and selflessness.

Abstract

The immense proliferation of research papers in journals and conferences poses challenges for researchers wanting to access relevant scholarly papers. Recommender systems offer a solution to this research problem by filtering all of the available information and delivering what is most relevant to the user.

Several approaches have been proposed for research paper recommendation, variously based on metadata, content, citation analysis, collaborative filtering, etc. Approaches predicated on citation analysis, including co-citation analysis and bibliographic coupling, have proven to be significant. Co-citation has been analyzed at content level and the use of citation proximity analysis has shown significant improvement in accuracy. However, co-citation presents the relationship between two papers based on their having been mutually cited by other papers, without considering the contents of the citing papers. Bibliographic coupling, on the other hand, considers two papers as relevant if they share common references, but traditionally does not consider the citing patterns of common references in different logical parts of the citing papers.

The improvement found in cases of co-citation when combined with content analysis, motivated us to analyze the impact of using proximity analysis of in-text citations in cases of bibliographic coupling. Therefore, in this research, three different approaches were proposed that extended bibliographic coupling by exploiting the proximity of in-text citations of bibliographically coupled articles. These approaches are: (1) DBSCAN-based bibliographic coupling, (2) centiles-based bibliographic coupling and (3) section-based bibliographic coupling. Comprehensive experiments utilizing both user study and automated evaluations were conducted to evaluate the proposed approaches. The results showed significant improvement over traditional bibliographic coupling and content-based research paper recommendation.

Contents

Author's Declaration	v
Plagiarism Undertaking	vi
List of Publications	vii
Acknowledgements	viii
Abstract	ix
List of Figures	xii
List of Tables	xiii
Abbreviations	xiv
Symbols	xv
1 Introduction	1
1.1 Background	2
1.2 Research motivation	4
1.3 Research Problem	5
1.4 Research Hypothesis	6
1.5 Research Methodology	6
1.6 Research Contributions	9
2 Literature Review	11
2.1 Meta Data Based Approaches	13
2.2 Citations Based Approaches	14
2.3 Content Based Approaches	20
2.4 Collaborative Filtering Based Approaches	23
2.5 User Profile Based Approaches	24
2.6 Data Mining Based Approaches	28
2.7 Hybrid Approaches	30
2.8 Critical Analysis of State-of-the-art	31

2.9 Comparison of Approaches	34
3 Methodology	38
3.1 Dataset Selection	42
3.2 Content Extraction	43
3.3 Proposed Approaches	45
3.3.1 DBSCAN Based Approach	45
3.3.2 Centiles Based Approach	46
3.3.3 Sections Based Approach	47
3.4 Evaluation	48
4 DBSCAN Based Citation Proximity in Bibliographic Coupling	51
4.1 Background	51
4.1.1 Co-Citation Analysis	52
4.1.2 Bibliographic Coupling	52
4.2 What is DBSCAN?	53
4.3 DBSCAN for Bibliographic Coupling	55
4.3.1 Modules of Proposed Approach	55
4.3.1.1 Data Acquisition Module	55
4.3.1.2 Data Normalization Module	55
4.3.1.3 DBSCAN Clustering and Similarity Score Measuring	56
4.4 Results and analysis	61
4.5 Summary	69
5 Centiles Based Approach	72
5.1 Background	73
5.2 Methodology	74
5.3 Results and Evaluation	80
5.4 Summary	84
6 Section Based Bibliographic Coupling	87
6.1 Background	87
6.2 Methodology	89
6.2.1 Data Acquisition	91
6.2.2 XML Conversion	91
6.2.3 Section Extraction	91
6.2.4 Similarity Score Measuring	94
6.3 Evaluation	96
6.4 Summary	101
7 Conclusion and Future Work	105
7.1 Conclusion	105
7.2 Future Work	109
Bibliography	111

List of Figures

1.1	Research methodology	7
3.1	System Architecture	39
4.1	Clustering using DBSCAN	54
4.2	DBSCAN results produced by WEKA	59
4.3	Proposed Approach vs Bibliographic Coupling	62
4.4	Proposed Approach vs Content Based Approach	62
4.5	Performance of proposed approach	63
4.6	Comparative Evaluation of DBSCAN for query 'Automatic generation'	65
4.7	Total no. of queries for which each approach performed better than others in Top 5, Top 10 and Top 15 rankings.	70
4.8	Average Correlation of All Approaches	71
5.1	System Architecture for centiles based approach	76
5.2	Number of in-text citation pairs with different value of centile differences	78
5.3	Comparative Evaluation of centiles based approach for query 'Automatic Generation'	80
5.4	Total no. of queries for which each approach performed better than others in Top 5, Top 10 and Top 15 rankings.	85
5.5	Average correlations of all queries	86
6.1	System Architecture for Section Based Bibliographic Coupling	90
6.2	Weights for citations from the sections and cross sections.	95
6.3	Comparative Evaluation of Sections based approach for query 'Automatic Generation'	97
6.4	Total no. of queries for which each approach performed better than others in Top 5, Top 10 and Top 15 rankings.	102
6.5	Total no. of queries for which the three approaches shared the top position with one or both approaches.	103
6.6	Average correlations of all queries	103
6.7	Comparison of Proposed Approaches	104

List of Tables

2.1	Comparison of Approaches	34
3.1	Queries used for dataset-2	44
4.1	Relationship between the number of clusters produced by subsets and the values of ε	58
4.2	DBSCAN clustering for citation proximity	60
4.3	Top 5 Rankings Comparison	66
4.4	Top 10 Rankings Comparison	67
4.5	Top 15 Rankings Comparison	68
5.3	Top 5 Rankings Comparison	81
5.4	Top 10 Rankings Comparison	82
5.5	Top 15 Rankings Comparison	83
6.1	Weights of different sections.	95
6.2	Top 5 Rankings Comparison	98
6.3	Top 10 Rankings Comparison	99
6.4	Top 15 Rankings Comparison	100

Abbreviations

DBSCAN	Density-based spatial clustering of applications with noise
CPA	Citation Proximity Analysis
J. UCS	Journal of Universal Computer Science
NSI	Normalized Similarity Index
WDC	Weighted direct citation
CL	Combined Linkage
TF/IDF	Term Frequency/Inverse Document Frequency
CBF	Content Based Filtering
KPEM	Key phrase Extraction Module
FP	Frequent Pattern
CTR	Click through Rate
MAP	Mean Average Precision
JSD	Jensen Shannon Divergence

Symbols

$MinX$	Minimum length of paper
$MaxX$	Maximum length of paper
$NewMinX$	New minimum length of paper
$NewMaxX$	New maximum length of paper
v	proximity of citation
New_v	normalized proximity of citation
ε	Epsilon
$minPts$	Minimum Points

Chapter 1

Introduction

In recent times, recommender systems for scientific papers have gained stature and importance, due to a colossal increase in the number of published research papers. Over the past few decades there has been a surge in the number of research papers published in conferences and journals [1]. According to one study, there are almost 25 million freely available scholarly documents on the Web [2]. Researchers today can access a huge quantity of knowledge. While no doubt worthwhile, this creates the problem of 'information overload'. As a result, academic researchers need to go through many different research papers in order to gain background knowledge in a particular area. Thus, the task of retrieving related papers becomes a tedious one.

This phenomenon of information overload has compelled many researchers to address the issue of research paper recommendation. The scientific community has proposed several approaches categorized as: collaborative filtering [3], [4], content-based filtering [5], citation analysis approaches [6], User Profile based approaches [7] or hybrid techniques [1]. A myriad of papers continues to be published about research paper recommender systems [8]. The scientific community, bearing in mind the value and utility of research paper recommendation, continues to propose and implement paper recommendation techniques. These include meta-data, content-based filtering, collaborative filtering, co-citations and bibliographic coupling etc. Content-based and citation-based approaches are the most popular [8].

According to a comprehensive survey of this whole field, 55 percent of paper recommendation approaches use content-based filtering [8]. Approaches based on citation analysis also tend to be very important. These include co-citation analysis, bibliographic coupling and direct citations. A significant amount of research has been done into co-citation analysis. Recent research by Gipp et al. showed that using the content of research papers (Citation Proximity Analysis) for co-citations can improve the accuracy of paper recommendations [9]. Citation Proximity Analysis (CPA) considers two papers relevant if they are cited near to each other in the text of citing papers. Bibliographic coupling, by contrast, is the primitive approach to citations that considers two papers to be relevant to each other if they share common references. Therefore, bibliographic coupling gives us the benefits of recommending relevant papers without, however, considering the citation patterns of common references in different logical parts of the citing papers. In this research, we focus on determining the effect of proximity analysis on paper recommendations produced by bibliographic coupling.

This chapter includes the background of citation analysis in the field of research paper recommender systems. We also discuss some state of the art approaches and, based on a critical analysis of the literature, we analyze the current state of play in the field. Once we have clearly laid out the dimensions of the problem we describe our motivation for doing the research and how that feeds into our research hypothesis. In the end, our research methodology is laid out in a step-by-step description of the process of research.

1.1 Background

In this section, we provide an overview of the background concepts including citation analysis, co-citation, bibliographic coupling and the use of proximity analysis.

Citations have long been referred to as a productive and potentially fruitful source in many different areas of scientific research. The applications of citation analysis range from research evaluation to paper recommendation. Kuhn et al. conducted

comprehensive experiments and analyzed the inheritance patterns in citation networks to determine the memes in scientific literature [10]. They discovered a relation between the occurrence of scientific memes and the degree to which they propagate along the citation graph. Similarly, Perc and Matja analyzed how the Matthew effect, by which fame itself attracts more recognition and fame, applies to citation data [11]. They performed extensive experimentation to conclude that publications with a larger number of initial citations will receive many more in the future, as compared to publications with a smaller number of original citations. This preferential attachment can be used to recommend papers. Papers that acquire more citations tend to be favored more.

Approaches based on citation analysis tend to be the quintessential ones. These include co-citation analysis, bibliographic coupling and direct citations.

Co-citation analysis considers two papers similar if both of them have been cited by one or more common papers [12]. Numerous techniques have been proposed that use co-citation analysis [9], [12]. However, in co-citation analysis, papers are recommended based only on the fact that one or more common citing papers have cited the recommended papers. For example, if paper A cites two papers B and C, papers B and C are considered to be relevant or similar. This similarity or relevance is calculated using the citing papers only. The contents or any other features of the cited papers that were identified to be relevant are completely overlooked while determining their similarity.

Gipp et al. proposed an approach called Citation Proximity Analysis (CPA) to find related papers [9]. Along with citation analysis, the authors used the distance between the citations to discover their relatedness. They discovered that the closer the citations are to each other, the more related the two papers are. The citation proximity analysis increased the accuracy of co-citation by 55 percent.

Bibliographic Coupling uses citation analysis to determine the relationship between documents [13]. Bibliographic Coupling occurs between two research papers if they both cite one or more common research papers. Coupling strength represents the number of common citations from both papers. For example if

papers A and B both cite papers C, D and E, then papers A and B have a bibliographic coupling strength of 3. The larger the number of common papers, the higher is the value of bibliographic coupling strength between them. Similarly, the higher the bibliographic coupling strength, the more similarity exists between the papers (A and B in the above example). Unlike the co-citation approach, in bibliographic coupling, the references of the cited papers are taken into account while determining the similarity.

In the traditional bibliographic coupling approach, only the strength of the coupling is considered to determine the similarity between the papers and the logical structure of the paper: the occurrence of citations in the full text of the papers is ignored. Another problem with traditional bibliographic coupling is that there are significant cases in which the references are included in the references section of the paper but are never referred to within the full text. Shahid et al. identified that there were more than 10 percent of such references in more than 16,000 references of the J. UCS which were part of the reference section but were never used in the text of citing documents [14]. Such citations are called false citations. Therefore, an exclusive reliance on the strength of bibliographic coupling, when it refers only to the references section or bibliography, may lead to incorrect results.

Although the proximity analysis of in-text citations in co-citation has improved the accuracy of paper recommendations, no research has been carried out to identify the impact of proximity analysis in bibliographic coupling.

1.2 Research motivation

We did a comprehensive review of the literature, which covered the existing research in the field and gave us our underlying research motivation.

1. In the last 15 years, more than 55 percent of paper recommendation approaches worked on the content of the papers [8]. In this context, one of the old citation-based approach known as: co-citation [12] was extended to

include content analysis [1], [9]. In approaches based on co-citation, the measure of relevance only comes from the fact that one or more common citing papers have cited some common research papers. For example, if a paper A cites two papers B and C, the papers B and C are considered to be relevant or similar. This similarity or relevance has been calculated using the citing papers only. The cited papers that were identified to be relevant are completely overlooked while determining their similarity, whereas bibliographic coupling identifies related papers based on their common references [13]. However, bibliographic coupling does so without using the in-text citation occurrences, proximities, and patterns which remained very helpful in the context of co-citation based extended approaches. This led us to explore the in-text citation occurrences, proximities and patterns in the bibliographically coupled papers.

2. In the traditional bibliographic coupling approach, content is not used. Content is not analyzed whether the references available in the bibliography section of the papers are actually cited in the content of the paper or not. According to a recent study, it has been found that more than 10 percent of references were never cited in the full text of the papers and were just part of the reference section of the papers. Therefore, unless we make sure that references are actually cited in the full-text by analyzing that text, we should not include such references for bibliographic coupling. This limitation also motivated us to analyze the full-text of the papers.

1.3 Research Problem

Discovering related research papers is of utmost importance for the scientific community. A plethora of approaches have been proposed to recommend research papers. Among these approaches, two of the most important research paper recommendation approaches that use citation analysis are co-citation analysis and bibliographic coupling. Researchers have made significant improvement in the

accuracy of recommendations produced by co-citation analysis by taking advantage of the proximity analysis of in-text citations. However, the in-text citations and their proximity have not been explored in the case of bibliographic coupling. Therefore, this thesis investigates the impact of using the proximity of in-text citations in bibliographically coupled papers in recommending the relevant research papers.

1.4 Research Hypothesis

This thesis arises from the above observations that we think are of great interest to the scientific community. In order to augment the utility and relevance of recommender systems, our research sets out to respond to the following hypothesis:

The accuracy of research paper recommender systems based on Bibliographic coupling can be improved by exploiting the in-text citation occurrences and their proximities between the bibliographically coupled papers.

1.5 Research Methodology

From the viewpoint of application this research can be considered to be applied research, since its principle aim is to resolve a functional problem: the pressing need for a functional recommender system that will return relevant and useful results to the scientific community that is drowning in a sea of papers. This means that the results of our research should be applicable in practice. From the viewpoint of objectives our research can be considered exploratory research, since it is conducted to explore the area of bibliographic coupling and citation proximity analysis, which has not been explored in much detail in the past.

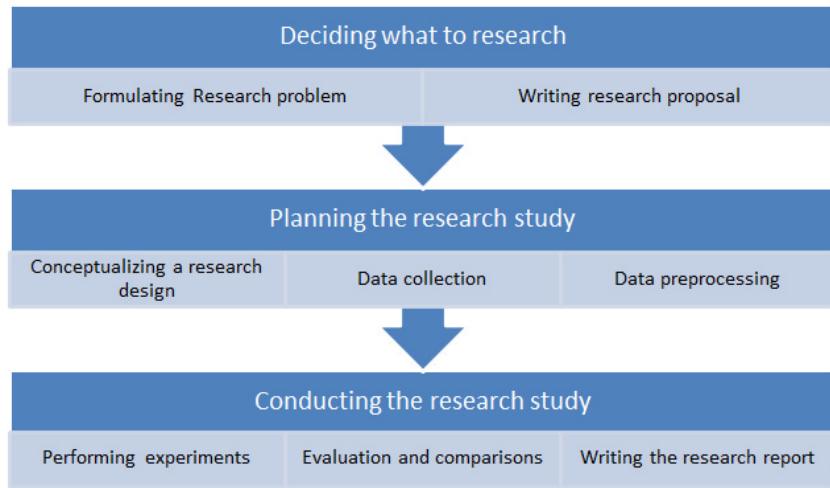


FIGURE 1.1: Research methodology.

In this research, we have used the three-phase, eight-step research model proposed by [15] as shown in Fig. 1.1. In the following paragraphs, we discuss these phases and steps in detail.

Phase 1 Deciding what to research

Step 1: Formulating a research problem:

The following tasks were performed in this step:

1. Literature review
2. Identification of the research gap

Step 2: Writing the research proposal: After identifying the research gap, , we devised the proposal for new bibliographic coupling approaches that could use the proximity analysis in-text citation occurrences. These include:

1. A DBSCAN-based proximity approach.
2. A Centile-based Proximity approach.
3. A Section-wise proximity approach.

Phase 2 Planning the research study

Step 3: Conceptualizing a research design:

We formulated the design for this research in the form of the research hypothesis.. We proposed three different solutions that we will discuss in their respective chapters.

Step 4: Data collection for proposed approaches:

During this step, we constructed instruments for collecting data sets for our experiments and proposed approaches. We designed a dedicated crawler to fetch two different datasets containing the bibliographically coupled papers and their metadata such as title, authors and citations-id. We used this crawler to gather the datasets from Citeseerx.

Step 5: Data Preprocessing:

This step encompasses the pre-processing of the data. After the data collection phase, we cleansed the data by getting rid of those papers in which the correct identification of in-text citations was not possible. Headings in the research papers were mapped onto the logical sections of research papers.

Phase 3 - Conducting the research study

Step 6: Performing experiments:

In this step, different experiments were performed in order to evaluate the performance of our proposed approaches in comparison with other currently existing approaches.

Step 7: Evaluation and Comparisons:

This step involved the evaluation of the performance of our proposed approaches using two different datasets. This step also included the automated evaluation of the performance of different approaches.

Step 8: Writing the research report:

In this step, we produced the dissertation that included the details of all the above mentioned steps in writing. In the dissertation we explain, analyze and critically discuss our proposed approaches.

1.6 Research Contributions

Followings were the main objectives of this research:

1. Explore the proximities of in-text citations of bibliographically coupled papers to recommend research papers.
2. Cluster the in-text citations based on their proximities and patterns in the full text of the papers.
3. Use the proximities of in-text citations without clustering to recommend scientific paper.
4. Use the logical sections of the bibliographically coupled papers to recommend research papers.

Our main contributions in this thesis are as follows:

1. We proposed a DBSCAN based approach that clusters the in-text citations using their proximities.
2. We proposed a centiles based approach that determines the centile values for all the in-text citations and then clusters the in-text citation pairs based on the distance between the centile values.
3. We proposed a sections based approach that exploits the distribution of in-text citations in different logical sections of the paper.
4. Through extensive experiments, we found the optimal value of Epsilon for DBSCAN algorithm for paper recommendation.

5. We proposed two new schemes for assigning weights to different citations pairs in bibliographically coupled papers.

The rest of the thesis is organized as follows. Chapter 2 discusses a comprehensive state of the art literature review. In this chapter we discuss various paper recommendation approaches along with their advantages and limitations. We conclude chapter 2 with a summary and critical analysis of this literature review. Chapter 3 discusses the methodology of this research that includes the discussion of dataset selection, content extraction and the evaluation techniques for our proposed approaches. Chapter 4 discusses our first proposed approach i.e. the DBSCAN based approach. In this chapter, we discuss the traditional DBSCAN approach. We also discuss the experiments we performed in order to find an optimal value for the parameters used in the DBSCAN based approach for research papers. Chapter 5 discusses the second proposed approach i.e. the centiles based approach. Chapter 6 discusses the third proposed approach i.e. the sections based approach. The chapters 4, 5 and 6 also present comprehensive analysis and comparisons of the proposed approaches. In the end, Chapter 7 discusses the conclusion and future work of our research.

Chapter 2

Literature Review

There has been a tremendous growth in the number of research papers that are being published online. Millions of new research papers are added to the scientific knowledge every year [16]. This makes it very difficult to find the relevant information from this humungous collection [8], [17], [1], [18].

Finding relevant information from these huge scientific repositories is a challenging task. Many different techniques and algorithms have been proposed by the scientific community in the past to address this issue. Some of such techniques have also been implemented in the digital libraries (e.g. CiteSeerx, ACM DL, CiteULike and PubMed Central etc.). These solutions, proposed by the scientific community find the relevant articles using certain similarity measures. These techniques can be placed into different categories based on the similarity measures. In this research thesis, more than 150 state-of-the-art papers were reviewed and the available approaches have been classified in the following categories which also serve as sub-headings in this chapter: (1) metadata based approaches [19], [20], (2) citation based approaches [6], [21], [12], [13], (3) content based approaches [5], [22], (4) collaborative filtering based approaches [3], [4], [23], (5) User Profile based approaches [7], [24], [25], [26], [27], (6) Data Mining based Approaches, and (7) Hybrid approaches [28], [29].

In the metadata-based approaches, the similarity between scientific articles is discovered by matching the metadata of papers. The metadata normally used for relevant paper recommendation includes the title of the paper, author's name, venue, date of publishing and keywords etc., whereas the techniques based on citations use the reference list available at the end of each research paper to find the similarity. The citation network is exploited in different ways like Bibliographic Coupling [13], co-citation analysis [12]. The content-based techniques use some content similarity techniques to measure the relatedness of two papers. Another approach to find similarity between research papers is to use collaborative filtering. This approach is being used in many items recommender systems [30], [31], [32]. In this technique, a user-item matrix is generated. Suppose we have the information that two users 'A' and 'B' like three items 'X', 'Y' and 'Z'. Now if 'A' likes another item 'P', it is assumed that 'B' would also like 'P'. So the item 'P' is recommended to user 'B' too. In case of scientific paper recommendation, the citation network is converted into paper-citation matrix which is analogous to the user-item matrix. However, collaborative filtering suffers from certain limitations such as data sparsity, cold start problem, and scalability etc.

Many different hybrid approaches have also been proposed. Some of these use both the content and citations to find out the similarity between two papers. There are some content-citation network based techniques too, which are used to determine the similarity between papers. In these techniques, context enrichment is incorporated to the citation network. Another approach uses the user profiles and access-log history to recommend scientific papers based on the interests and usage behaviors.

Some of the approaches that are based on the above mentioned techniques are explained in the following sections.

2.1 Meta Data Based Approaches

Metadata is defined as "data about data". In the context of research papers, the metadata consists of elements that help us understand different aspects of research papers. Different digital libraries store different pieces of information as metadata. For example ACM Digital Library maintains following information as metadata for their articles: author, title, journal, issue-date, volume, number, month, year and keywords.

Although different digital libraries keep different information as metadata, most of these elements are the same.

Afzal et al. proposed a metadata based approach for relevant paper recommendations [20]. The metadata they used to find the related literature included venue, title, authors and year of publication of research papers. Two papers are said to be relevant if they are written by the same team of the authors, on the same topic, one cites the other one etc. For identifying the citations, a citation mining technique known as 'TIERL' has been proposed. In this approach, Firstly, it is checked if both articles (for which the relationship need to be identified) have the same venue and are written by the same team of authors. Then titles of both the documents are matched directly. If direct match fails, a partial match is applied. This helps in finding if both papers are similar. Furthermore, the year of publication helps in ranking the related papers in a chronological order.

Bollacker et al proposed another approach using a Web based information agent to find the related research papers [33]. A set of keywords is given as an input. The agent uses this information to find the research papers. The metadata of these papers such as the authors, citations etc. is then placed in a SQL database. The database contains the information about document, document words, citations, citation words, cite clusters and cluster weights. Their proposed approach uses a semantic distance measure (citations of paper by other papers) to measure the relatedness of papers.

However, the metadata based approaches have certain limitations. An author may publish articles in different fields. Authors may not necessarily work in the same field. Similarly metadata elements comprise of small set of terms which makes it difficult to achieve good accuracy.

2.2 Citations Based Approaches

A citation is a reference to another paper. An author cites another related paper within the body of a paper, or in the bibliographic section of the paper. A citation is basically an acknowledgment from the author that the cited paper has some relevance to the citing paper. Many techniques for scientific paper recommendation have been proposed based on the citations of the scientific papers.

Kessler and Maxwell proposed a citation analysis approach using bibliographic coupling [13]. This approach uses citation analysis to determine the relationship between documents, for instance if two research papers both cite one or more common research papers. Bibliographic coupling uses coupling strength, which has, at the basis of its calculations, the number of common citations from both papers. For example, if papers 'A' and 'B' both cite 'C', 'D', and 'E', then papers 'A' and 'B' have a bibliographic coupling strength of three. The larger this number, the higher is the overlap between two papers' bibliography and the bibliographic coupling strength.

Small et al developed a paper recommendation technique called co-citation which solved the static nature problem of bibliographic coupling [12]. In this approach, two papers are considered related, if both are cited by the same paper. For example if a paper 'A' cites two different papers 'B' and 'C', the papers 'B' and 'C' are considered to be co-cited by paper 'A'. If papers 'A' and 'B' have 100 co-citations and papers 'B' and 'C' have 50 co-citations, then the paper 'B' is considered to be more related to the paper 'A' as compared to paper 'C'.

The major disadvantage of this approach is that it uses only the co-occurrences of papers and not the actual content of the two papers citing a certain paper. Another disadvantage of this approach is that new papers normally don't have many other papers citing them. So even if they are related, they will not be shown as related in this approach.

Gipp et al proposed an approach called Citation Proximity Analysis (CPA) to find the related papers [9]. Along with the citation analysis, the authors used the proximity or the distance between the citations within the text of the citing document to discover the relatedness between cited and citing paper. Results showed that the closer the citations are to each other, the more related the two papers are. For example if two citations occur within the same sentence, they are considered more related to each other, than if they occurred in two different paragraphs or in two different sections. One limitation to this approach lies in the fact that each author has his unique way of writing. Some authors explain each citation in a different paragraph. Others mention more than one citation in a single sentence. So the results may vary from author to author.

Bethard et al proposed another approach to retrieve related documents using different factors which are important to researchers [34]. The proposed system learns the weights of these factors using the citation patterns. These factors include similar terms, cited by others, recency, cited using similar terms, similar topics and social habits. All of these factors were combined to create a scoring function which is used to rank the articles. A linear classifier was then used to learn the weights of all these features. The system was evaluated by using 10,921 papers from Anthology Reference Corpus. Results showed that the proposed approach produced better mean average precision compared to that of the systems using only features from related work and also compared to the systems which don't use iterative learning.

Shahid et al. proposed an in-text citations based approach to discover and recommend the related papers [14]. The links between the citing and cited paper were

explored, which appear within the text of the research paper. The citation frequency of a cited paper is an important measure of finding the relatedness among two papers. The higher the frequency of in-text citations in the cited paper, the stronger is the relation between the cited and the citing paper. The authors identified that cited paper is relevant to citing paper if cited paper is cited more than five times in the text of the citing paper.

Nassiri et al proposed an approach based on citation networks [35], called Normalized Similarity Index (NSI) to measure the similarity between two papers. In NSI, three types of citation relationships are considered. These include co-citations, bibliographic coupling and longitudinal coupling. Longitudinal coupling refers to the indirect citations between two papers, i.e. the two papers are connected through some other inter-connected papers. The NSI for five different citation networks was calculated. The results were compared with the peer reviews. There was a high correlation between the two results. NSI was also compared with combined linkage (CL) and weighted direct citation (WDC) of those 5 networks. NSI provided much better results than the combined linkage and the weighted direct citation.

Krapivin et al proposed an approach to find the related papers and to rank the related papers by using the famous PageRank algorithm [36]. PageRank algorithm is used by search engines to rank the web documents. The important aspect of this algorithm is that it uses the inbound links to determine the importance and rank of a paper. The larger the number of incoming links to a document 'A', and the higher the pagerank of the incoming links, the higher is the ranking of 'A'. In pagerank, the outbound links also matter. The larger the number of outbound links, the lower is the ranking of the given document. This may be a problem in case of scientific document, since the survey papers normally have a lot of outbound links i.e. citing other papers. This large number of outbound links should not decrease the rank of the paper. To resolve this problem the authors proposed the new approach called Focused PageRank. Also, the recently published papers normally don't have a lot of inbound links, but that doesn't mean that they are not related to a given paper. The authors evaluated their approach

by exploring 266,788 papers published from 1950 to 2007 in ACM conferences and journals. The authors found out that applying the Focused PageRank on papers gives almost the same results as those of Citation Count and are much better than applying the simple PageRank algorithm.

Gori et al proposed a Google PageRank based algorithm for ranking the research papers [37]. This approach is based on the Random Walk approach. Two properties called attenuation and propagation are used in this approach. Propagation suggests that if a paper is connected to a good paper, then this paper itself is a good one too. Attenuation means that a good paper should also spread its influence to other papers to which it is connected. But this reduces the importance of the paper itself. To control this, a decay factor is introduced. This approach assumes that a user has an incomplete paper with some bibliography already written. This is used as an input. A citation graph is used which is an un-directed citing/cited graph. This graph consists of nodes that represent the research papers. An edge between nodes shows that one of the two papers cites the other one. Therefore the graph formed is un-directed. From this graph a connectivity matrix, a stochastic, and a correlation matrix is generated. These matrices help in finding the relationship between papers. Authors used a specialized web crawler to collect dataset from ACM. Furthermore, this dataset was used for the evaluation. Both online and offline evaluation approaches were used, which showed a 100% ranking of most related papers. The limitation of this approach is that it has not been tested for the negative examples i.e. those papers that have been selected by the user but are not actually relevant to the topic.

El-Arini et al. have proposed that instead of using keywords based query to discover the related scholarly papers, using a list of references of different papers yields better results [38]. The notion of influence between the papers was used to capture the flow of ideas from the cited papers to the citing papers and from the previous papers to the subsequent papers. A user study was performed which indicates that the proposed technique performs better than Google Scholar and other paper recommendation systems. However this approach needs the users to provide a list of trusted papers as the query, in order to find the relevant papers.

Using the citation networks alone to recommend the related scientific papers has certain limitations. As we have already discussed, not all the references are always related to the content of the original papers. This problem can be resolved by using the techniques that combine both the citation networks and the content of papers.

Singla et al proposed a hybrid approach for finding the related papers [39]. This approach used a hybrid of content and citation based approaches. In this approach, the authors firstly created a summary for the papers using the keywords, which are extracted from the title and the references of the papers. 15% sentences of total summary are from the abstract, and the remaining 35%, 20% and 30% are from Introduction, Related work and the Implementations sections respectively. The summary of citing papers is also constructed in the same manner. The summaries are then compared using some linguistic measure. This produces a similarity score. The paper rank is then obtained by dividing this similarity score with the number of papers that cited the paper.

The authors applied this approach on 10 papers and found out that there is significant improvement in the rankings. To get better insights into the results of this approach, a better experimentation on a bigger dataset needs to be performed.

Strohman et al proposed an approach to rank the related document by assuming that a user submits his un-published or incomplete document as a query to the search system [40]. The text based, citation based and feature based similarity were used to find the related papers. 6 different features were used such as: (1) publication year, (2) text similarity, (3) co-citation coupling, (4) same author, (5) Katz and (6) citation count. This approach is divided into two steps. In the first step, the user provides a query, and the system returns 100 most similar papers out of over 1 million papers. This set of 100 papers is called 'R'. All papers cited by any of these 100 papers are also added to 'R' in the 2nd step, 'R' now contains almost 1000-3000 papers. The papers in 'R' are then ranked using the above mentioned 6 features. The documents are then ranked based on a document score which is obtained by combining the features in a weighted linear model.

This approach was applied on Rexa database. Results showed that Katz feature is very important. Without this feature, the ranking performance drops by 50%. This approach outperformed the text based similarity.

Reyhani et al proposed an algorithm called SimCC to find the similarity between two papers [41]. Similarity between two papers is considered as a contribution score of the cited paper into the citing paper. According to this approach, the number of citations (alone) received by a paper doesn't depict the relation between papers truly. Content and citations both need to be used at the same time. SimCC score of a cited paper is calculated for each term. To calculate this, contribution score is added to the relevance score. Relevance score is the TF/IDF value of the term. To calculate contribution score, a complex recursive process is executed offline. The authors compared the results of SimCC with cosine, Dice, BM25 and Kullback Leibler Distance and found that their proposed approach performed better than these four.

Bichteler J. et al proposed a hybrid approach for related paper recommendation [42]. This approach uses a combination of bibliographic coupling and co-citation to find the related papers. A user study on 1712 papers was performed. The results showed that using both the bibliographic coupling of documents and their co-citations i.e. cited and citing papers, give better results as compared to using only the co-citations or bibliographic coupling. The main limitation of this approach is that it doesn't take into consideration, the actual contents of the research papers to measure the relatedness.

Boyack et al also presented another approach that uses the proximity of in-text citations for finding the related papers [1]. This technique also uses the distance between the citations. But instead of using the sentence structure, the character or byte offset and centiles positions were used. 4 schemes (B, O , P1 and P2) were proposed for this purpose. Using the 1st scheme 'B', each co-citation pair is assigned a weight of 1. This scheme doesn't take the distance between the in-text citations into consideration. In the 2nd scheme represented by 'O', if the two in-text citations are within the same byte position, they are assigned a weight of 4. If

references are within 375, 1500 and 6000 bytes, they are given weights of 3, 2 and 1 respectively. If the distance is more than 6000 bytes, a weight of 0 is assigned. In the 3rd scheme P1, the paper's text is divided into 20 equal parts which are considered as 5 centiles. The weights are assigned based on these centiles. In the 4th scheme P2, the byte range of centiles is changed. The similarity between the two papers is then discovered based on these weights.

The authors used a dataset containing 270,521 articles from 1,606 different computer science related journals, which contained 12,569,686 references of 4,484,815 different papers. The results of this evaluations showed that there was a 62% improvement in the results. But these results were for the computer science domain only. More experimentation on other domains is also required to validate the technique.

2.3 Content Based Approaches

Many paper recommendation techniques have been proposed to use the content of the research papers to find the similarity. Before the similarity between papers can be found using the content based approaches, some pre-processing needs to be performed, for example conversion of documents to text, removal of stop words, expansion of abbreviations, generalizing the synonyms and stemming of words etc.

Another frequently used approach for recommending items to the users is Content Based Filtering (CBF) [5], [22],. In this approach the items are recommended to users based on the similarities between the items. Content based filtering has been used successfully for movie recommendation and item recommendations in ecommerce websites such as Amazon. This technique builds the user profiles and item profiles based on the items that a user likes. For example if a user 'A' likes a movie with actors 'B' and 'C', the system will add 'B' and 'C' in the user profile. Based on this user profile, the system will check the movie catalogue and recommend him the movies that match his user profile i.e. the movies containing the actors 'B' and 'C'. This technique has some limitations in case of research

paper recommendations, since the number of research papers is huge as compared to the number of users.

Nascimento et al. proposed an approach for recommending related papers [43]. This approach used content based filtering. Article's titles were used to generate the user profiles. The titles and the abstracts of papers were used to generate the feature vectors of the papers. The user provides one research paper as an input. Using this input, the system generates different keywords and queries using the abstract and title of that paper. These queries are then submitted to different sources containing research papers. A set of candidate research papers is generated. Content based recommendation algorithms are then applied to rank these candidate papers. The advantage of using this approach is that it uses only title and abstract of papers which are publicly available. But the limitation of this approach is that a title and abstract cannot effectively represent a user's interests or candidate papers.

Beel et al proposed a paper recommendation system called Docear [44]. This system uses content based filtering to recommend papers to users. This system uses the users mind maps to generate users profiles. Based on these, the system recommends the papers to the user. The system, over the time, keeps collecting the user information like, which paper user is working on, in which field user has written papers previously, which section he is working on right now and so on. Using this information, Docear provides very personalized research paper recommendations.

Daud et al proposed an approach for discovering related documents [45]. This approach uses the semantic information inside the contents of a paper. This approach uses the latent topics information present inside the paper, to discover the semantically related papers. Author-topic modeling was used for this purpose. The topic includes the short description of the document. Experiments were performed on CiteSeer corpus. The dataset included 3,335 papers. 80% of this dataset was used as training data and remaining 20% as the testing data.

Ferrara et al have proposed a content based approach for recommending the papers [46]. In this approach the key-phrase extraction was used to build the user profiles and the document profiles. For this purpose the Key-phrase Extraction Module (KPEM) was used. The user profile is built using the tags which the user assigned to the previous papers. With the help of the profile of users and documents, the relevant documents are discovered for the users. This approach used unigrams, bigrams and trigrams for measuring similarity. This approach was evaluated using publicly available dataset which contains 597 full papers extracted from the ACL Anthology Reference Corpus. This dataset contains information of 28 researchers. It also included their tags. The experimentation showed that the bigram and trigram perform better as compared to the unigram.

Jing et al have proposed a technique for citation recommendation [32]. This technique helps in recommending papers in case of queries which have noise. The citing and the cited papers may contain different terms, and this may make the recommendation difficult. So, they proposed an innovative approach that uses position-aligned translation model. This model aligns the query to the most relevant parts of the document and makes the estimated translation probabilities more accurate. For experimentation, the authors collected 29,353 computer science research papers. These were published between 1988 and 2010. The results showed that the position-aligned translation model helps in improving the effectiveness of recommendations.

Ratprasartporn et al proposed another approach which uses ontology to determine the related papers [5]. In this approach, the contexts have been incorporated in order to measure the relatedness of papers. The existing ontology terms are used as the context of the papers. A paper is assigned to its relevant contexts. Two papers are considered to be related if they both belong to same contexts. The authors performed evaluation using the biomedical ontology terms as contexts for genomics related papers.

Pruitikanee et al proposed an approach to find related papers [47]. This approach is based on fuzzy clustering and consists of four steps. In the first step the user

provides a search query. All the papers, that contain at-least one keyword from the user query, are returned. In the second step, the papers are grouped based on the topic similarity. These clusters are made by using the fuzzy clustering. The third step helps in simplifying the user interaction by computing the representative papers. These representative papers are then ranked by using different rankings algorithms such as PageRank algorithm.

2.4 Collaborative Filtering Based Approaches

Collaborative filtering (CF) is one of the most frequently used techniques for recommending items to the users. In this technique, the items are recommended to a user based on the interests of other users who have the similar interests [30], [31], [32]. For example, if two users 'A' and 'B' like a movie starring Tom Hanks and the user 'B' also likes a movie starring Kate Winslet, then the movies starring Kate Winslet will also be recommended to the user 'A' as well. The users who have similar interests to each other are called neighbors. This technique has many applications in many different fields like ecommerce [48], movie recommendations and research paper recommendations etc.

McNee et al proposed a collaborative filtering based technique for recommending related papers [23]. This approach used the citation graphs of research papers that can also be treated as a social network of papers. The paper-citations relationship was mapped to users-item framework. This relationship between papers and citations is then transformed into a rating matrix. Similarity between papers is then computed based on this rating matrix, and six different algorithms were used for this purpose. The experiments showed that the k-nearest neighbor's algorithm outperforms others. Both the online and offline experiments were used to test this approach.

Agarwal et al proposed another collaborative filtering based approach for recommending research papers to the authors/users [49]. The authors have explained that the recommendation for research papers is different from that for ecommerce.

The main difference is that in case of research papers, the number of users is much smaller than the number of articles. Another difference explained by authors is the large number of dimensions/features of articles. This approach makes sub-clusters of the authors based on the papers they selected or read. This information was collected through the access logs of digital libraries. Using these clusters, new papers are recommended to a user based on what other users in that cluster have liked or read. This approach was evaluated using synthetic data and the MovieLens benchmark data. The evaluation showed that this approach performs better than the normal collaborative filtering approaches. The main limitation of this approach is the preprocessing.

Bogers et al proposed an approach which uses the bibliographical reference managers to recommend related papers to the researchers [50]. CiteULike was used for this purpose. The dataset consisted of 803,521 unique articles, 25,375 authors and 232,937 unique tags. Three different collaborative filtering algorithms were used on one day data dump from CiteULike. Experiments showed that the user-based algorithm performs better than the two item-based techniques. Like most collaborative filtering approaches, this approach also suffers from the cold start problem.

2.5 User Profile Based Approaches

Pohl et al have proposed employed the access logs of digital libraries to determine the relatedness of two papers and recommending related papers [51]. The digital library called arXiv was used. It contains the data of over 650 million accesses to over 350,000 scientific documents between 1994 and July 2006.

From the access logs, the information of time, date, source IP address and documents downloaded by the users. From this information, this approach calculated the number of times each pair of documents was co-downloaded (two papers

downloaded in the same session by the same user). If two papers have been co-downloaded a lot by different users, this means that the papers are related to each other.

The citations data is often difficult to extract in case of newly published papers and in case of images audio and video. But the information of co-downloads is easily available from access logs. Furthermore, no complex information extraction techniques are needed for co-download data. However, such recommendations can only be made in a specialized environment like arXiv, or the specialized digital libraries where users interact with the digital library. But the recommendations cannot be made for the literature available outside these specialized environments.

Lee et al proposed a user profile based research paper recommendation system [24]. The system proposed in this approach takes the name of author and a few keywords from the user as input. The system then presents the related papers to the user, based on these keywords. A web crawler was also introduced which retrieves indexed papers from different conferences and journals.

It was assumed that a user who is searching for related papers, is going to like the papers similar to the paper she herself has written previously. Using the information of an author's published papers; this approach finds similarity between the user's papers and the indexed papers. Vector cosine similarity is used for this purpose. K-Nearest Neighbors algorithm is then applied to recommend the K most related paper.

Sugiyama et al. proposed a technique for scholarly paper recommendation based on researcher's past publications [25]. A weighted vector based on term frequency is created to model the user's profile. Cosine similarity is then used to determine the relevance between the user profile and the document. Experiments were performed to measure the effectiveness of this approach for two different categories of researchers: senior researchers who have published multiple papers recently and the junior researchers who have published only one paper.

Baez et al have conducted a study and shared their results about how most researchers find the relevant research papers [52]. A significant portion of the researchers are influenced by the social factors which include the co-author relationships or the conferences where the researchers met one another. A personalized approach for the recommendation of research papers was proposed. The data from Microsoft Academic Search was used to evaluate the recommendations made by their system. This data included 7,465,398 research publications by 5,726,226 unique authors. They found out that considering the social factor improves the results of recommendations.

Chandrasekaran et al proposed a user profile based approach for recommending research papers [53]. The user profiles and the document profiles were used. Trees of concepts were generated from these and then the tree-edit distance measure was used to find the documents that match the interests of a user. This system consists of three modules. The classifier module classifies the documents into the categories. The ACM categories were used for this purpose. The profile module builds the user profile based on their names, publications and interest areas. The recommender module represents the user profiles and document as trees and uses tree-edit distance measure to find the related documents. This system was evaluated through a user study which involved 8 authors who have their papers published in CiteSeer collection. This approach was compared with the CitesSeer's built in recommender system, and it was discovered that the proposed approach gives much better results. There was an 8% improvement. The authors need to evaluate their approach for other digital libraries too.

Choochaiwattana proposed using a user profile based approach for recommending scientific documents [54]. This approach uses CiteULike, which is a Web 2.0 application. CiteULike allows the researchers to tag the research papers. This approach uses the tags/annotations to build the user profiles. These user profiles are then used to recommend the documents. Through experimentation, it was discovered that the accuracy of the proposed research paper recommendation is 79% with f-measure value at 82

Dong et al utilized the combination of traditional recommender systems and the social recommendation techniques [55]. The social recommendation includes the social media presence of a user/researcher. This includes his friends and their interests. The authors performed experiments and found out that the fusion of collaborative and content based filtering technique works the best. The results of this technique are then combined with the social behavior of the researcher to recommend relevant papers to him. A dataset consisting of 68,625 papers from CiteSeer was used for the experiments and it was discovered that this technique performs better as compared to using the collaborative filtering alone.

Geisler et al have proposed another technique for recommending scientific papers for changing data [56]. This technique was proposed for a digital library called iLuminia. This digital library contains the data which is uncertain and keeps changing. Some users get registered and others don't, so the data of users is different. For this purpose, certain rules were proposed. If a user makes a suggestion, other similar resources can be suggested to him. Resources can be suggested to users based on the user's profile. If a user has a profile and he has also read some papers, then more resources are recommended based on his profile and his previous downloads. This approach is using only the user's information, which may lead to some incorrect recommendations, since a user may have a variety of interests.

Kodakateri et al proposed an approach for recommending research papers, using the user profiles [57]. This approach consists of three parts. The first part is a classifier that is used to classify papers into different categories using the ACM categories. The second part is a profiler that creates user profiles for the registered users, based on the links they clicked after using a query on CiteSeerX. The third part is the Recommender that matches the user profiles with the different categories and suggests papers to the user. The system was evaluated by using a subset of CiteSeerX documents, which consisted of 1,000,000 documents from Computer Science field. A user study was also performed which included 7 professors and graduate students. It was discovered that using the top 3 concepts from user profiles gives the best results. However a comprehensive user study needs to be performed which should include more professors/graduate students.

2.6 Data Mining Based Approaches

Theeramunkong et al proposed association rule mining based approach to discover related papers [58]. Association mining is a data mining technique which helps in discovering the patterns which have strong association among them. In this approach, first of all the text of paper is filtered to find the scientific terms from the paper. Unigram and bigram techniques are then applied to these terms to add semantics to these terms. A candidate set is generated which consists of these terms. A candidate set of papers is also generated. Association mining techniques like Apriori algorithm and FP Tree are then applied to discover the most frequent terms in the documents. The documents that represent the frequent item set are considered to be the most related. However, using APriori or FP Tree algorithms have their own disadvantages especially if the number of transactions (papers) is huge.

Cazella et al proposed an approach to recommend related papers [59]. This approach uses the data mining technique called association rule mining and an intelligent agent to recommend the related papers. The agent crawls different article directories to build a user profile. Based on this user profile, the association rule mining discovers the rules regarding the preferences of the users. For evaluation, a dataset was used which contained 10,000 researchers and their interests. The experiments proved that using association rule mining and intelligent agents together improves the paper recommendation and solves the 'new user' problem. This algorithm helps discover hidden areas of interests for a user.

Beel et al performed an experiment on a paper recommender system to determine if organic paper recommendation is better than the sponsored paper recommendation [60]. This experiment was performed on a paper recommender system called Docear. This system recommended 22,452 papers to 587 users. The labels of these research papers were modified to Sponsored, Free and Partnered. The parameters that were used to evaluate these were (1) Click through Rate (CTR) and (2) Mean Average Precision (MAP). Experiments showed that the recommendations labeled as organic (free) performed better than the sponsored/paid recommendations.

Beel et al performed another experiment to study the user behavior about paper recommendations [61]. The system used 31,942 research paper recommendations for this purpose. These were shown to 1,155 researchers. Docear literature management software was used for this purpose. This experiment showed that the chances of a researcher/user to click on a recommendation were double for fresh (previously unknown) recommendations, as compared to the recommendations which are shown multiple times. The authors performed this study only on Docear. They need to perform this study on other digital libraries too. Other factors such as the placement of recommendations and design factors also need to be studied.

Beel et al performed another experiment to study the impact of demographics and other user characteristics on the Click Through Rate (CTR) for the research paper recommendations [62]. As many as 37,352 recommendations were presented to 1,028 researchers and it was discovered that elder users (age between 50-54 years) clicked more often on the recommendations and had a CTR of 9.26%. The younger users (age between 20-24 years), on the other hand, had a lower CTR of 2.73%. Gender of the users had a little impact on the recommendations CTR. It was concluded that the recommender systems need to maintain the detailed information of the users. This can help in improving the usefulness and effectiveness of the recommendations.

Bollen et al proposed a methodology to evaluate different recommendation systems [63]. They proposed that such evaluation can be used to collect useful empirical data about the knowledge of researcher communities within certain institutions. A Research Library at the LANL was used. This digital library keeps the web logs of the users. A similarity matrix for the journal titles was constructed. Then hierarchical clustering was applied on this matrix. From these clusters the user communities were derived based on the frequently downloaded articles by the researchers.

2.7 Hybrid Approaches

Torres et al proposed a hybrid approach for scientific paper recommendation [64]. The use of collaborative filtering and content based filtering approaches together was proposed. A set of tools was developed to search the related papers. This approach was implemented in the form of two modules which were called CF module and CBF module. For the experiments, CF and CBF modules were then run together in 5 different combinations. CF and CBF were run independently, together and in the form of a fusion. In fusion, both the CF and CBF run in parallel and their result lists are merged together. If both lists contain a paper, then that paper is added to the list of recommended papers. The rank of this paper is the sum of its rank in both the lists. The papers which are not present in both the lists are appended to the final list. The fusion technique performs much better than using CF and CBF separately.

For experiments, the authors used a dataset from CiteSeer. This dataset contained 500,000 research papers and 2 million citations. After some pruning and pre-processing, 102,295 research papers remained. Online and offline experiments were conducted on these papers. For experiments and evaluation, they applied this technique to the papers from Computer Science field only. More experiments need to be carried out on other domains as well.

Wang et al proposed a hybrid approach for scientific paper recommendation [65]. In this approach, collaborative topic regression model was used. This approach uses both CF and content analysis based on probabilistic topic modeling. This approach uses two types of data to generate the recommendations. These two types of data are: (1) the other users' libraries and (2) the content of the articles. This approach helps to find both the old papers and the recent papers that are related to the user's interests. The older papers are recommended based on number of other users who have seen them already. The new papers are recommended based on their contents. A study was conducted which showed that their approach works better than traditional matrix factorization methods. The abstract and title of the

paper were used for modeling the user and to find candidate papers to recommend. This often results in irrelevant recommendations.

Ekstrand et al have proposed a hybrid approach for recommending papers [66]. This approach uses collaborative filtering, content based filtering and the paper's influence measure to discover the related papers. To measure the influence of papers, the link ranking techniques like SALSA, PageRank and HITS were used. For evaluation, a dump of ACM Digital Library was used. It included 256,937 articles and 201,145 reference lists. This evaluation showed that Salsa performed better than PageRank and HITS to rank the documents. Furthermore, CBF-CF performed the best for filtering the documents.

2.8 Critical Analysis of State-of-the-art

Meta Data Based Approaches:

The main advantage of using the meta-data based approaches for determining relevance among papers is that the meta-data of papers is available freely and openly. The information like venue, author-name, year-of-publication etc. is easily available in different digital libraries.

But the meta-data approaches have certain limitations too. For example, we cannot determine the relationship between two papers using the author names. The same author may have published papers in totally opposite research areas. Similarly venue and year of publication may also not be very helpful for determining the similarity between research papers. Two papers may have been published in the same year and on the same venue but one may be in the field of data mining and the other one in computer networks. Furthermore, the metadata is represented in small number of features which does not give a fair chance to identify related papers.

Citations Based Approaches:

Using citations based approaches to determine the relevance between papers can produce much better results as compared to meta-data based approaches. The research papers normally cite the papers which are related to them, so the relationship found using the citations are usually more meaningful.

However, these approaches also suffer from some limitations. Using the citations alone, while ignoring the actual content of the research papers, may lead to incorrect results. Some researchers cite a paper in the references section without actually using them in the main content of the paper. And such citations may often prove to be less useful. Similarly, the relevant papers which have not been cited by the authors of the papers may not be discovered from such approaches.

Content Based Approaches:

Using content to determine the similarity between research papers provides much better results as compared to using only the citations. We can get much better relevant papers, if we use the content of the papers for this purpose.

But these approaches have certain limitations too. One of the main limitations is that the full content of the papers is usually not openly available. Even if the full content of the papers is available, processing the whole content of the papers can prove to be very costly. Therefore using the content alone for the paper recommendations is not recommended.

Furthermore, these approaches fails when related papers use different vocabulary or different papers use the similar vocabulary, and when one of the paper use abbreviations and other use full terms, and when context remains important for example the term apple could have two contexts such as: fruit or company.

Collaborative Filtering: Collaborative filtering suffers from certain problems. One of these problems is the Cold Start problem. Collaborative filtering is based on the user-item matrix. Items are recommended to users based on their previous preferences. So, if a new user is added to the system, he/she would need to rate a certain number of items before the items could be recommended to other users based on his/her ratings. Similarly if a new item is added to the system, it will

need to be rated by a certain number of users before it can be recommended to the other users.

Another issue faced by the collaborative filtering based approaches is the scalability. The number of users and items is usually massive. This leads to huge computation costs using collaborative filtering.

User Profile Based Approaches:

These approaches are dependent upon the availability of usage profile information of the digital libraries. So, without the access to enough usage information, these approaches do not provide the required results. Some digital libraries may provide access only to a small or absolutely no part of the usage data.

Data Mining Based Approaches:

The Data Mining based approaches are very good at discovering hidden patterns, which is not possible using the other techniques. But these approaches suffer from the scalability problems. Since the number of papers and researchers is huge, the processes of discovering frequent patterns, classification and clustering may prove to be very costly.

As we can see from this literature review, researchers have proposed many different approaches for paper recommendation. These approaches such as content based approaches, collaborative filtering based approaches and meta-data based approaches etc. have certain limitations due to the fact that they do not consider the in-text citations and their proximity while recommending research papers. Researchers have improved the co-citation based approaches by incorporating the content analysis and citation proximity analysis and this resulted in better and more accurate paper recommendations. However, the impact of using content analysis and citation proximity analysis in bibliographic coupling is yet to be explored. In order to bridge these gaps, we have proposed three different approaches that are based on bibliographic coupling. In these approaches, the in-text citations and their proximity in bibliographically coupled papers are considered in order to recommend scientific papers.

2.9 Comparison of Approaches

TABLE 2.1: Comparison of Approaches

Category	Research Work	Methodology	Strengths	Limitations
Meta Data Based Approaches	[20], [33]	Meta data of scientific papers (e.g. author name, year of publications) is used to find the related papers.	Meta data of the scientific articles is generally very conveniently available.	These approaches will not work when: (1) The same author may be publishing papers in two different domains. (2) There are some venues which are generic and publish papers related to many diversified fields. (3) Very few features of metadata do not generally produce better results.

Continued on next page

Table 2.1 – *Continued from previous page*

Category	Research Work	Methodology	Strengths	Limitations
Citations Based Approaches	[12], [13], [9], [34], [14], [35], [36], [37], [39], [1]	The related papers are discovered using the citation analysis. This includes co-citation analysis, bibliographic coupling and citations networks etc.	(1) Citations are normally freely available on different digital libraries. (2) Citations are handpicked by the authors of the papers and make a good candidate for relevant papers already.	These approaches do not work when: (1) Authors cited a paper in the reference section but do not cite it in the full text of the papers. (2) Authors have missed certain relevant papers while making citations to relevant papers. (3) Two papers may cite certain number of common papers for different reasons.

Continued on next page

Table 2.1 – *Continued from previous page*

Category	Research Work	Methodology	Strengths	Limitations
Content Based Approaches	[5], [43], [44], [45], [32], [47]	The related scientific papers are recommended using the content of the research papers.	Using the content based approaches, the recommended papers are normally semantically related too because the results are based on actual content of the papers.	These approaches do not work when: (1) Two relevant papers are using different vocabularies. (2) two irrelevant papers are using similar vocabulary (3) vocabulary context, for example apple fruit and apple computers, USA and United States of America
Collaborative Filtering Based Approaches	[30], [23], [49], [50]	The user-item matrix is created for the users and the research papers. In these approaches the papers are recommended to researchers based on the interests of other similar researchers.	The techniques based on Collaborative filtering generally provides better results and coverage when there is dedicated digital library where users collaborations are being recorded.	Collaborative filtering approaches suffer from many issue: (1) Works for a specialized digital library and may not recommend papers outside this collaborative network (2) Cold start problem (3) Grey Sheep

Continued on next page

Table 2.1 – *Continued from previous page*

Category	Research Work	Methodology	Strengths	Limitations
User Profile Based Approaches	[51], [24], [25], [52], [53], [55],	These approaches use the access logs of digital libraries and recommend the papers based on the actual usage profiles of the researchers.	These approaches recommend the papers based on the actual usage profile of the researcher, so the recommended papers are more personalized.	Some digital libraries may provide access only to a small or absolutely no part of the usage data.
Data Mining Based Approaches	[58] , [59] , [60] , [61] , [63]	These approaches dis- cover the hidden patterns and recommend related pa- pers using the techniques like frequent pattern mining, clas- sification and clustering.	These approaches find out the hidden patterns and hence can make serendipitous discoveries	Since the number of research papers and authors are large in numbers, these approaches suffer from scalability problems. A large number of patterns may immerse due to large number of papers and authors.

Chapter 3

Methodology

This research has proposed three approaches for research paper recommendation such as: (1) DBSCAN based approach, (2) Centiles based approach (3) and sections based approach. These are discussed in chapters 4, 5, and 6 respectively. Before going into the details of each approach, we want to give the reader an overview of what we have done and why we have proceeded in this particular direction and how these approaches came into existence. To avoid repetition, the common steps required to understand the flow in chapters 4, 5, and 6 are explained beforehand in this chapter. It is therefore advisable to read this chapter before going on to read the subsequent chapters.

This research attempts to evaluate bibliographic coupling at content level. In our efforts to provide a more comprehensive evaluation we examined the literature and found that co-citation was extended with respect to content in an approach known as Citation Proximity Analysis, or CPA [9]. Therefore, we have implemented the same approach for bibliographic coupling. Furthermore, we have proposed two new approaches based on the contents of bibliographically coupled papers. The three approaches can be outlined as:

1. DBSCAN based approach

2. CPA based approach

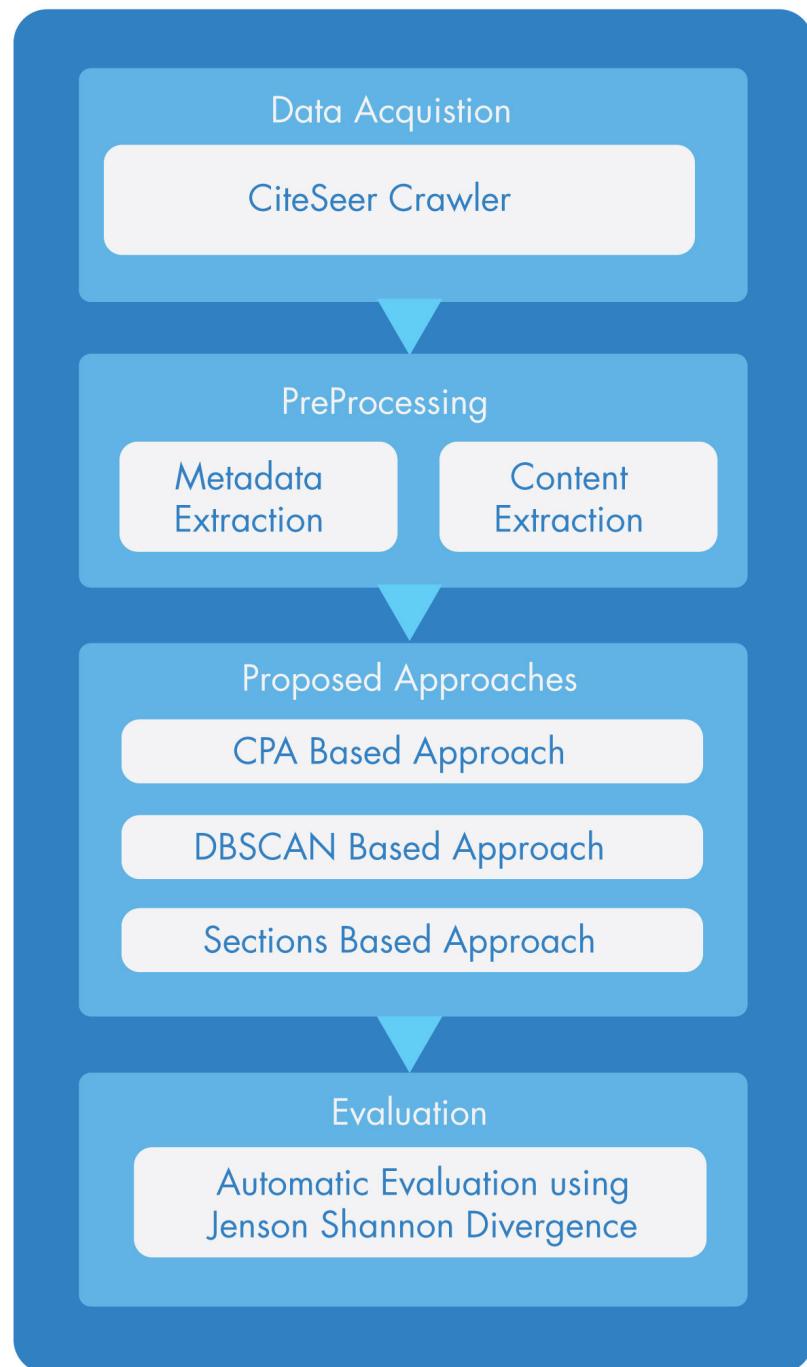


FIGURE 3.1: System Architecture

3. Section based approach

Fig. 3.1 shows the overall flow of the research, highlighting the different steps.

In preparation for our experiments, we amassed two datasets of different sizes using a crawler to gather the data from CiteSeer¹. We collected two datasets of different sizes. We performed initial experiments on the smaller dataset, and then comprehensive experiments on the larger dataset. The details of these datasets and how they were gathered are provided later in this chapter.

The first approach we used was DBSCAN-based. Since the lengths of scientific papers can vary, some being perhaps 15,000- 20,000 words long and others 3,000-5,000 words long, comparing the proximity of citations using the position of citations may lead to incorrect results. To compensate for this, we normalized the values of citation positions within the full text of documents. For our research, we use the Mix-Max Normalization. This algorithm performs a linear transformation on the data values. For the interval $[MinX, MaxX]$ for a feature X, we used Min-Max normalization to transform it into a new interval $[New-MinX, New-MaxX]$. Similarly each value v in the original interval is converted into a new value $New-v$ using the Eq. 3.1.

$$New-v = \frac{v - MinX}{MaxX - MinX} * (New-MaxX - New-MinX) + New-MaxX \quad (3.1)$$

In our proposed algorithm, we used a density based clustering approach called DBSCAN to discover the clusters of citations. DBSCAN discovers the clusters based on the density of the items in the item set. The two parameters used in DBSCAN are ε and $minPts$. This approach is discussed in full detail in Chapter 4.

In our second proposed approach, we used Citation Proximity Analysis (CPA) in bibliographic coupling for recommending papers. This approach extends traditional bibliographic coupling by also integrating the proximities of in-text citations. It mines the patterns of in-text citations in bibliographically coupled papers and

¹<http://citeseerx.ist.psu.edu/index>

recognizes clusters based on their normalized proximity using the centile positions. We discuss this approach in detail in Chapter 5.

Our third proposed approach is based on the intuition that authors cite certain papers in particular sections for certain reasons. Citations from different sections have different weights in determining the similarity between the documents. In this approach we analyzed the in-text citations from different sections of various research papers and tried to discover whether the existence of an in-text citation in a particular section has any impact on the accuracy of paper recommendations. We discuss this approach in Chapter 6.

Our proposed approaches take as input a query from the users. The user can pose any query such as 'data mining', 'collaborative filtering' etc. Our database contains a dataset of 5000 bibliographically coupled papers. Each of our proposed approaches applies their respective algorithm on these queries and the dataset and returns a list of ranked research papers in order of their relatedness.

After performing the experiments, we evaluated our proposed approaches and compared their accuracies with the traditional bibliographic coupling and content based approached. We evaluated our approaches in two steps:

1. User Study
2. Automated Approach

For the user study, we used a dataset that consisted of 320 bibliographically coupled pairs. Every paper was evaluated by two individual users. For each paper, the inter-rater agreement was calculated using Spearmans correlation coefficient. Using these rankings, the proposed approach was compared with the bibliographic coupling approach and the content based approach. This comparison was done using Spearmans correlation coefficient. The comparison showed that our proposed approaches performed better than the traditional bibliographic coupling and content based approaches. We discuss the evaluation and the results in Chapter 7.

We also used an automated instrument to evaluate the performance of our proposed approaches. For this we made us of the Jenson Shannon Divergence (JSD). JSD finds the distance between two probability distributions. In the case of research papers, the word-distribution of individual research papers formed one probability distribution and the word-distribution of the entire cluster formed the second probability distribution. The results of this automated instrument also suggested that our proposed approaches provided greater accuracy than existing approaches. We discuss this process and the results in detail in Chapter 7.

3.1 Dataset Selection

In order to comprehensively evaluate our proposed approaches, a comprehensive dataset was required. There are many different digital libraries and online resources that offer the datasets. For example, PubMed provides access to almost 27 million citations for biomedical literature. Scopus is another huge repository of research papers. However, few of these repositories provide access to the datasets for free. Users have to pay for it. Another issue with some of these repositories is that it is a challenging task to extract the references from the papers. The process of downloading bibliographically coupled papers is complicated.

For this study, we used a digital library called CiteSeer to gather our dataset. CiteSeer is a huge repository that has around 2 million publications indexed. It provides access to the metadata (authors name, venue and year of publication, etc.) and the full texts of research papers. Researchers have used CiteSeer data in the past for various tasks, including text classification, collective classification and citation recommendation etc. [67]. There are two main reasons for using this digital library. The first is that it provides free access to the datasets, which can also be accessed in many different ways. The second is that it retains all the cited papers in a special table, and citing articles can be linked to them using a key attribute CID. In other words, CiteSeer simplifies the process of downloading datasets of bibliographically coupled papers.

We developed a focused crawler to download two different datasets. We used the first dataset for initial experiments, and the second for more extensive and comprehensive experiments. We called them dataset-1 and dataset-2. Initially, we collected dataset-1, containing 320 bibliographically coupled papers. Later, we collected the larger dataset-2, containing 5,000 bibliographically coupled papers from different domains.

We used the 17 queries mentioned in the Table 3.1 to collect the dataset-2. These queries were chosen in order to provide a comprehensive and diversified dataset.

CiteSeer contains millions of freely accessible research papers. However, we downloaded only the research papers that were bibliographically coupled. For this purpose we used the queries mentioned in Table 3.1. For each query, the top 20 results were considered. For each of these 20 papers, we downloaded the citing papers. We included the top 20 citing papers for each case as the bibliographically coupled papers in dataset-2.

dataset-1 consisted of 320 bibliographically coupled papers which were divided into 32 subsets. Each subset consisted of 10 papers that were bibliographically coupled based on a certain query paper. dataset-2 was divided into 226 subsets. These subsets were generated based on the combination of the search query used and the cited-paper-id. These subsets were later combined into 17 groups each representing a query.

3.2 Content Extraction

These datasets contained the research papers in PDF format. While papers in PDF format can provide useful information, we had to convert them into XML in order to fetch other important aspects of content. We converted the papers into XML using an online tool called PDFx². This is a specialized tool for the

²<http://pdfx.cs.man.ac.uk/> (accessed on 14 January 2018)

TABLE 3.1: Queries used for dataset-2

QID	Query
1	Social network
2	Information retrieval
3	Bayesian networks
4	Feature selection
5	Collaborative recommendation
6	Recommendation system
7	Content based filtering
8	Black box testing
9	Automatic generation
10	Regression testing
11	Query processing
12	Sensor networks
13	Wireless communications
14	Opinion mining
15	Subjectivity analysis
16	Online marketing
17	Graph theory

conversion of research papers from PDF to XML format [68] and is useful for converting files in bulk.

The XML files contain certain important elements, the most important of which are the section, *ref* and *xref*. The element *xref* with the attribute *ref-type="bibr"* represents the in-text citations and can be linked to the *ref* tags through the attribute *rid*. This *rid* attribute proves to be very helpful in counting the frequency of in-text citations within sections.

The section element refers to all the sections inside the research paper. This element consists of a nested heading element denoted by *h1*. This heading tag refers to the heading of each section. Two further levels of headings are also provided,

namely $h2$ and $h3$. In our section-based approach, we used the Document Object Model (DOM) to traverse the XML files and fetch the section headings.

Both of these datasets were stored in an SQL database. The information stored in the database includes the metadata of research papers, DOIs of all the cited papers, DOIs of all the citing papers, the positions of in-text citations in all the citing papers, the section headings of all the citing papers and the centiles to which the in-text citations belong.

Although our proposed approaches explore the full text of the papers in order to fetch the in-text citations, these can't be considered to be content based approaches. Our proposed approaches use certain XML tags and elements to discover the proximities of in-text citations. Our approaches don't use any of the techniques (such as cosine similarity or TF-IDF etc) used by the content based approaches to recommend papers. Therefore these can't be considered as content based approaches.

3.3 Proposed Approaches

We proposed three approaches:

1. DBSCAN based approach,
2. CPA based approach and
3. Sections based approach.

In the following sections, we will explain all of these approaches.

3.3.1 DBSCAN Based Approach

The first approach that we proposed for recommending research papers uses a density-based clustering algorithm called DBSCAN (Density-based spatial clustering of applications with noise). Our examination of the literature showed that,

while researchers had found that the use of citations proximity analysis for co-citation can help improve accuracy, the impact of using proximities of in-text citations in bibliographic coupling has not been analyzed extensively in the past, . Therefore, we decided to analyze the impact of using proximity analysis and the positions of in-text citations in full texts in cases of bibliographic coupling.

We first extracted all the in-text citations from the bibliographically coupled papers. Next, we found the proximities of all the in-text citations. To compensate for the varying lengths of papers, we normalized the proximities of the in-text citations using min-max normalization.

In the traditional DBSCAN algorithm, clusters are formed using two parameters: ε and $minPts$. ε represents the radius and $minPts$ represent the number of minimum points required within the ε . The values of these parameters are inputted. In our case, we performed an extensive experiment on dataset-1 to determine the value of ε that produced the most accurate recommendations. We found that the best value of ε was 150. Later on, we used this value of ε on our dataset-2.

We discuss the details of this approach, and the details of its evaluation and comparison with other approaches, in Chapter 4.

3.3.2 Centiles Based Approach

As we mentioned in the previous chapter, more than 55% of approaches to research paper recommendation utilize the content of papers. Co-citation is one of the oldest citation based approaches in this regard. Researchers have performed content analysis on co-citation and have found improvements in accuracy [1], [9]. Gipp et. al. proposed a CPA-based approach for co-citation and their results showed that the relevance between two papers is higher in cases where the citation is from the same sentence. The relevance decreases when the citation is from the same paragraph instead of the same sentence. However, according to Boyack at al, the

reference positions in the full text can be specified without the sentence, paragraph, and section demarcations [1]. In both studies, results showed an improved accuracy in paper recommendation as compared to simple co-citation.

The improvement in accuracy associated with the use of centile locations of in-text citations in co-citation analysis motivated us to explore the in-text citation occurrences, proximities and patterns in bibliographic coupling. The proposed approach clusters the in-text citations based on their centile positions.

Initially we used dataset-1 for this approach. First, we found the positions of the in-text citations. Then, we calculated the centile location of each in-text citation. Next, the distance between the centile values of all the in-text citations pairs were calculated. These values were stored in the database and used by five different citation proximity schemes that cluster the citation percentile values using different thresholds. We used two weighting schemes proposed by Boyack et al [1]), and also used dataset-1 to test and evaluate three new weighting schemes for our proposed approach.

We discuss the details of this approach and these weighting schemes in the Chapter 5.

3.3.3 Sections Based Approach

The way in-text citations are distributed in the full text of a research paper varies from author to author. The way authors place the citations is subjectively decided. However, studies [69], [12] suggest that authors generally follow a certain set of procedural standards when referencing other papers. Another study [70] highlights the fact that authors normally tend to prefer certain sections over others when distributing in-text citations. According to this study, citations are most common in literature review sections, followed by methodology sections.

This raised our interest in exploring section-based bibliographic coupling for paper recommendation. In this approach, we used dataset-2 to fetch the sections from

the citing papers. In the next step, we mapped these sections to a set of generic sections that were determined using previous studies [71], [72]. Then we assigned weights to the in-text citations from all sections. Literature shows that in-text citations from methodology and results sections are given more weight than those from introduction sections, and in-text citations from related work carry the least weight [73], [25].

In the first proposed approach (DBSCAN based approach), we used a density based clustering algorithm called DBSCAN to cluster the in-text citations. As mentioned earlier, DBSCAN algorithm can identify the arbitrarily shaped clusters too. In the second proposed approach (centiles based approach), the five different schemes assign different weights to the in-text citation pairs based on the difference between their centile values. In case of four out of these five schemes, the in-text citations are assigned to their respective clusters based on these weights. In case of one remaining schemes, K-Means clustering was used to cluster the in-text citations. These two approaches take the proximity of in-text citations into consideration for recommending relevant papers. However, our third approach (sections based approach) utilizes the section structure of research papers. This approach clusters the in-text citations and assigns weights to them based on the distribution of in-text citations in different sections of research papers.

The details of this approach are further discussed in the Chapter 6.

3.4 Evaluation

There are three main methods of evaluating the research paper recommendation systems [44]. These are: (1) user studies, (2) online evaluation and (3) offline evaluation. Beel et al reviewed 176 papers and found that 69% of the approaches used offline evaluation, 34% used the user studies and 7% used the online evaluation for paper recommendation approaches. In the case of the offline evaluation, 29% of the approaches used the CiteSeer data.

In order to evaluate our proposed approaches, we needed to use a benchmark dataset. Unfortunately, there is no such benchmark dataset available for evaluating approaches to research paper recommendation. Many researchers have used the user studies to evaluate their research paper recommendation approaches [44], [24]. Therefore, we also used the user study to evaluate one of our proposed approaches (DBSCAN-based approach). For this purpose, we used our smaller dataset consisting of 320 bibliographically coupled papers and carried out a user study. The 320 papers were assigned to 10 experts. These experts included 4 PhD scholars and 6 MS students. Each paper was evaluated and manually ranked by 2 unique users.

Then we calculated the inter-rater agreement between these rankings. The correlation between two rankings can be determined using different correlation coefficients e.g. Kendall rank correlation coefficient, Spearman's rank correlation coefficient and Cohen's kappa coefficient.

The Kendall rank correlation coefficient also known as Kendalls tau coefficient is used to measure ordinal association between two rankings. It is calculated by first determining the concordant and discordant pairs among the set of observations. A higher value for the coefficient means there is a higher correlation among the rankings and vice versa. Kendalls tau coefficient performs better if there is impulsive noise in the data. Similarly if the sample size is large, the Kendalls tau coefficient is preferred.

Another popular statistic used to measure the correlation between rankings is the Cohens kappa coefficient. It is used to measure the inter-rater agreement for categorized or classified instances. Since in the case of research paper recommendation, the papers are being ranked based on their relatedness and are not categorized, the Spearmans correlation coefficient is the more suitable measure for determining the inter-rater agreement.

Based on the above observations, we used the Spearman's correlation coefficient to measure the level of agreement between the two users for each paper.

The user study is an effective way of evaluating research paper recommendation systems, but it suffers from the problem of scalability. A user study can be conducted only with smaller datasets. It can become very costly where larger datasets are used and is not the best option for comprehensive evaluation.

In order to overcome the limitations of user study evaluation, we used offline evaluations and employed an automated instrument for evaluating all of our proposed approaches. For this purpose, we used Jenson Shannon Divergence (JSD), which measures the similarity between two probability distributions. It is based on the Kullback Leibler divergence. JSD finds the distance between two probability distributions. In the case of research papers, the word distribution of individual research papers forms one probability distribution and the word distribution of the entire cluster forms the second probability distribution.

We also performed in depth and extensive evaluation using the top 5 rankings, top 10 rankings and top 15 rankings. This helped us determine which approach performed better in different cases. We have discussed these in detail in the respective chapters of each approach.

Using this automatic evaluation, we were able to extensively evaluate the three approaches that we proposed. We discuss the details of the evaluation of each approach in their respective chapters.

Chapter 4

DBSCAN Based Citation Proximity in Bibliographic Coupling

One version of this chapter has been published in the Turkish Journal of Electrical Engineering [74].

4.1 Background

Over last few decades, research paper recommendation has emerged as a very hot research area with a wide range of applications. One of the approaches for paper recommendation is Bibliographic coupling [13]. The traditional Bibliographic Coupling is afflicted with certain limitations. The primary reason for these limitations is the fact that traditional Bibliographic coupling is entirely based on the number of bibliographic coupling units and doesn't give consideration to the proximity and patterns of the in-text citations. To get rid of these limitations, this research proposes different approaches that extend the traditional bibliographic coupling by making use of the proximity of in-text citations of bibliographically

coupled papers. One of these proposed approaches takes into account the proximity of in-text citations by clustering the in-text citations using a density based algorithm called DBSCAN.

4.1.1 Co-Citation Analysis

Co-citation analysis considers two papers similar if both of them have been cited by one or more common papers [12]. Numerous techniques have been proposed that use the co-citation analysis. One limitation of co-citation is that it measures the relevance between papers based on their co-occurrences in other papers but doesn't take into account the content or other features of the cited papers (the ones which being recommended as relevant papers). Bibliographic coupling, on the other hand, presents a relationship between two papers based on their common references.

4.1.2 Bibliographic Coupling

Bibliographic Coupling uses citation analysis to determine the relationship between documents. Bibliographic coupling occurs between two research papers if they both cite one or more common research papers. Coupling strength represents the number of common citations from both papers. For example if papers 'A' and 'B' both cite papers 'C', 'D' and 'E', then papers 'A' and 'B' have a bibliographic coupling strength of 3. The larger the number of common papers, the higher is the value of bibliographic coupling strength between them. Similarly, the higher the bibliographic coupling strength, the more similarity exists between the papers ('A' and 'B' in the above example). Unlike the co-citation approach, in the bibliographic coupling, the references of the cited papers are taken into account while determining the similarity.

As we can see, in the traditional bibliographic coupling approach, only the bibliographic coupling strength is considered to determine the similarity between the papers and the logical structure of the paper and the occurrence of citations in

the full text of the papers are ignored. Another problem with the traditional bibliographic coupling is that there are significant cases in which the references are included in the References section of the paper but are never used inside the full text of paper. Shahid et al identified that there were more than 10% such reference in more than 16,000 references of the JUCS which were part of the reference section but were never used in the text of citing documents [14]. Such citations are called false citations. Therefore using only the bibliographic coupling strength may lead to incorrect results.

4.2 What is DBSCAN?

Clustering is used in data mining to group similar objects into same clusters and the dissimilar objects into different groups. Various clustering approaches exist that can be categorized as hierarchical clustering, density based clustering and grid based clustering.

K-means clustering algorithm partitions the data into k clusters. Each observations falls into the cluster with the nearest mean. K-Means clustering has certain limitations. In K-means clustering, the prediction of number of clusters produced is difficult. Similarly the clusters produced may vary depending on the initial seeds. Another limitation of K-Means clustering is that the normalization or scaling of the dataset may also change the results.

DBSCAN is a density based clustering algorithm and has certain advantages. Unlike K-means clustering, DBSCAN doesn't need one to specify the number of clusters to be produced in advance. Another advantage of DBSCAN is that it can discover the arbitrarily shaped clusters. DBSCAN determines the clusters using only two parameters and is independent of the ordering of the points in the database.

DBSCAN determines the clusters by finding the core objects. The core objects are the data items that have dense neighborhoods. DBSCAN forms clusters by

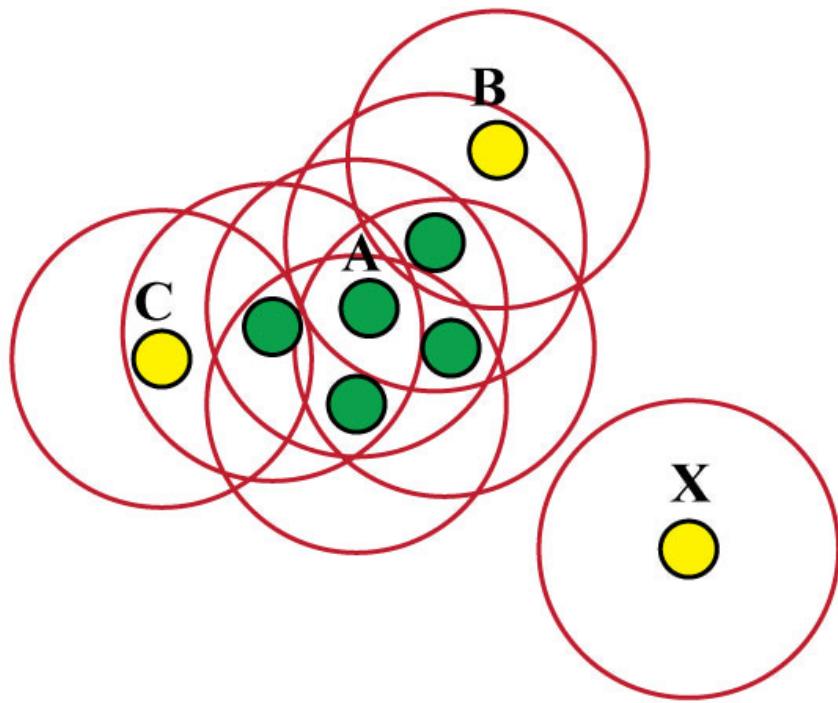


FIGURE 4.1: Clustering using DBSCAN

joining these core objects with their neighborhoods. The two parameters used in DBSCAN are ε and $minPts$. ε represents the radius and $minPts$ represents the minimum points required within ε . A point is called a core point if it has at least $minPts$ number of points within its ε . The points within the ε of a core point are called directly density reachable points. The point is called indirectly density reachable if it doesn't lie within the ε of a core point, but there exists a path of points from the core point to it. A point that is not reachable from the core point and is not a core point itself is called an outlier. The core points and the reachable points make the clusters.

In Fig. 4.1 suppose $minPts = 3$. The green points shown in the figure are core points since they have at least 3 points within ε radius. All these points make a single cluster since they are all reachable from one another. Points B and C are not core points but are indirectly reachable from point A, so they are also included in the same cluster. The point X, however, is an outlier since it is neither a core point nor is it density reachable.

4.3 DBSCAN for Bibliographic Coupling

In this section, we will discuss how the DBSCAN clustering can be used in Bibliographic Coupling. We will discuss the different modules used in this approach which include the data acquisition, normalization and similarity measuring etc. We will also discuss how we modified the traditional Bibliographic Coupling to determine a value of ε that could improve the accuracy of the paper recommendation.

4.3.1 Modules of Proposed Approach

The main modules of our system include: (1) Data Acquisition, (2) Data Normalization, and (3) DBSCAN Clustering and Similarity Score Measuring.

4.3.1.1 Data Acquisition Module

The first module is Data Acquisition Module. In this module, we gather data from CiteSeerX which is an online digital library as explained in Chapter 3. We used a focused web crawler to fetch the DOIs (Digital Object Identifier) for all the research papers that showed in the list of results, when we posed certain queries (keywords such as data mining, computer architecture etc.) on CiteSeerX. Using regular expressions and simple string matching, we also fetched the metadata elements such as title, authors, and year of publishing, number of received citations, abstract, context and references section. All of this information is stored in a database. After downloading the papers, we extracted the proximities of all the in-text citations.

4.3.1.2 Data Normalization Module

As we know, the length of scientific papers varies. Some papers may be 15,000 to 20,000 words long and others may be 3,000 to 5,000 words long. Comparing the

proximity of citations using the position of citations may lead to incorrect results. So, to fix this issue, we normalize the values of citations' positions within the full text of documents.

For our research, we use the Min-Max Normalization. This algorithm performs a linear transformation on the data values. For the interval $[MinX, MaxX]$ for a feature X, we use Min-Max normalization to transform it into a new interval $[New-MinX, New-MaxX]$. Similarly each value v in the original interval is converted into a new value $New-v$ using Eq. 3.1.

In case of the research papers in our approach, $MinX$ will be 1, $MaxX$ is the total number of words in the research paper, v is the position of the in-text citation, $New-MinX$ is 1 and $New-MaxX$ is set to 1000.

As explained in the upcoming sections, we performed experiments in order to find an optimal value for ε . In those experiments we applied our approach for different values of ε starting at 50. We increased the value by 50 going up to 1000. We found out that the number of the clusters produced for $\varepsilon = 700$ upto $\varepsilon = 1000$ was 1. Therefore, using the value of $New-MaxX = 1000$ helped us to perform experiments on different values of ε .

4.3.1.3 DBSCAN Clustering and Similarity Score Measuring

In the next step, the normalized positions of in-text citations are given as an input to the DBSCAN Clustering module. We used WEKA¹ to perform the clustering of in-text citations [75]. The WEKA machine learning workbench assists the researchers by providing a general purpose environment for the data mining tasks like classification, regression, and clustering etc. It consists of a humongous set of machine learning algorithms. The normalized positions of in-text citations produced in the previous step are provided as an input to WEKA. We provided the list of normalized in-text citations of the bibliographically coupled papers to WEKA.

¹<https://www.cs.waikato.ac.nz/ml/weka/> (accessed on 14 January 2018)

We used two different datasets for our experiments in the case of DBSCAN based approach. We used these datasets in two steps.

In the first step, we used a dataset containing 320 bibliographically coupled research papers as mentioned in Chapter 3. This dataset is further divided into 32 different subsets, each containing 10 research papers. Each subset consists of 9 research papers that are bibliographically coupled according to 1 query research paper. We applied the DBSCAN clustering algorithm on the proximities of in-text citations of this dataset. This algorithm uses two parameters for creating clusters. These parameters are ε and $minPts$. ε represents the radius and $minPts$ represent the number of minimum points required within the ε .

The minimum value of the normalized proximity of in-text citations in this dataset is 16 and the maximum value is 950. In order to determine the best value of ε , we used different values of ε on this dataset of 310 research papers.

We started with $\varepsilon = 1000$, which is higher than the maximum value of the normalized proximity of in-text citations. When we used this value, all the data points fell within one cluster. So we could not determine the rankings of bibliographically coupled research papers with this value. Then, we reduced the value of ε to 950, 900, 850, 800, 750 and 700. However, with these values of ε too, only 1 cluster was produced.

When we set the value of ε to 650, we got only one cluster for 30 subsets. In one subset though, we got 2 clusters. 8 of the 9 documents fell into first cluster and the 9th document was the only document in the second cluster.

In case of $\varepsilon = 600$, $\varepsilon = 550$ and $\varepsilon = 500$, 29 subsets produced only one cluster while 2 subsets produced 2 clusters. In case of $\varepsilon = 450$, 28 subsets produced only one cluster while 3 subsets produced 2 clusters. In case of $\varepsilon = 400$ and $\varepsilon = 350$, 25 subsets produced only one cluster while 6 subsets produced 2 clusters.

For $\varepsilon = 300$, 20 subsets produced only one cluster and 11 subsets produced 2 clusters. When we change the value of ε to 250, 12 subsets produced only 1 cluster, 18 subsets produced 2 clusters and 1 subset produced 3 clusters. For $\varepsilon =$

TABLE 4.1: Relationship between the number of clusters produced by subsets and the values of ε

Epsilon Values	1000	650	600	450	400	300	250	200	150	100	50
-	-	-	-	-	-	-	-	-	-	-	-
700	700	500	350								
1 Cluster	31	30	29	28	25	20	12	6	5		
2 Clusters		1	2	3	6	11	18	20	18	7	2
3 Clusters							1	5	7	12	2
4 Clusters										6	6
5 Clusters										6	7
6 Clusters											6
7 Clusters											6
8 Clusters											2

200, 6 subsets produced 1 cluster, 20 subsets produced 2 clusters and 5 subsets produced 3 clusters. In case of $\varepsilon = 150$, 18 subsets produced 2 clusters, 7 subsets produced 3 clusters, and 6 subsets produced only 1 cluster. In case of $\varepsilon = 100$, 7 subsets produced 2 clusters, 12 subsets produced 3 clusters, 6 subsets produced 4 clusters and 6 subsets produced 5 clusters. For $\varepsilon = 50$, 2 subsets produced 2 clusters, 2 subsets produced 3 clusters, 6 subsets produced 4 clusters, 7 subsets produced 5 clusters, 6 subsets produced 6 clusters, 6 subsets produced 6 clusters and 2 subsets produced 8 clusters. Table 4.1 shows the relationship between the number of clusters produced by subsets and the values of ε .

We also tried the values of ε below 50. But those produced even more clusters. The lower the value of ε gets, the more the number of clusters produced. Too many clusters mean each document can belong to a different cluster, the results of which match those in case of all the documents belonging to one cluster.

We used all the above mentioned values to determine the accuracy of DBSCAN algorithm. We compared the rankings produced from these variations with the rankings produced by the Jenson Shannon Divergence for the entire dataset of 320 research papers. We found out that the accuracy remained the same for $\varepsilon = 200$ to $\varepsilon = 1000$. And the accuracy remained the same for $\varepsilon = 50$ to $\varepsilon = 150$. The accuracy for $\varepsilon = 50$ to $\varepsilon = 150$ was higher than $\varepsilon = 200$ to $\varepsilon = 1000$. The average accuracy for $\varepsilon = 50$ to $\varepsilon = 150$ was 92% while the average accuracy for $\varepsilon = 200$

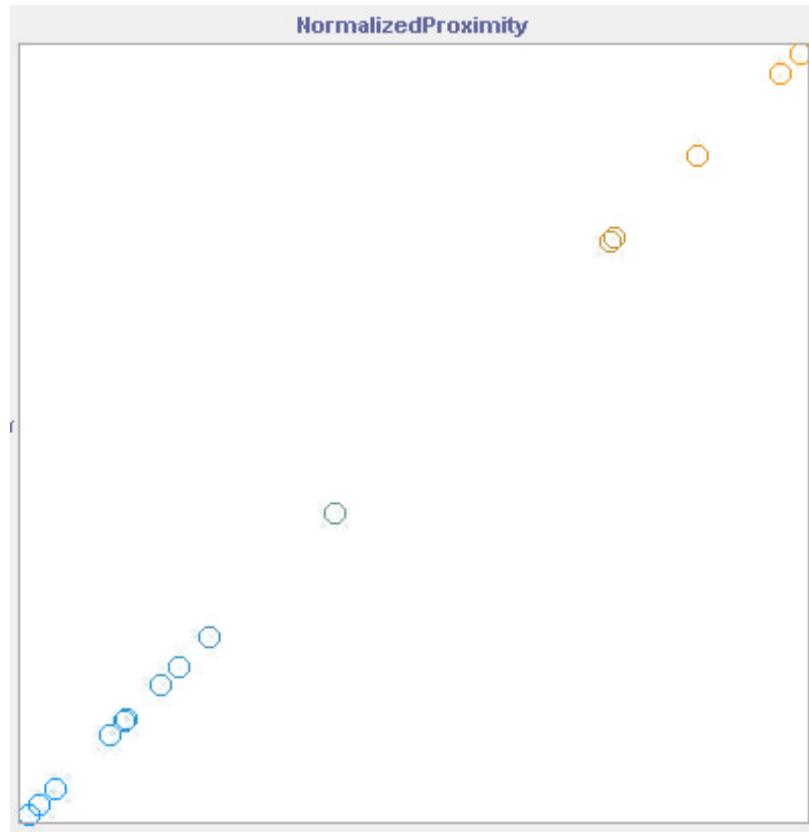


FIGURE 4.2: DBSCAN results produced by WEKA

to $\varepsilon = 1000$ was 70%. Therefore the average increase in accuracy for $\varepsilon = 50$ to $\varepsilon = 150$ was 31% compared to $\varepsilon = 200$ to $\varepsilon = 1000$.

For this experiment, $minPts = 2$ and $\varepsilon = 150$. As an output, we received the clusters. WEKA provides a visual output of clusters too. Fig. 4.2 represents the output of DBSCAN clustering for a sample dataset. We used this value of $\varepsilon = 150$ for applying DBSCAN clustering on the dataset-2.

Using these clusters, we calculate the similarity among the research papers, based on the proximity of the in-text citations. Suppose the papers 'x1', 'x2' and 'x3' cite papers 'A', 'B' and 'C' as shown in the Table 4.2. A1 represents the first citation of paper 'A' in papers 'X1', 'X2' and 'X3'. A2 represents the second citation of paper 'A' from the papers 'X1', 'X2' and 'X3' and so on.

From Table 4.2, we can see that points (X1,A1), (X2,A1), (X3,A1), (X1,A2), (X3,A2) and (X2, A2) are core points with respect to paper 'A'. These are all core

TABLE 4.2: DBSCAN clustering for citation proximity

	A1	A2	A3	B1	B2	B3	C1	C2	C3
X1	100	600	2000	400			700		
X2	200	1800					400		
X3	500	800	1500	600	800	1500			

points because they have 2 minimum points within a distance of 150. We get the following three clusters for paper 'A':

1. (X1,A1), (X2,A1)
2. (X3,A1), (X1,A2), (X3,A2) and
3. (X2,A2) and (X1, A3)

Similarly for paper 'B', we have the following clusters:

(X1,B1), (X3,B1), (X3,B2)

The points (X3, A3), (X3, B3), (X1, C1) and (X2, C1) are considered to be outliers. We can see from the clusters based on citations of paper 'A' that there are more citations from paper 'X1' and 'X2' that are within the same proximity as compared to citations from paper 'X3'. 'X1' and 'X2' are considered to be more similar to each other with respect to paper 'A'. Similarly based on citations of paper 'B', we can see that papers 'X1' and 'X3' are more similar to each other since they are within the same proximity, as compared to paper 'X2'. Suppose there are two more papers 'C' and 'D' and for both of them 'X1' and 'X2' turn out to be within same proximity as compared to paper 'X3', we will consider 'X1' and 'X2' to be more similar to each other.

4.4 Results and analysis

As we discussed in Chapter 3, the research paper recommendation approaches can be evaluated using three techniques [8]. The three evaluation techniques are: (1) user study, (2) online evaluation and (3) offline evaluation. Beel et al reviewed 176 papers and found that 69% of the approaches used offline evaluation, 34% used the user studies and 7% used the online evaluation for paper recommendation approaches.

User studies have been conducted by some researchers to evaluate paper recommendation systems [8], [25]. Therefore, we also evaluated the performance of our approach using the user study. We compared the performance of our approach with the traditional Bibliographic Coupling approach and with the content based approach.

We compared our approach with the traditional bibliographic coupling approach and with the content based approach using the dataset-1. For the user study we used the dataset-1 mentioned in Chapter 3. Every paper was evaluated by two distinct users. For each paper, the inter rater agreement was calculated between the users by using the Spearman's correlation coefficient [76].

Using this aataset, the proposed approach is compared with the bibliographic coupling approach and the content based approach. This comparison is also done using the Spearman's correlation coefficient. Fig. 4.3 shows the comparison between the proposed approach and the bibliographic coupling approach. We can see that there is higher correlation between the users' opinion and our proposed approach as compared to the traditional bibliographic coupling approach for majority of the documents used. For one document, both techniques performed the same. We can see that the average correlation improved by 22% in the proposed approach as compared to the traditional bibliographic coupling approach.

We compared the performance of our approach with content based approach too. We used the Spearman's coefficient for this comparison too. Fig. 4.4 shows the comparison between the two approaches. We can see from this figure that our

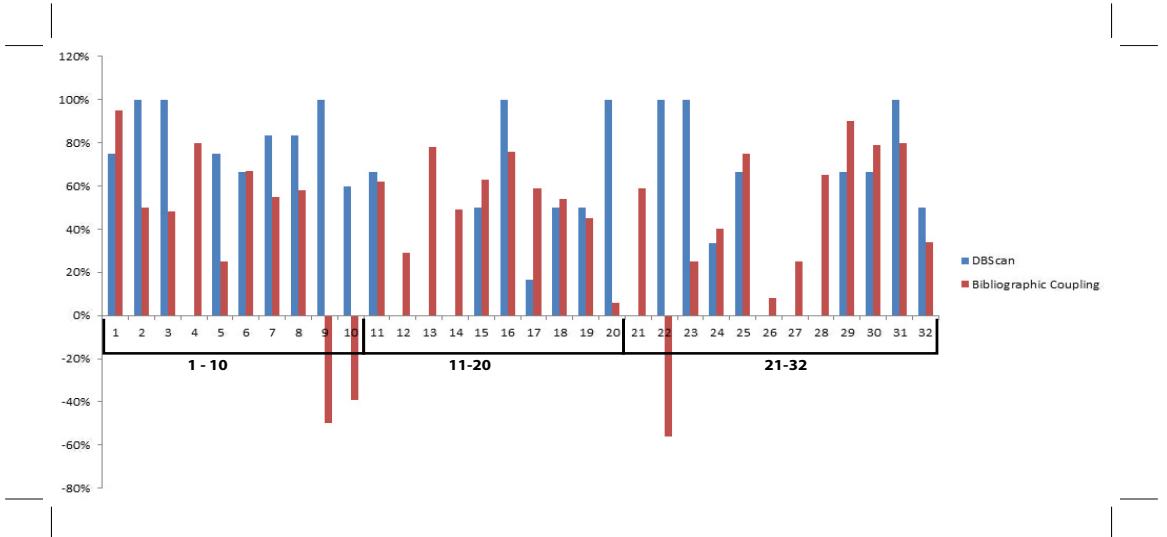


FIGURE 4.3: Proposed Approach vs Bibliographic Coupling

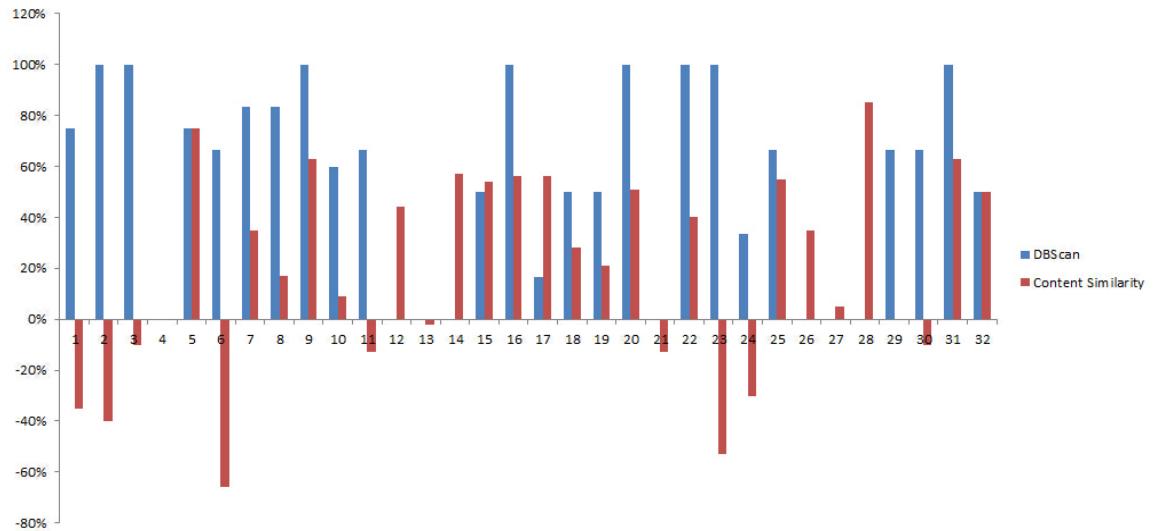


FIGURE 4.4: Proposed Approach vs Content Based Approach

proposed approach performed better for majority of the document as compared to the content based approach. The average correlation of the proposed approach with the dataset-1 was 0.55, whereas the average correlation of the content similarity with the dataset-1 remained as 0.20. We can see that the average correlation improved by 175% in the proposed approach as compared to the content similarity approach.

The above experiments and results have shown that by clustering the proximities of in-text citations using DBSCAN clustering algorithm in bibliographic coupling can achieve significant improvement as compared to the other approaches. Fig. 4.5

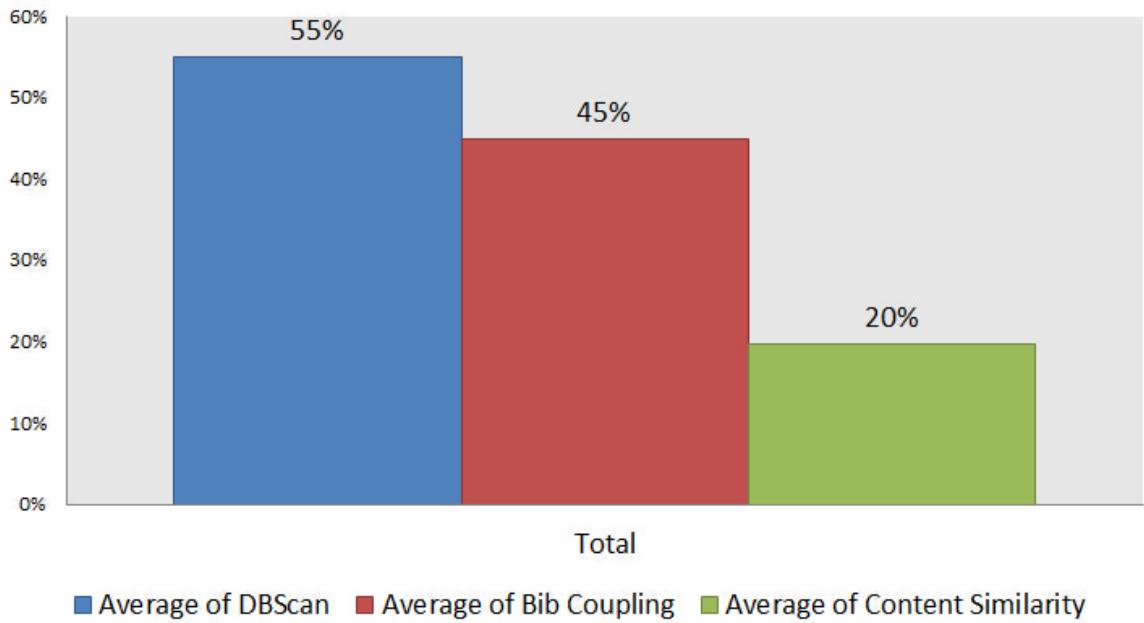


FIGURE 4.5: Performance of proposed approach

shows these results using a graph. In the X-axis, all three approaches are shown and in the Y-Axis, the values of correlation of each approach with the dataset-1.

Although user studies are a popular method of evaluating the research paper recommendation systems, they have certain limitations as well. For instance [77] found out that the results of user studies depend on what the users were asked for. That is whether they asked for the "perceived relevance" or "global satisfaction" of recommendations. Another study showed that the evaluation performed through user study gets affected by if the users were asked to rate the novelty or relevance of recommendations [78]. Another limitation for the user studies is that the user studies often need a lot of time as compared to offline evaluation. Another major drawback of user studies is that they cannot be performed in case of large datasets.

Considering the above mentioned drawbacks and limitations of user studies, we decided not to perform the user studies on the dataset-2 in case of DBSCAN based approach. We decided to use automatic evaluation on the dataset-2. We used the Jenson Shannon Divergence for this purpose. We will discuss the results of our experiments in the next paragraphs.

JSD finds the distance between two probability distributions. In case of research papers, the words distribution of individual research papers formed one probability distribution and the words distribution of the entire cluster formed the second probability distribution. We used the K-Means clustering algorithm to rank the distances produced by JSD. Then we used the Spearman correlation coefficient to determine the correlation between the rankings produced by JSD and those produced by the DBSCAN approach. According to Mukaka et al, there exists a strong correlation if the value of Spearmans correlation coefficient is between 0.7 and 1. The correlation is negligible if it is between 0.0 and 0.30. The correlation is low if it is between 0.30 and 0.50 and it is moderate if its value is between 0.50 and 0.70 [79]. We used the same ranges for the automatic evaluation of our approach. We compared our approach with traditional bibliographic coupling and content similarity as well. The following graphs provide an insight into how our approach behaved as compared to the other approaches for various different queries. The focus of these graphs is on the subsets that showed high correlation with the JSD i.e. the value of their correlation coefficient was higher than 0.7.

Fig. 4.6 shows a comparison between the correlations of the rankings produced by our proposed approach (DBSCAN based approach) and the content based and the traditional bibliographic coupling approaches.

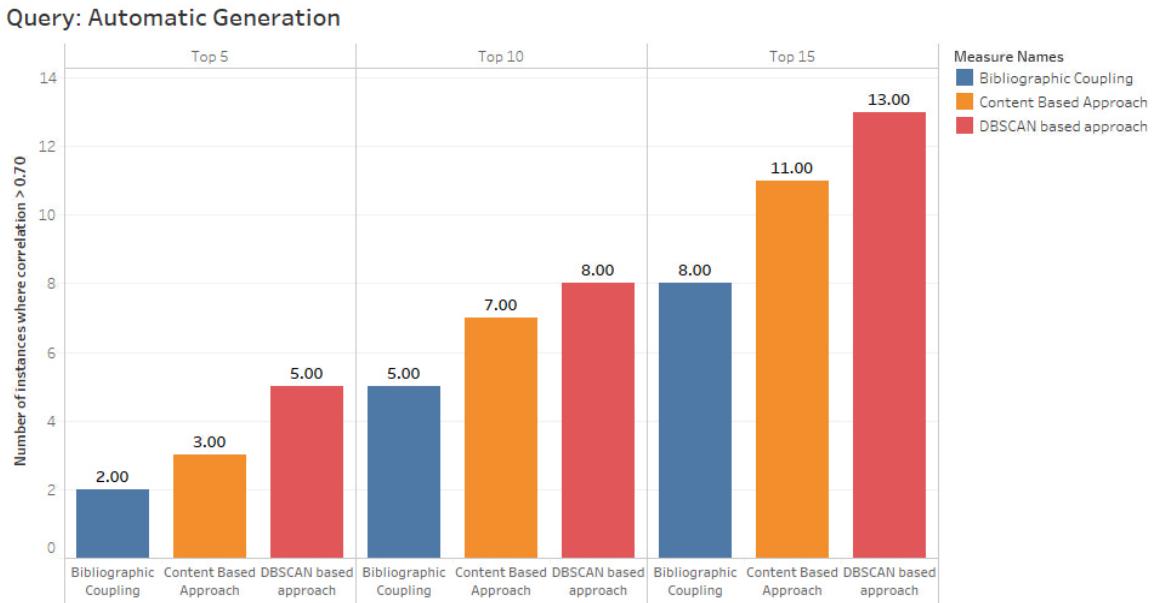


FIGURE 4.6: Comparative Evaluation of DBSCAN for query 'Automatic generation'

The graph shows the comparison for the search query 'Automatic Generation'. The X-axis represents the different approaches that are being compared. The X-axis also represents the Top 5, Top 10 and Top 15 segmentation of the correlation strengths based on the Spearman's Coefficient. The Y-axis represents the subset instances where the correlation is higher than 0.7. As can be seen from the Fig. 4.6, the DBSCAN approach performs better than the other two approaches in all cases for the given query.

We performed the similar analysis for the remaining 16 queries too. Table 4.3 shows the top 5 rankings for the number of instances where correlation was higher than 0.7 for all the 17 queries.

TABLE 4.3: Top 5 Rankings Comparison

Query	DBSCAN Based Approach	Content Based Approach	Bibliographic Coupling
Automatic generation	5	3	2
Bayesian networks	5	3	1
Black box testing	5	3	3
Collaborative recommendation	5	5	5
Content based filtering	5	4	2
Feature selection	5	5	5
Graph theory	5	5	5
Information retrieval	4	2	1
Online marketing	5	3	3
Opinion mining	3	3	1
Query processing	5	5	4
Recommendation system	5	5	5
Regression testing	5	3	1
Sensor networks	5	4	4
Social Network	5	3	2
Subjectivity analysis	4	4	3
Wireless communications	4	4	4

Table 4.4 shows the top 10 rankings for the number of instances where correlation was higher than 0.7 for all the 17 queries.

TABLE 4.4: Top 10 Rankings Comparison

Query	DBSCAN Based Approach	Content Based Approach	Bibliographic Coupling
Automatic generation	8	7	5
Bayesian networks	10	7	3
Black box testing	8	4	3
Collaborative recommendation	9	9	9
Content based filtering	7	5	2
Feature selection	9	10	10
Graph theory	9	10	9
Information retrieval	9	5	3
Online marketing	9	8	8
Opinion mining	7	7	5
Query processing	9	9	7
Recommendation system	9	9	7
Regression testing	10	7	3
Sensor networks	10	7	7
Social Network	10	4	4
Subjectivity analysis	9	7	5
Wireless communications	9	9	8

Table 4.5 shows the top 10 rankings for the number of instances where correlation was higher than 0.7 for all the 17 queries.

TABLE 4.5: Top 15 Rankings Comparison

Query	DBSCAN Based Approach	Content Based Approach	Bibliographic Coupling
Automatic generation	13	11	8
Bayesian networks	15	11	7
Black box testing	8	4	3
Collaborative recommendation	13	13	11
Content based filtering	7	5	2
Feature selection	14	14	15
Graph theory	13	14	14
Information retrieval	13	6	4
Online marketing	12	12	10
Opinion mining	12	12	10
Query processing	14	14	12
Recommendation system	14	12	10
Regression testing	15	12	8
Sensor networks	11	8	7
Social Network	15	4	7
Subjectivity analysis	13	10	6
Wireless communications	13	12	9

As we can see from the above discussion, our proposed approach produced much better results compared to the bibliographic coupling and content based approach. However, there are certain cases in which DBSCAN based approach may not

produce good results. One of the disadvantages of the DBSCAN based approach is that there are cases with border points that may be reachable from more than one cluster. Such points may become part of any cluster depending on the order in which DBSCAN performs clustering. DBSCAN is also not suitable for data that has high variance in density. DBSCAN clustering is also sensitive to the values of ε and $minPts$. But in the case of research paper recommendation, we discovered the optimal value for these parameters using comprehensive experiments.

DBSCAN produced better results compared to the other two approaches. Since there was less variance in the density of the in-text citation pairs in our dataset, our proposed approach produced good results. Our proposed approach performed better compared to the traditional bibliographic coupling because the bibliographic coupling uses only the bibliographic strength to recommend papers. It assigns a weight of 1 to all the in-text citation papers in the bibliographically coupled papers. However, in the case of DBSCAN based approach, the in-text citation pairs are assigned different weights depending on the clusters in which they fall and on how many clusters do they belong to.

4.5 Summary

In this chapter, we discussed our first approach i.e. the DBSCAN based approach. In this approach, firstly we extracted all the in-text citations and their proximities. In next step we performed an extensive experiment on dataset-1 to determine that the best value of ε that produced the most accurate recommendations was 150. Later on, this value of ε was used on our dataset-2.

In Fig. 4.7 shows a summary of comparison of the proposed approach (DBSCAN based), content based approach and the bibliographic coupling approach in case of top 5, top 10 and top 15 rankings. The X-axis represents the three categories i.e. Top 5, Top 10 and Top 15. The Y-axis represents the number of queries for which each approach outperformed the remaining two approaches. The three approaches are represented by the circles of different colors. As we can see from the figure

Number of queries for which each approach performed the best.



FIGURE 4.7: Total no. of queries for which each approach performed better than others in Top 5, Top 10 and Top 15 rankings.

4.24, our proposed approach performed significantly better than the other two approaches in the case of top 5, top 10 and top 15 rankings. There were 17 queries in total.

When considering the top 5 rankings, our proposed approach produced better results than the other two approaches in 9 out of 17 cases. The other two approaches didn't perform better than our proposed approach in case of any query. In the remaining 6 out of 17 cases, if one of the two other approaches won, our proposed approach shared the top position with it.

In case of top 10 rankings, our proposed approach performed better than the other two approaches in case of 10 queries out of a total of 17 queries. The content based approach produced more highly correlated instances than our approach and the bibliographic coupling approach in case of only one query.

In the case of top 15 rankings, our proposed approach produced instances that had a correlation of higher than 0.7 in case of 11 queries. The bibliographic coupling produced better results than our approach in case of only query. Content based

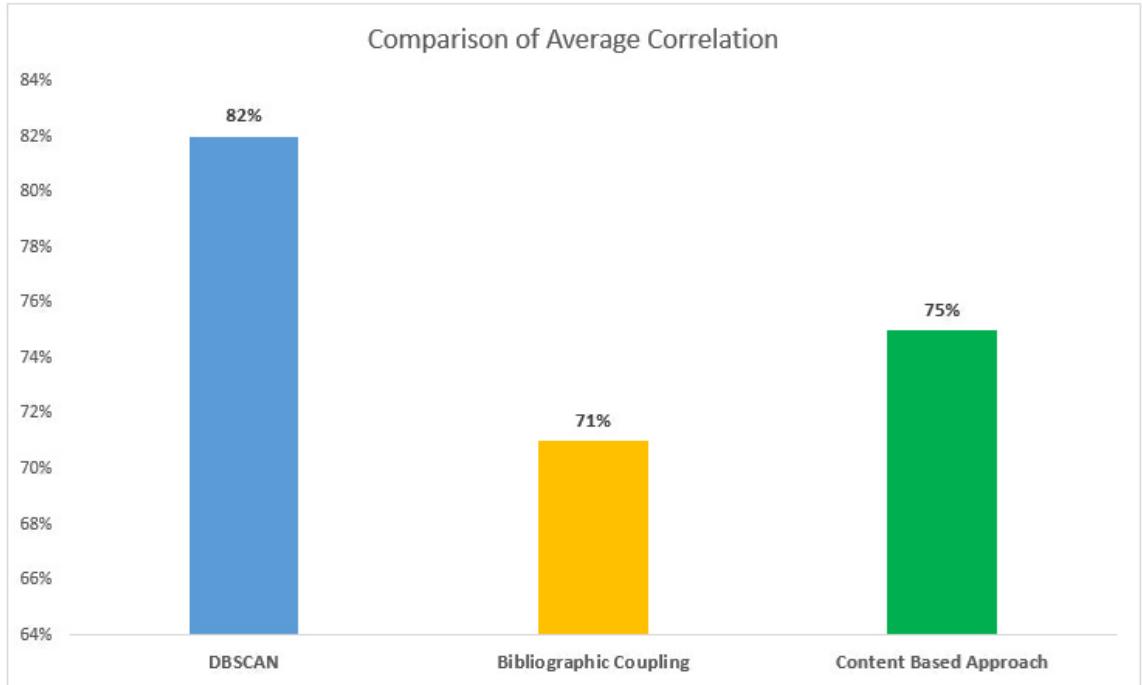


FIGURE 4.8: Average Correlation of All Approaches

approach could not produce better results than our proposed approach in case of any query.

Fig. 4.8 shows the comparison of the average correlations of the proposed approach with bibliographic coupling and content based approaches. As we can see our proposed approach performed better than the other two approaches. The average increase in accuracy of our proposed approach was 11% and 7% as compared to bibliographic coupling and content based approaches respectively.

Hence the following hypothesis discussed in the chapter 1 is proved: **The accuracy of research paper recommender systems based on Bibliographic coupling can be improved by exploiting the in-text citation occurrences and their proximities between the bibliographically coupled papers.**

Chapter 5

Centiles Based Approach

One version of this chapter has been submitted to the Journal of Information Science.

In this chapter we discuss one of the proposed approaches to incorporate in-text analysis into bibliographic coupling. This approach is inspired by a recognized approach in the area of co-citation analysis known as: Citation Proximity Analysis [9]. However, it was identified that there were some other variants of the same approach in the literature [1]. The main similarity in all these approaches is the proximity analysis of in-text citation patterns. The closer the co-citation exists, the more weight is assigned to those co-cited papers and, consequently, the more relevance is assigned to the co-cited papers. However, the definition of proximity (closeness) has been discussed by people with different approaches in many differing ways. In this chapter, we apply the similar concept of bibliographically-coupled papers with all different variants and have also proposed some new variants. These five different schemes are used to assign weights to in-text citation pairs. At the end of the chapter, we evaluate the proposed approach in comprehensive detail and compare it with state-of-the-art approaches, such as standard bibliographic coupling and a content-based approach.

5.1 Background

As we say in Chapter 1, more than 55 percent of approaches utilize the content of papers to recommend research papers. Similarly, citation-based approaches are considered very important in paper recommendation, since the citations are hand-picked by the authors themselves, which gives them a great potential in terms of recommending the papers to other researchers. Amongst citation-based approaches, co-citation is one of the most well-known [12]. The accuracy of co-citation has been increased by using content analysis in different approaches [1], [9].

Gipp et al. proposed a CPA-based approach for co-citation [9]. This approach analyzed the co-citation at certain levels such as, sentence level, paragraph level, and document level. The results showed that the relevance between two papers is higher if a common citing paper cites them from the same sentence. The relevance decreases if the common citing paper cites them from the same paragraph instead of the same sentence.

However, according to Boyack at al., the reference positions in the full text can be specified without the sentence, paragraph, and section demarcations [1]. Instead, they divided the documents into byte offsets and centiles. They divided the papers into 20 parts each consisting of 5 centiles. Four schemes (B,O,P1 and P2) were proposed for this purpose. Using the first scheme 'B', each co-citation pair is assigned a weight of 1. This scheme doesn't take the distance between the in-text citations into consideration. In the second scheme represented by 'O', if the two in-text citations are within the same byte position, they are assigned a weight of 4. If references are within 375, 1500 and 6000 bytes, they are given weights of 3, 2 and 1 respectively. If the distance is more than 6000 bytes, a weight of 0 is assigned. In the 3rd scheme P1, the paper's text is divided into 20 equal parts which are considered as 5 centiles. The weights are assigned based on these centiles. In the fourth scheme P2, the byte range of centiles is changed. The similarity between the two papers is then discovered based on these weights. The results showed that

the schemes P1 and P2 that used the centile locations performed better than the other two.

This improvement in accuracy by using the centile locations of in-text citations in co-citation analysis motivated us to explore the in-text citation occurrences, proximities and patterns in bibliographic coupling by using the similar and enhanced centiles based approach. The approach we propose clusters the in-text citations based on their centile positions.

5.2 Methodology

Initially we used the dataset-1 for this approach too. In this approach, firstly we find the position of the in-text citations. In the next step, we calculated the centile location of each in-text citation. In the next step, we calculate the distance between the centile values of all the in-text citations pairs. These values are stored in the database and are used by five different citation proximity schemes that cluster the citation percentile values using different thresholds. We used two weighting schemes which were proposed by Boyack et al [1]. And we used the dataset-1 in order to propose and evaluate three new weighting schemes for our proposed approach.

Fig. 5.1 shows the system architecture of centile-based proposed research paper recommender system. The main modules of the proposed approach are data acquisition, preprocessing, citations proximity schemes, and evaluation. This architecture is almost similar to the system architecture we discussed in the Chapter 3 with the difference being the citations proximity schemes.

As we discussed in Chapter 3, the data acquisition module is used to gather data from an online digital library called CiteSeer. We used 17 different queries to download bibliographically-coupled papers and their metadata. This data is stored in an SQL database. As discussed in previous chapters, we convert the papers into XML to gain access to the in-text citations and their positions. The

TABLE 5.1: Percentile values of in-text citations.

	No. of Words	No. of Centiles	Citation Position	Citation Centile
Paper A	2000	20	500	25
Paper B	3000	30	1800	60
Paper C	7500	75	3500	47

data normalization module is then used to normalize the values of these in-text citation positions. Since research papers vary in length, it becomes essential to normalize the proximity values of the in-text citations. In the first step, the data normalization module calculates the total number of words 'w' in the research papers. In the second step, 'w' is divided by 100 to find the total number of centiles 'p' in each research paper. 'p' can be given as mentioned in Eq. 5.1:

$$p = w/100 \quad (5.1)$$

For example, if three research papers 'A', 'B' and 'C' contain 2000, 3000 and 7500 words respectively and they cite a research paper 'D', the values of 'p' for 'A', 'B' and 'C' will be 20, 30 and 75 respectively.

In the next step, the proximity values of in-text citations are divided by the value of 'p' to determine the centile values of the in-text citations. In the case of the above example, if the positions of in-text citations from research papers 'A', 'B' and 'C' to the research paper 'D' are 500, 1800 and 3500 respectively, the centile values of in-text citations will be 25, 60 and 47 respectively. This means that the in-text citations from paper 'A', 'B' and 'C' to the paper 'D' belong to the 25th, 60th and 47th centiles respectively. The Table 5.1 shows the summary of this example.

In the above example, we can see that the distance between the values of citation percentiles of paper 'B' and paper 'C' is less compared to that of paper 'A' and 'B' or 'A' and 'C', so we can infer from this that papers 'B' and 'C' have more similarity between them compared to paper 'A'.

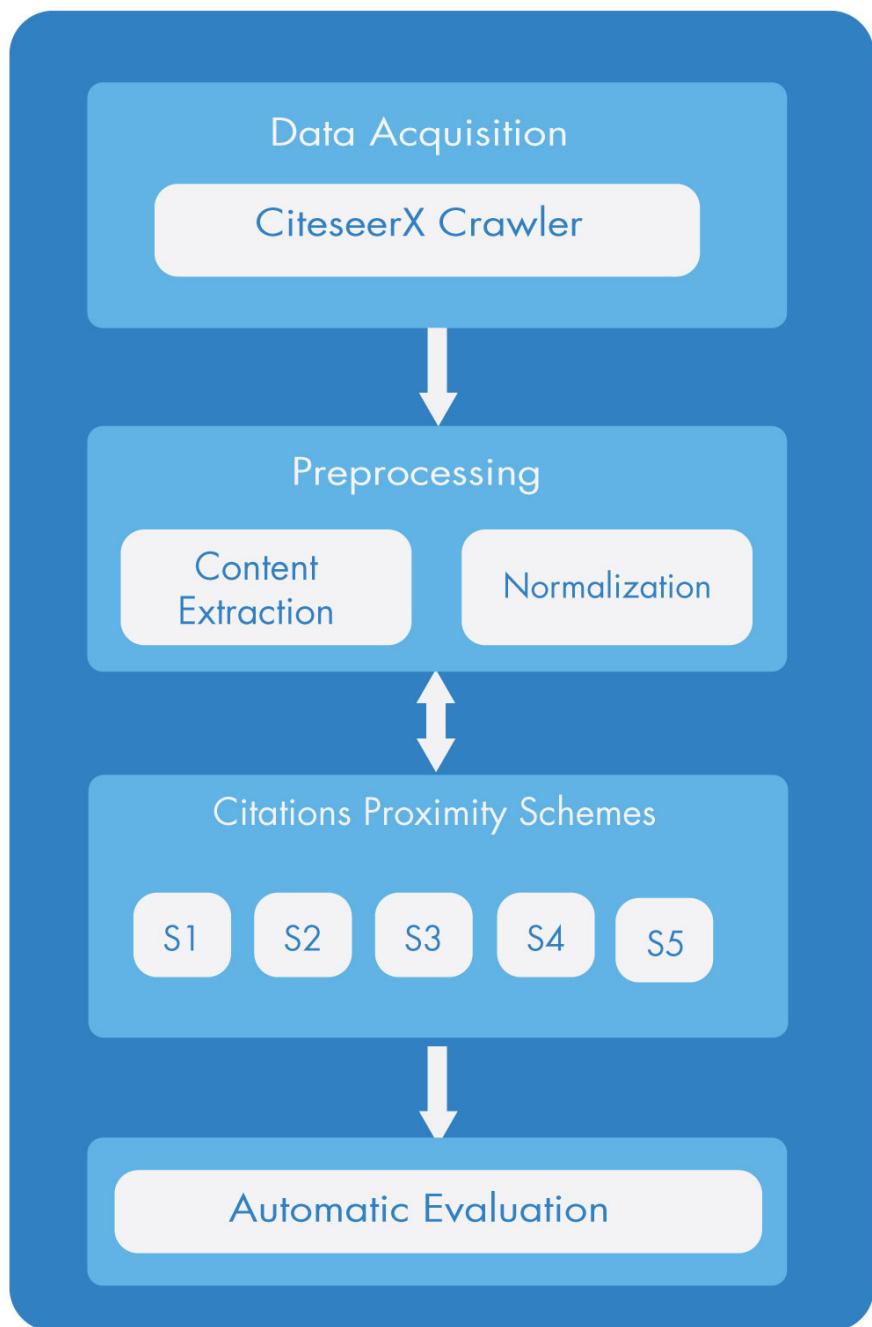


FIGURE 5.1: System Architecture for centiles based approach

These values are stored in the database and are used by five different citation proximity schemes that cluster the citation percentile values using different thresholds. We called these schemes S1, S2, S3, S4 and S5. The intuition behind these clustering schemes is that the value of similarity between two papers is larger if the distance between their citation percentile values is smaller and vice versa.

In the first scheme S1, the citing in-text citation pairs that are within 5 centiles are given a weight of 3. The citing in-text citation pairs that are within 15 and 25 centiles are given weights of 2 and 1 respectively. The citing in-text citation pairs that are more than 25 centiles apart are given a weight of 0. These in-text citation pairs were clustered according to their weights. The pairs having weights 3, 2, 1 and 0 were assigned to the clusters 3, 2, 1 and 0 respectively.

In the second scheme S2, the citing in-text citations pairs that are within 10 centiles are assigned a weight of 3. The citing in-text citation pairs that are within 25 and 40 centiles are given weights of 2 and 1 respectively. The citing in-text citation pairs that are more than 40 centiles apart are given a weight of 0. Like scheme S1, S2 assigns the citing in-text citation pairs to the clusters 3, 2, 1 and 0 according to their weights.

In the third scheme S3, k-means clustering algorithm (Hartigan et al, 1979) is used to assign the clusters to the in-text citation pairs. This algorithm aims at grouping the values into clusters such that each value belongs to a cluster with the nearest mean. We use WEKA to cluster the in-text citations. WEKA helps researchers perform a set of data-mining tasks such as classification, clustering and regression etc. on the datasets. We provide the list of centile distances between the in-text citation pairs to WEKA and it generates the clusters using the k-means clustering algorithm.

To determine the thresholds for the fourth and fifth schemes denoted by S4 and S5 respectively, we performed a detailed analysis of the in-text citations. We determined the values of all the centile differences. In the next step, for all the values of centile differences, we calculated the number of in-text citation pairs. From this analysis, we discovered that most of the in-text citation pairs are 4

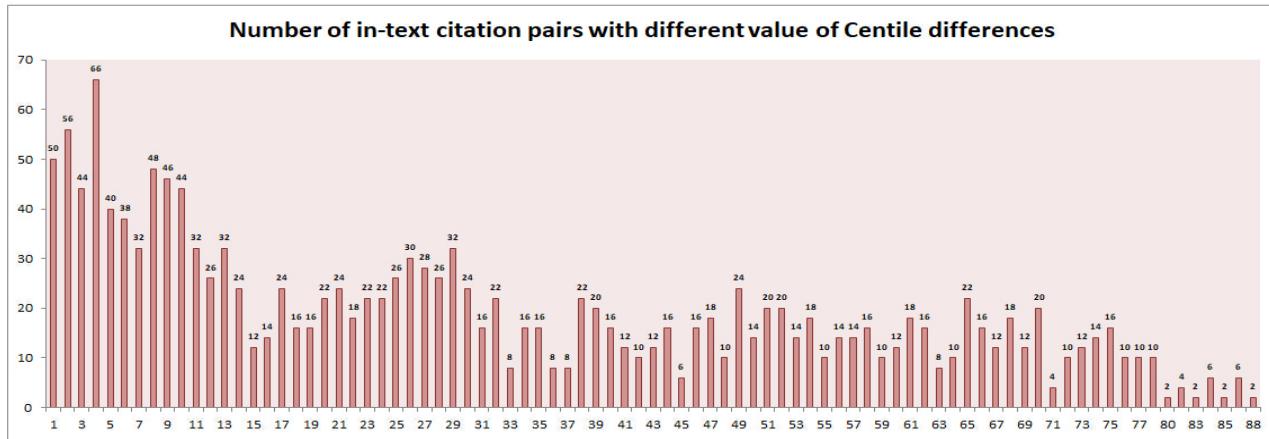


FIGURE 5.2: Number of in-text citation pairs with different value of centile differences

centiles apart. There were 56 in-text citation pairs that were 2 centiles apart. Fig. 5.2 shows the number of in-text citation pairs with different value of centile differences. S1 and S2 use the predetermined thresholds without considering how the in-text citations have been used in the research papers. Unlike S1 and S2, the proposed schemes S4 and S5 determine the thresholds and weights based on the actual distribution of the in-text citations in the full text of the paper.

In the next step, we applied K means clustering on these values. We clustered the centile ranges into four clusters, based the number of in-text citations falling in each range. Using this clustering, we determined the most frequent centile ranges in our data set. From each cluster, we selected the top 2 most frequent centile ranges i.e. the centile ranges with the largest number of in-text citations. Based on this analysis, we determined the two schemes denoted by S4 and S5.

In the scheme S4, the citing in-text citations pairs that are within 4 centiles are assigned a weight of 3. The citing in-text citation pairs that are within 7 and 22 centiles are given weights of 2 and 1 respectively. The citing in-text citation pairs that are more than 22 centiles apart are given a weight of 0. Like S1 and S2, S4 assigns the citing in-text citation pairs to the clusters 3, 2, 1 and 0 according to their weights.

In the scheme S5, the citing in-text citations pairs that are within 2 centiles are assigned a weight of 3. The citing in-text citation pairs that are within 11 and

TABLE 5.2: Weights for S1, S2, S4 and S5

S1	S2	S4	S5	Weights
<5	<10	<4	<2	3
5 to 15	10 to 25	4 to 7	2 to 11	2
16 to 25	25 to 40	8 to 22	12 to 47	1
>25	>40	>22	>47	0

47 centiles are given weights of 2 and 1 respectively. The citing in-text citation pairs that are more than 47 centiles apart are given a weight of 0. The scheme S5 assigns the citing in-text citation pairs to the clusters 3, 2, 1 and 0 according to their weights.

The Table 5.2 depicts a summary of schemes S1, S2, S4 and S5 and their corresponding weights for citations pairs in different percentile ranges.

In the above mentioned schemes, the in-text citation pairs are clustered based on their similarity with each other. The papers that are the most related to each other are assigned to the cluster 0 in all approaches. The papers that are the least related to each other are assigned to the cluster 2 in all approaches. The cluster 1 contains those papers that are less related to each other as compared to those in cluster 0 and more related to each other as compared to those in cluster 2.

We applied S1, S2, S3, S4 and S5 scheme on dataset-1 in order to determine the best scheme. As mentioned in the Chapter 3, dataset-1 consists of 320 papers divided into 32 subsets. All five schemes were used to rank the ten papers in each subset. In the second step, Jenson Shannon Divergence was used to rank the same paper. In order to compare the five schemes, we used the Spearman's correlation coefficient.

The results showed that scheme S5 performed better than the remaining 4 schemes. The value of average correlation for S5 was 0.93. Since S5 performed better than the remaining 4 schemes, we preferred to use this scheme on the dataset-2.

5.3 Results and Evaluation

In this section, we discuss the evaluation and results of the proposed approach. As explained in previous chapters, user study is not a feasible option for evaluation of paper recommendation systems in case of large datasets. Therefore, in order to evaluate the clusters produced by our proposed approach, we used the automatic approach i.e. the Jenson Shannon Divergence. In the next step we compared the results of JSD with those of our proposed approach using the Spearman's correlation coefficient. We compared our proposed approach with the traditional bibliographic coupling and the content similarity approach.

We explain performance of our approach in comparison with the other two approaches with the help of the following graphs. These graphs show the subsets that showed high correlation with the JSD when the value of their correlation coefficient was higher than 0.7.

Fig. 5.3 shows a comparison between the correlations of the rankings produced by our proposed approach (centiles based approach) and the content based and the traditional bibliographic coupling approaches.

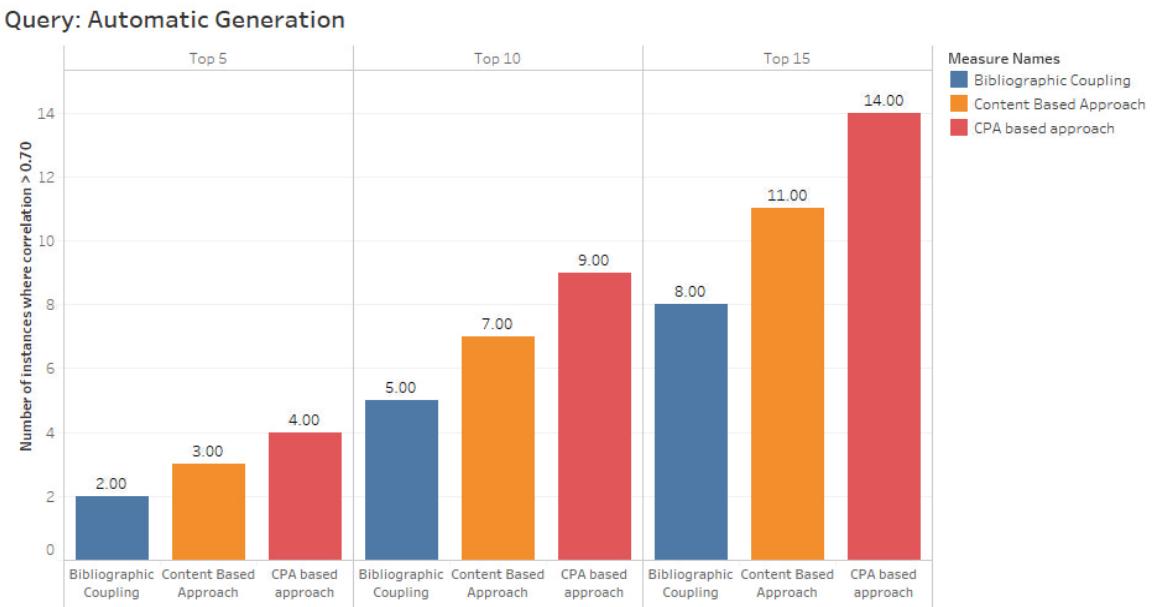


FIGURE 5.3: Comparative Evaluation of centiles based approach for query 'Automatic Generation'

TABLE 5.3: Top 5 Rankings Comparison

Query	Centiles Based Approach	Content Based Approach	Bibliographic Coupling
Automatic generation	4	3	2
Bayesian networks	5	3	1
Black box testing	5	3	3
Collaborative recommendation	5	5	5
Content based filtering	5	4	2
Feature selection	5	5	5
Graph theory	5	5	5
Information retrieval	5	2	1
Online marketing	4	3	3
Opinion mining	5	3	1
Query processing	5	5	4
Recommendation system	5	5	5
Regression testing	5	3	1
Sensor networks	5	4	4
Social Network	5	3	2
Subjectivity analysis	4	4	3
Wireless communications	4	4	4

Fig. 5.3 shows the results for the query 'automatic generation'. The graph has all the approaches and the three categories i.e. top 5, top 10 and top 15 rankings along the X-axis. The Y-axis represents the total number of instances or subsets for which the value of Spearman's correlation coefficient is higher than 0.70. In figure 5.3, we can see that our proposed approach performed better than the other two approaches in case of top 5, top 10 and top 15 rankings. Our proposed approach produced more subsets with value of correlation higher than 0.70.

We performed the same analysis for all the remaining 16 queries too. Table 5.3 shows the top 5 rankings for the number of instances where correlation was higher than 0.7 for all the 17 queries.

Table 5.4 shows the top 10 rankings for the number of instances where correlation was higher than 0.7 for all the 17 queries.

TABLE 5.4: Top 10 Rankings Comparison

Query	Centiles Based Approach	Content Based Approach	Bibliographic Coupling
Automatic generation	9	7	5
Bayesian networks	10	7	3
Black box testing	8	4	3
Collaborative recommendation	10	9	9
Content based filtering	7	5	2
Feature selection	10	10	10
Graph theory	10	10	9
Information retrieval	10	5	3
Online marketing	9	8	8
Opinion mining	10	7	5
Query processing	10	9	7
Recommendation system	10	9	7
Regression testing	10	7	3
Sensor networks	10	7	7
Social Network	9	4	4
Subjectivity analysis	8	7	5
Wireless communications	9	9	8

Table 5.5 shows the top 15 rankings for the number of instances where correlation was higher than 0.7 for all the 17 queries.

TABLE 5.5: Top 15 Rankings Comparison

Query	Centiles Based Approach	Content Based Approach	Bibliographic Coupling
Automatic generation	14	11	8
Bayesian networks	15	11	7
Black box testing	8	4	3
Collaborative recommendation	15	13	11
Content based filtering	7	5	2
Feature selection	15	14	15
Graph theory	15	14	14
Information retrieval	14	6	4
Online marketing	14	12	10
Opinion mining	15	12	10
Query processing	15	14	12
Recommendation system	15	12	10
Regression testing	15	12	8
Sensor networks	12	8	7
Social Network	13	4	7
Subjectivity analysis	11	10	6
Wireless communications	13	12	9

Since the centiles based approach uses the centile values of the in-text citations, this approach doesn't need the min-max normalization of the research papers. Since this approach also uses the proximities of the in-text citations within the full text of the paper, it produces better results compared to the traditional bibliographic coupling and the content based approach. However there were certain cases in which the accuracy of bibliographic coupling or the content based approach was the same as the centiles based approach. These were mostly the cases in which there was a larger centile difference between most of the in-text citation pairs.

5.4 Summary

In this chapter, we discussed one of the proposed approaches i.e. the centiles based approach. In first step, firstly we found the positions of the in-text citations and the distance between the centile values of all the in-text citations pairs were calculated. We used two weighting schemes (S1, S2) which were proposed by Boyack et al [1]. And we used the dataset-1 in order to propose and evaluate 3 new weighting schemes (S3, S4 and S5) for our proposed approach. S5 performed better than the remaining 4 schemes. Therefore, we preferred to use this scheme on the dataset-2.

Fig. 5.4 shows a summary of comparison of the proposed approach (centiles based approach), content based approach and the bibliographic coupling approach in case of top 5, top 10 and top 15 rankings. The X-axis represents the three categories i.e. Top 5, Top 10 and Top 15. The Y-axis represents the number of queries for which each approach outperformed the remaining two approaches. As we can see from the figure 5.20, the proposed approach performed significantly better than the other two approaches in the case of top 5, top 10 and top 15 rankings.

There were 17 queries in total. Our proposed approach produced better results than the other two approaches in 10 out of 17 queries in case of top 5 rankings. In the remaining 7 out of 17 queries too, the other two approaches didn't perform better than the proposed approach in any of the case. In these 7 cases either

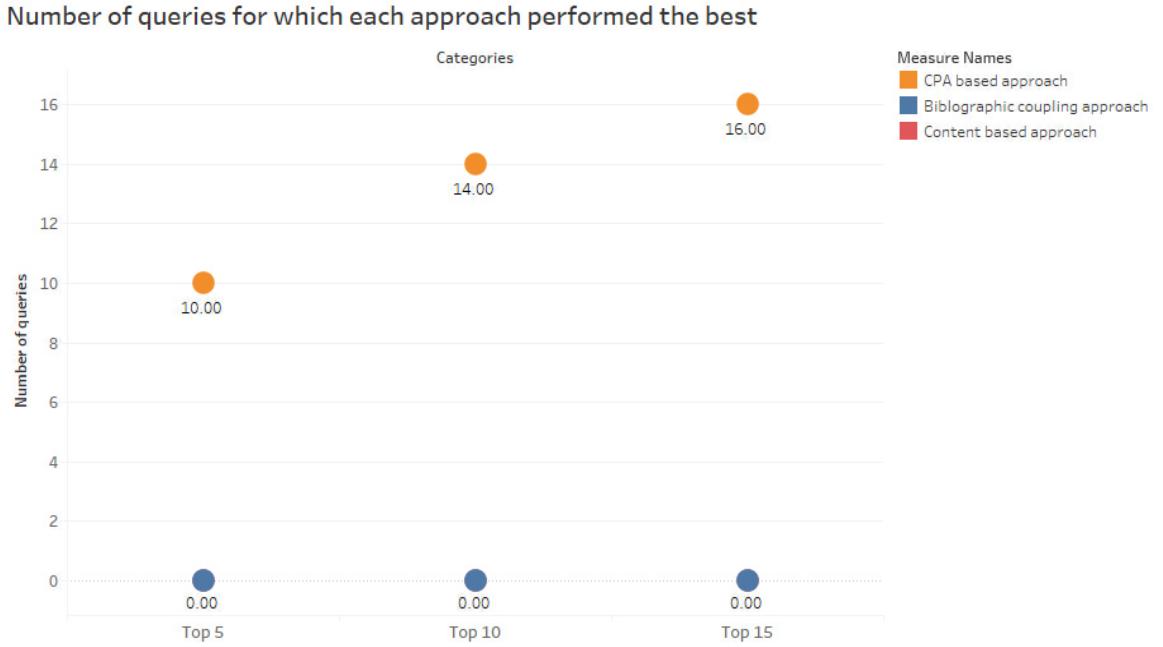


FIGURE 5.4: Total no. of queries for which each approach performed better than others in Top 5, Top 10 and Top 15 rankings.

bibliographic coupling or the content based approach or both were on top but our proposed approach also share the top spot with them in those queries too.

Similarly, in the case of top 10 rankings, our proposed approach produced better results than the other two approaches in 14 out of 17 queries. Just like in case of top 5 rankings, the other two approaches could not outperform our proposed approach in any query. Our proposed approach shared the top spot with them if any of them was on the top in any query.

In the case of top 15 rankings, our proposed approach produced instances that had a correlation of higher than 0.7 in case of 16 queries. The other two approaches could not outperform our proposed approach in any query in case of top 15 rankings.

Fig. 5.5 shows the comparison of the average correlations of the proposed approach with bibliographic coupling and content based approaches. As we can see our proposed approach performed better than the other two approaches. The average increase in accuracy of our proposed approach was 12% and 8% as compared to the bibliographic coupling and content based approach.

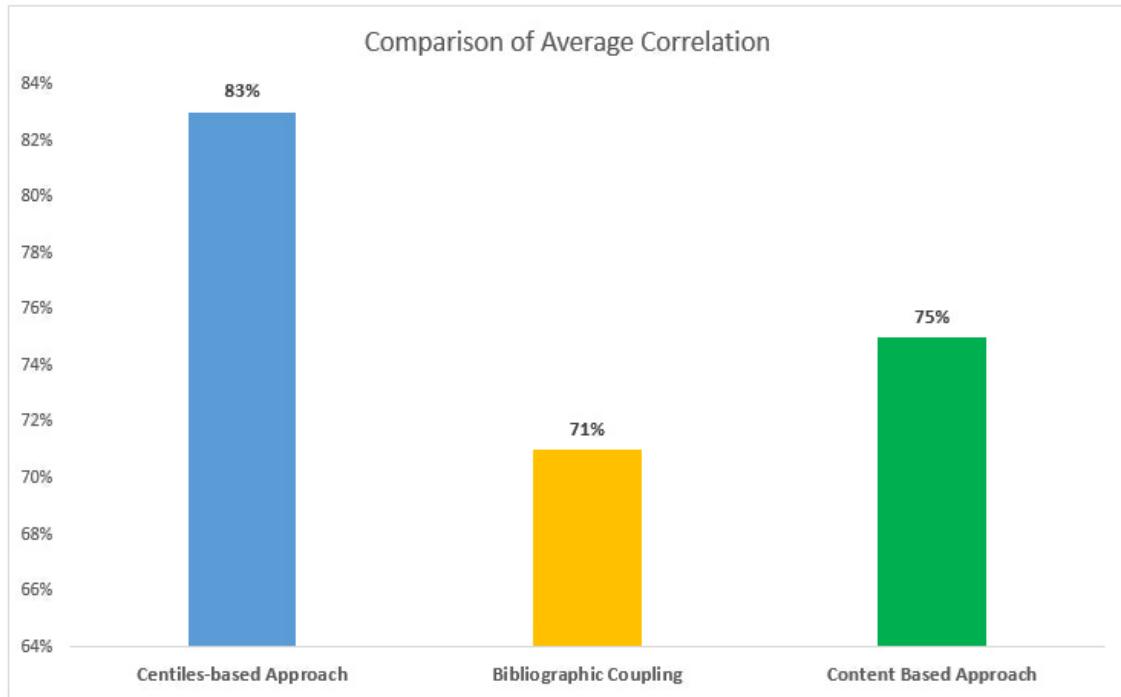


FIGURE 5.5: Average correlations of all queries

Hence the following hypothesis discussed in the chapter 1 is proved: **The accuracy of research paper recommender systems based on Bibliographic coupling can be improved by exploiting the in-text citation occurrences and their proximities between the bibliographically coupled papers.**

Chapter 6

Section Based Bibliographic Coupling

One version of this chapter has been submitted to Scientometrics.

In this chapter, we have discussed our third proposed approach to exploit the content of the papers for bibliographic coupling. The importance of sections has been discussed in the context of citation analysis. Furthermore, the detailed methodology has been formulated and described for the proposed approach. In the end, the evaluation has been shown and comparisons with state-of-the-art approaches have been depicted.

6.1 Background

Studies have shown that almost all the authors follow a certain set of procedural standards when referencing other papers [69], [80]. For example, most relevant papers are cited in the methodology and result sections. Papers belonging to background knowledge are normally cited in the introduction or related work section. This makes the exploration of logical structure of research papers an area of interest for many researchers. In recent times, many researchers have shown interest in exploring the importance of position of in-text citations within the full content

of research papers. The availability of full-text of research papers has made it possible for the researchers to develop innovative approaches for citation analysis, citation recommendation and paper recommendation [70], [81], [82]. This full-text access to research papers has also provided possibilities for studying the distribution of in-text citations in the full content of research papers. Many researchers have shown interest in the localization of in-text citations in past as well. For example, Voos and Dagaev (1976) conducted a manual study in order to find out if two citations can be given the same weight during citation analysis [83]. They used a very small dataset consisting of four research papers.

McCain and Turner proposed the idea for the first time that the section structure plays an important role in determining the function of in-text citations [84]. They studied and analyzed the in-text citations in different sections and proposed a scheme to assign different weights to the citations from different sections. Similarly Marii et al. analyzed set of 357 research papers and concluded that location of in-text citations, along with their level and age, plays a vital role in citation analysis [85]. They suggested that the in-text citations belonging to different sections have different values. Based on their analysis, they assigned different weights to different sections (Introduction: 10, Methods: 30, Results: 30, Discussion: 25). Another study [70] highlights the fact that authors normally tend to prefer certain sections over the others while distributing the in-text citations. According to this study the Introduction section contains the largest number of in-text citations. The Literature Review section makes for the second most citing section followed by the Methodology section.

The way, an author distributes the in-text citations in a particular fashion across the full text of a research paper, is subjective. But studies show that the authors follow a set of norms and procedural standards when distributing the citations in the citing papers [69], [80]. According to a study, the highly cited papers get cited the most from the Introduction section [70]. The Literature Review section makes for the second most citing section followed by the Methodology section. Moreover, the citations from the Results and the Methodology sections are more important as compared to those from the Related Work section [73], [25]. The papers that

are cited from the Results and the Methodology sections are usually more relevant to the citing papers. The authors usually cite the most relevant papers in these sections. However, the Related Work section and the Introduction may contain the citations to generic papers which may not be very relevant to the citing papers.

As we discussed above, some sections are more important than the others in terms of the in-text citations. For example, the citations in the methodology sections are more important compared to those from the related work section. However, in the case of the centiles based approach, the logical sections of the paper are ignored. For example a citation pair may have smaller distance but they may actually exist in two different sections. Such a citation pair may get a higher weight despite the fact that they may exist in two sections that have relatively little importance. The centiles based approach may not produce good results in such cases.

As shown above, a lot of work has been done in the past and in the recent time to show that authors follow a certain pattern when distributing the in-text citations. Different weighting schemes for sections have been proposed as well. However, not much research has been done to exploit the distribution of citations in sections in the context of citation analysis. In this chapter, we propose a paper recommendation system approach that exploits the sections in bibliographically coupled papers to recommend relevant papers.

6.2 Methodology

Fig. 6.1 shows the system architecture for this approach. The important modules for this system are Data Acquisition, XML Conversion, Sections Extraction and Similarity Score Measuring. In the next sub-sections, we will discuss each of these modules in details.

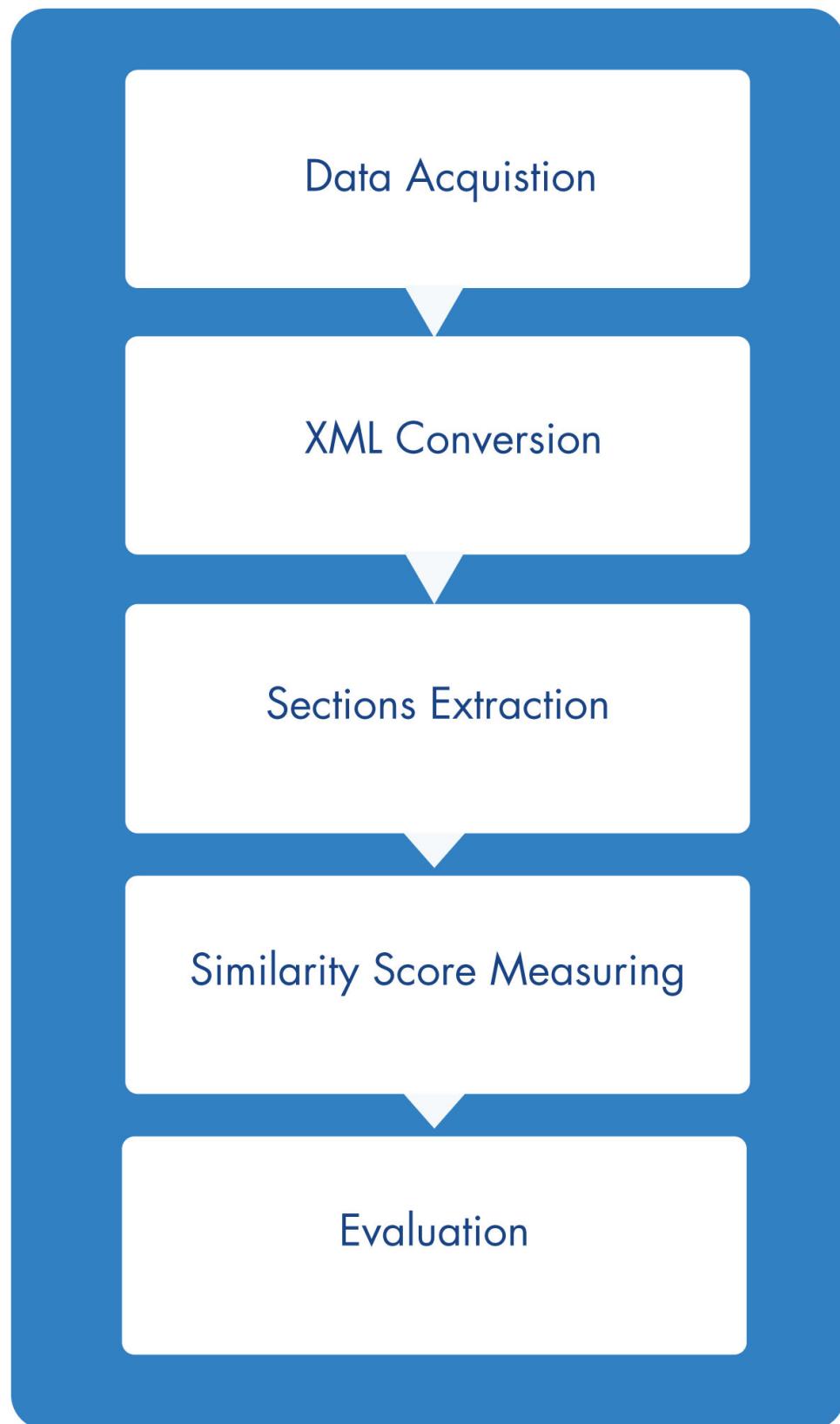


FIGURE 6.1: System Architecture for Section Based Bibliographic Coupling

6.2.1 Data Acquisition

This module is used to collect two datasets for our experiment. The details of these two datasets i.e. dataset-1 and dataset-2 have been discussed in Chapter 3. As discussed in Chapter 3, we used 17 different queries to collect a dataset called dataset-2 that provides good coverage of different sub-domains of computer science. Using a crawler, we fetched the metadata for these research papers that includes DOIs, abstract, author-names, and venue etc. We stored all this information in an SQL database. To download all of these bibliographically coupled papers, we built a utility using Python. These research papers were downloaded in PDF format.

6.2.2 XML Conversion

Since the web crawler downloaded all the papers in PDF format, they needed to be converted into XML format in order to fetch the information related to sections and in-text citations. A freely available online tool called PDFx was used to convert our dataset of 5000 research papers in PDF format to XML format. PDFx is a tool designed specifically for conversion of scientific articles [68]. The converted XML files contain some very important elements such as section, ref and xref etc. The element xref with the attribute ref-type="bibr" represents the in-text citations and can be linked to the `jref` tags through the attribute rid.

6.2.3 Section Extraction

The XML documents from the previous module are passed on to the Section Extraction module. This module extracts the sections from the research papers using the special elements inside the XML documents denoted with the tag `jsection`/`section`. This section element refers to all the sections inside the research paper. This element consists of a nested heading element denoted by `jh1`/`h2`. This heading tag refers to the heading of each section. PDFx provides two more

levels of heading element i.e. h2 and h3. This module uses the Document Object Model (DOM) to traverse the XML files and to fetch the section headings.

Studies show that normally the research papers are organized in a standard way and contain specific sections. Studies show that most of the research papers contain certain sections [71], [72]. These sections are given as follows:

1. Introduction
2. Related Work
3. Architecture/Methodology
4. Results/Comparisons
5. Conclusion/Future Work

These studies helped us to determine the main sections for our research too and we decided to use the same main sections as mentioned above. Using the section element, we fetched the sections from the research papers. In order to map these fetched sections to the sections mentioned above, we used the suggestions of a study conducted by Ding et al [70]. Using this study, we can infer that the Introduction section contains the largest number of in-text citations followed by the Literature review section that contains the second largest number of in-text citations followed by the Methodology section. The sections with the fourth and fifth largest number of citations are Results and Conclusion respectively. After extracting the sections and the in-text citations from each section, we mapped the sections to the generic sections mentioned above, using the frequencies of the in-text citations.

In order to verify the section mapping of our system, we conducted a user study. We used the dataset-1 for this purpose. As we explained in the Chapter 3, the dataset-1 consists of 32 different subsets with 10 bibliographically coupled papers in each subset. The dataset was assigned to two experts who have advanced experience and knowledge in the field of Computer Science. The two experts, we assigned the dataset-1 to, were pursuing their PhDs in the area of paper recommendation using citation analysis as well. This made them the perfect candidates

for this user study, since they have the knowledge and hands on experience of the citations, research paper sections, and paper similarity.

The experts were assigned the task to manually map the sections of the papers in the dataset-1 to the generic sections that we mentioned above. This mapping produced by the experts was then compared with the mapping produced by our system. For this purpose we used the Spearman rank correlation coefficient. Other correlation coefficients like Pearson's coefficient and Kendall's Tau coefficient could have been used too. But we preferred to use the Spearman's coefficient because, unlike the other two above mentioned correlation coefficients, it doesn't need to make the assumption that the two variables are linearly related to each other. Moreover, it doesn't need the variables to be measure on interval scales [86].

The value of Spearman's correlation ranges between 0 and 1. Its value was 0.85 for the correlation between mappings produced by our system and those produced by the experts. According to Mukaka et al, there exists a strong correlation if the value of Spearmans correlation coefficient is between 0.7 and 1 [79].

Since there was high correlation between the mappings produced by our system and those by the experts in case of the dataset-1, we decided to use the same mapping criteria for the larger dataset i.e. dataset-2. Since the dataset-2 contains almost 5000 papers, it was not feasible to conduct user study for the dataset-2. However, we manually cross-checked randomly selected 100 papers from this dataset too, and found that the sections have been mapped with 90% accuracy.

There may be cases where the sections could not extracted correctly. So this approach may not produce good results in such cases. Authors can use different names for the same sections. Similarly, some digital libraries, conferences and journals may have predetermined sections to be used which may be different from those used by the others. Such cases can prove to be challenging.

6.2.4 Similarity Score Measuring

Many researchers have analyzed the distribution of in-text citations in research papers and their research suggests that the citations from different sections should be given different weights during citation analysis [73], [25]. These studies show that the citations from each section carry a different weight and have a different meaning. For example, the citations from Related Work and Introduction usually mean that the cited document might be a supporting document. The documents cited from the methodology and results sections, however, tend to be the most closely related ones. Similarly, the documents cited from the Related Work are considered to be the least important ones, since the Related Work may contain less related and more generic kind of citations too.

Considering the results of previous studies [73], [25], the relation among the weights of different sections can be given by the following equation:

$$\text{weight}(m/rs) > \text{weight}(i) > \text{weight}(rw) \quad (6.1)$$

In Eq. 6.1, $\text{weight}(m)$ denotes the weight of methodology section, $\text{weight}(rs)$ denotes the weight of results section, $\text{weight}(i)$ denotes the weight of introduction section and $\text{weight}(rw)$ denotes the weight of related work section. As is obvious from the above equation, the in-text citations from the methodology and results section are given more weight than those from the introduction section. And the in-text citations from the related work carry the least weight. We determined the weights for different sections in two steps. In the first step, we used the Jenson Shannon Divergence to rank the papers in dataset-1. We have explained the working of JSD in Chapter 3 and Chapter 4. In the second step, we generated the rankings for the dataset-1 using our system. For this purpose, we initialized the weight of Related Work section with a value of 1 and changed the weights of other sections by increasing the value by 0.5 for same sections and 0.2 for cross sections. We used the Spearman's coefficient to determine the correlation between our rankings and the rankings produced by the JSD for all the different weights of

TABLE 6.1: Weights of different sections.

Section	Weight
Methodology	3
Results	3
Introduction	2
Related Work	1

	Introduction	Related work	Methodology	Results
Introduction	2	0.5	0.24	0.24
Related work	0.5	1	0.24	0.24
Methodology	0.24	0.24	3	3
Results	0.24	0.24	3	3

FIGURE 6.2: Weights for citations from the sections and cross sections.

the sections. We found out that the weights mentioned in the Table 6.1 produced the best results. The value of correlation for these values of weights was 0.8.

Table 6.1 represents the weights for the citations from the same sections. For example if paper 'A' and paper 'B' cite a common paper from the Methodology chapter, the weight will be 3. The weights for citations from the cross sections were calculated in the same way as mentioned above. The weights for same section and cross sections citations are shown in Fig. 6.2.

In Fig. 6.2, the sections mentioned along the Y-axis represent the sections of paper 'A' and the sections mentioned along the X-axis represent the sections of paper 'B'. If paper 'A' and paper 'B' cite a common paper from sections Introduction and Results respectively, the value of weight will be 0.24.

6.3 Evaluation

In this section, we will discuss the evaluation of our proposed approach. In order to evaluate the accuracy of our proposed approach we compared its performance with the content based paper recommendation and the traditional bibliographic coupling approach.

As we discussed in Chapter 3 and Chapter 4, the research paper recommendation approaches can be evaluated using user study, online evaluation or offline evaluation (Beel et al, 2013). User studies have been useful way of evaluating paper recommendation systems [44], [25]. Despite being useful, user studies have certain limitations as well. Conducting a user study for a large dataset is not feasible since it requires many experts who are willing to evaluate such a large dataset. Since the dataset-2 had almost 5000 research papers, conducting a user study was not the preferred method of evaluation for this dataset. Therefore, we decided to use the automatic method of evaluation i.e. Jenson Shannon Divergence (JSD). As we have discussed in previous chapters, JSD finds the distance or divergence between two probability distributions. In the case of research papers, the two probability distributions were: (1) the contents of an individual paper (2) the contents of the cluster of papers it belongs to.

JSD produced the rankings for the bibliographically coupled papers automatically. Then we used the Spearman's correlation coefficient to determine the correlation between the results of our approach and those produced by the JSD. We also compared the results of our approach with those of the traditional bibliographic coupling and the content similarity.

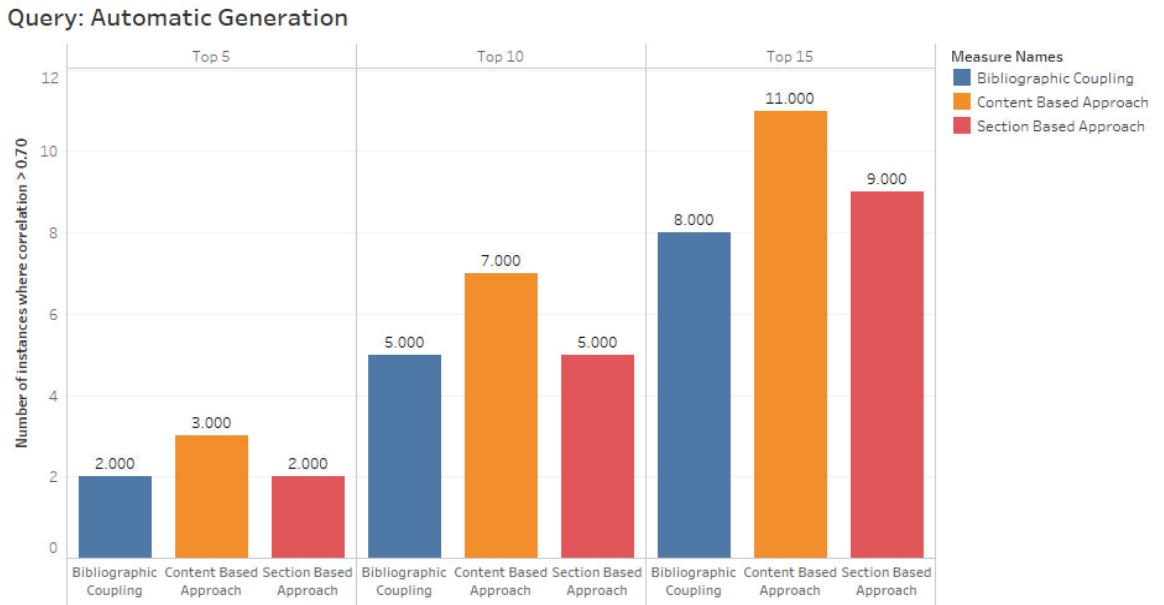


FIGURE 6.3: Comparative Evaluation of Sections based approach for query 'Automatic Generation'

The following paragraphs and figures describe the results and analysis of our proposed approach and the two other approaches.

Fig. 6.3 shows the comparison of our proposed approach with the bibliographic coupling and the content based approach in the case of top 5, top 10 and top 15 rankings. The X-axis represents the top 5, top 10 and top 15 categories. The Y-axis shows the number of instances for which the value of Spearman's correlation coefficient was higher than 0.70. As can be seen from the Fig. 6.3, content based approach performed better than the bibliographic coupling and our proposed approach in case of top 5, top 10 and top 15 rankings. The number of strongly correlated instances was the same for our approach and the bibliographic coupling for top 5 and top 10 rankings. Our proposed approach performed better than the bibliographic coupling in the case of top 15 rankings.

We performed the similar analysis for all the remaining 16 queries too. Table 6.2 shows the top 5 rankings for the number of instances where correlation was higher than 0.7 for all the 17 queries.

TABLE 6.2: Top 5 Rankings Comparison

Query	Section Based Approach	Content Based	Bibliographic Coupling
Automatic generation	2	3	2
Bayesian networks	1	3	1
Black box testing	3	3	3
Collaborative recommendation	5	5	5
Content based filtering	3	4	2
Feature selection	5	5	5
Graph theory	4	5	5
Information retrieval	5	2	1
Online marketing	3	3	3
Opinion mining	1	3	1
Query processing	4	5	4
Recommendation system	5	5	5
Regression testing	1	3	1
Sensor networks	4	4	4
Social Network	1	3	2
Subjectivity analysis	3	4	3
Wireless communications	4	4	4

Table 6.3 shows the top 10 rankings for the number of instances where correlation was higher than 0.7 for all the 17 queries.

TABLE 6.3: Top 10 Rankings Comparison

Query	Section Based Approach	Content Based	Bibliographic Coupling
Automatic generation	5	7	5
Bayesian networks	3	7	3
Black box testing	6	4	3
Collaborative recommendation	9	9	9
Content based filtering	5	5	2
Feature selection	10	10	10
Graph theory	8	10	9
Information retrieval	10	5	3
Online marketing	8	8	8
Opinion mining	5	7	5
Query processing	7	9	7
Recommendation system	7	9	7
Regression testing	3	7	3
Sensor networks	9	7	7
Social Network	2	4	4
Subjectivity analysis	5	7	5
Wireless communications	8	9	8

Table 6.4 shows the top 15 rankings for the number of instances where correlation was higher than 0.7 for all the 17 queries.

TABLE 6.4: Top 15 Rankings Comparison

Query	Section Based Approach	Content Based	Bibliographic Coupling
Automatic generation	9	11	8
Bayesian networks	7	11	7
Black box testing	6	4	3
Collaborative recommendation	12	13	11
Content based filtering	5	5	2
Feature selection	15	14	15
Graph theory	12	14	14
Information retrieval	13	6	4
Online marketing	12	12	10
Opinion mining	10	12	10
Query processing	12	14	12
Recommendation system	12	12	10
Regression testing	8	12	8
Sensor networks	11	8	7
Social Network	5	4	7
Subjectivity analysis	6	10	6
Wireless communications	12	12	9

The sections based approach performed better than the bibliographic coupling and content based approaches in most cases. However there were some cases where the bibliographic coupling and content based approaches performed the same or

better than our proposed approach. In these cases, most of the in-text citations pairs between bibliographically coupled papers were from the cross sections. In some of such cases, the in-text citation papers were mostly from the Introduction and Related Work sections.

6.4 Summary

In this approach, the logical structure of research papers exploited in order to improve the accuracy research paper recommendations. The initial experiments suggested that the sections Methodology, Results, Introduction and Related Work should have weights 3, 3, 2 and 1 respectively. These weights were then used for the dataset-2 and detailed evaluation was performed as discussed in previous section.

Fig. 6.4 shows a summary of comparison of our proposed sections based approach with the content based approach and the bibliographic coupling approach in case of top 5, top 10 and top 15 rankings. The X-axis represents the three categories i.e. Top 5, Top 10 and Top 15. The Y-axis represents the number of queries for which each approach outperformed the remaining two approaches. The three approaches are represented by the circles of different colors.

Our proposed approach performed better than the traditional bibliographic coupling approach in all three cases i.e. top 5, top 10 and top 15 rankings. Out of a total of 17 queries, our proposed approach produced better results than the bibliographic coupling in 1 query in case of top 5 rankings, and in 3 queries in case of top 10 and top 15 rankings. The figure shows that the content based approach produced better results than our proposed approach in case of top 5, top 10 and top 15 queries. In order to further investigate the results, we determined the average of correlation for all the queries.

Fig. 6.5 shows the number of queries for which the three approaches shared the top position with one or both approaches. We can see from this figure that the

Number of queries for which each approach performed the best.



FIGURE 6.4: Total no. of queries for which each approach performed better than others in Top 5, Top 10 and Top 15 rankings.

proposed approach was on the first spot for 7, 4 and 5 queries in case of top 5, top 10 and top 15 rankings.

Fig. 6.6 represents the average of the correlations for all the queries. In this diagram we did not select any top 'X' figure. Instead all of the papers were considered and complete ranking was compared between the proposed and state-of-the-art approaches. The X-axis represents the three approaches and the Y-axis represents the average of correlations for all the queries. As we can see from this figure, our proposed approach has an average correlation of 0.77 with the results of JSD. The average correlation of our proposed approach is higher than the content based approach and the bibliographic coupling approach. The average increase in accuracy for our approach is 8.5% and 2.7% as compared to bibliographic coupling and content based approaches respectively.

No. of queries for which approaches shared top position

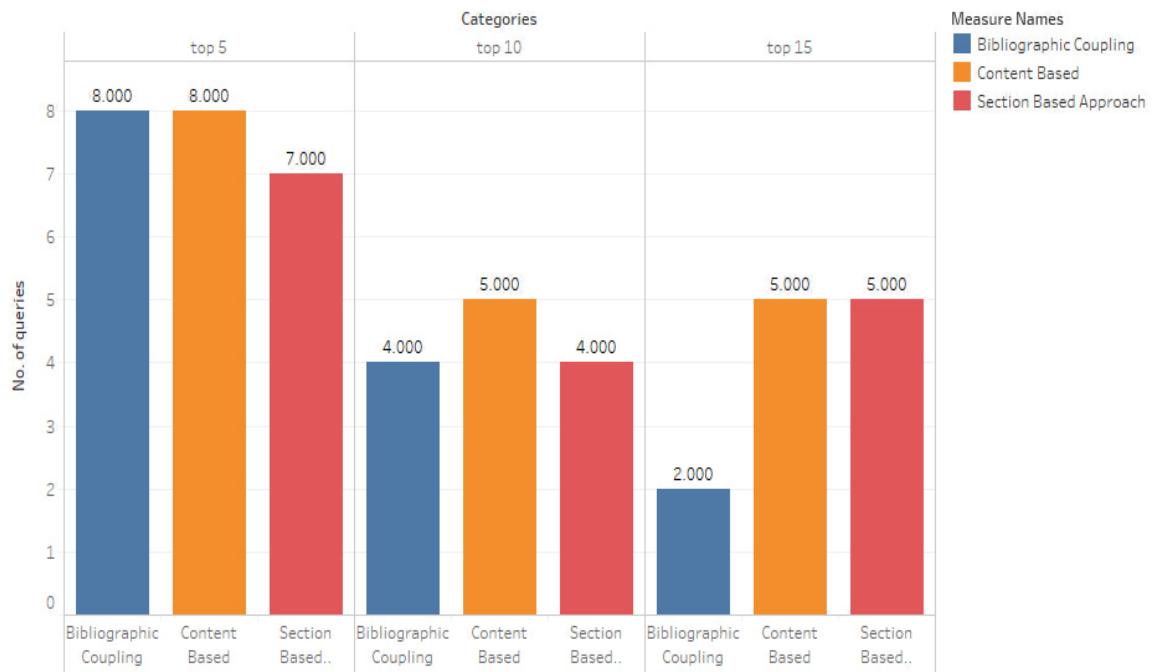


FIGURE 6.5: Total no. of queries for which the three approaches shared the top position with one or both approaches.

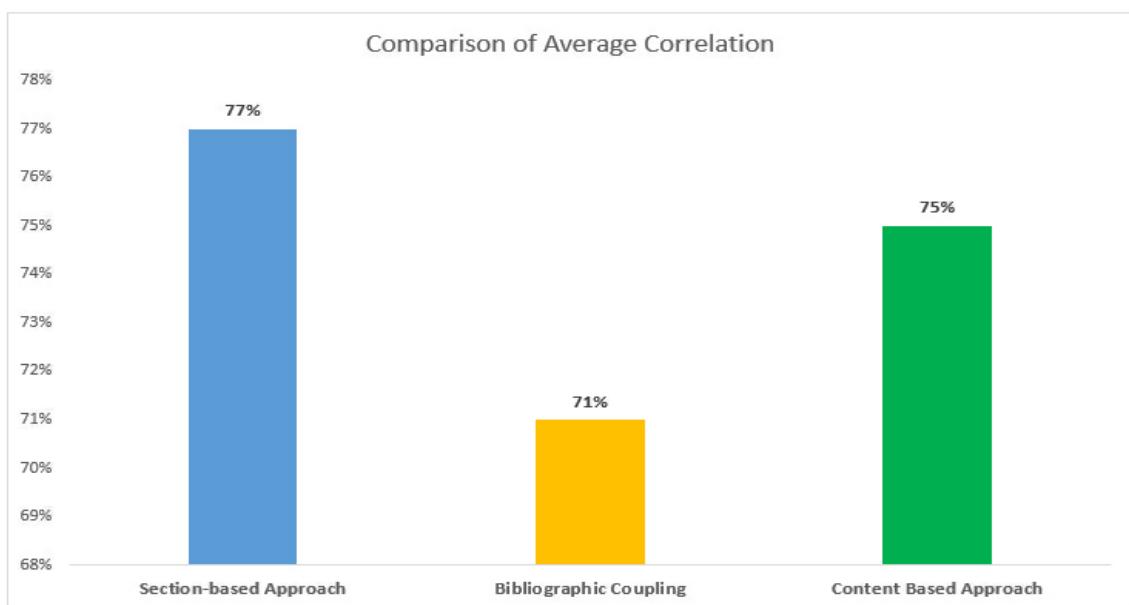


FIGURE 6.6: Average correlations of all queries

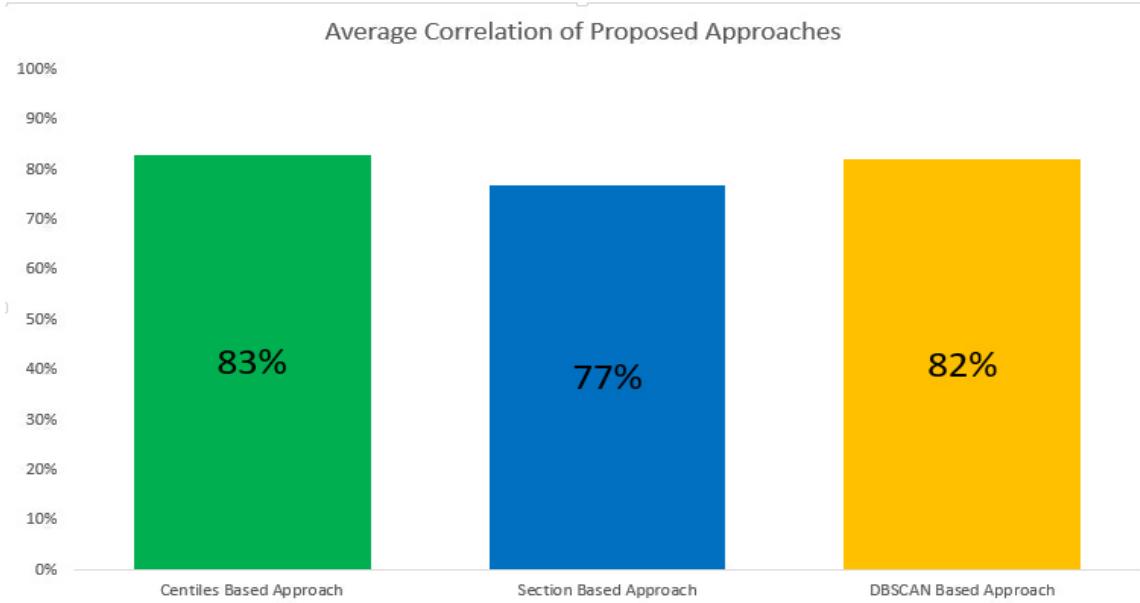


FIGURE 6.7: Comparison of Proposed Approaches

Fig. 6.7 shows the comparison of average correlation of our proposed approaches with each other. The centiles based approach performed better than the DBSCAN based approach and the sections based approach. The centiles based approach produced more accurate recommendation as compared to the other two approaches.

Hence the following hypothesis discussed in the chapter 1 is proved: **The accuracy of research paper recommender systems based on Bibliographic coupling can be improved by exploiting the in-text citation occurrences and their proximities between the bibliographically coupled papers.**

Chapter 7

Conclusion and Future Work

7.1 Conclusion

As a result of overwhelming advances in information technology, users face an arduous task when they are trying to access relevant information. Researchers today face the strenuous job of gaining access to the relevant research papers, due to information overload and the over-abundance of publications in conferences and journals. Research paper recommender systems have emerged as a revolutionary concept to help researchers get through this difficult situation.

Over the last few decades, many researchers have shown interest in proposing and developing innovative paper recommendation systems. Several techniques for paper recommender systems have been proposed and these can be placed into different categories based on the similarity measure they use. In this thesis, more than 150 state-of-the-art papers were reviewed and the available approaches were classified into the following categories: (1) metadata-based approaches, (2) citation-based approaches,(3) content-based approaches, (4) collaborative filtering-based approaches, (5) User Profile-based approaches, (6) Data Mining-based Approaches, and (7) Hybrid approaches. These approaches were critically analyzed and their limitations and drawbacks were discussed in the Chapter 2 of this thesis.

Research motivation was identified through this comprehensive literature review. Citation analysis-based approaches are very important. Researchers have extended the co-citation approach to include content analysis and citation proximity analysis and this had led to improvement in the accuracy of recommendations. Co-citation presents a relationship between two papers based on their co-occurrences in other papers, without considering the contents of the cited papers. However, bibliographic coupling considers two papers as relevant if they share common references. Therefore, bibliographic coupling has inherited the benefits of recommending relevant papers; however, traditional bibliographic coupling does not consider the citing patterns of common references in different logical parts of the citing papers. A recent study identified that more than 10 percent of references were never cited in the full text of the papers and were just part of the reference section of the papers. This limitation also motivates us to analyze the full-text of the papers.

In order to bridge these gaps, we proposed the following three new approaches for paper recommendation by making use of content and citation proximity analysis in bibliographic coupling: (1) DBSCAN-based bibliographic coupling, (2) CPA-based bibliographic coupling and (3) sections-based bibliographic coupling.

In order to conduct comprehensive experiments, two different datasets were collected using a focused web crawler. The first dataset called the dataset-1 contained 320 bibliographically coupled papers and the second dataset called the dataset-2 contained 5,000 bibliographically coupled papers from different domains. As discussed in the Chapter 3, 17 different queries were used to collect a comprehensive and diversified dataset. We used these two datasets to evaluate our proposed approaches. We used the dataset-1 in case of evaluation through user study. However, for the automatic evaluation using the JSD, we used the dataset-2.

Our first approach used DBSCAN clustering to analyze the impact of using proximity analysis of in-text citations to recommend research papers. In this approach, we first extracted all the in-text citations and their proximities from the bibliographically coupled papers. In next step we performed an extensive experiment on

dataset-1 to determine the value of ε that produced the most accurate recommendations for the DBSCAN clustering of in-text citations. Experiments showed that the best value of ε was 150. Later on, this value of ε was used on our dataset-2.

The evaluation showed significant improvement in the accuracy of paper recommendation. As discussed in Chapter 4, our proposed approach won in 9, 10 and 11 out of 17 queries in case of top 5, top 10 and top 15 rankings. The other two approaches won in one query each in case of top 10 rankings only. In case of top 15 rankings, the bibliographic coupling won in one query only. In all the remaining queries, our proposed approach shared the first position with any of the approach that won. When considering all the queries, the percentage increase in accuracy of our proposed approach was 16% and 10% as compared to the bibliographic coupling and content based approach respectively.

The DBSCAN based approach produced better results in the cases where the citations had less variance in density. The results of DBSCAN based approach varied based on the values of the ε and $minPts$. We optimized the value of the Epsilon using a training dataset. The optimized value of ε produced relatively accurate paper recommendations. This approach doesn't need the information of sections to recommend papers.

Our second approach used citation proximity analysis (CPA) in bibliographic coupling. CPA has been used by researchers in the past for the co-citation analysis and led to improvement in paper recommendation. Therefore, we decided to use it in bibliographic coupling. In this approach, we evaluated three existing approaches in CPA and proposed two new schemes for assigning weights to the in-text citation pairs based on their centiles.

In this approach, first we found the positions of the in-text citations. In the next step, we calculated the centile location of each in-text citation. In the next step, the distance between the centile values of all the in-text citations pairs were calculated. These values were stored in the database and were used by five different citation proximity schemes that cluster the citation percentile values using different thresholds. As explained in Chapter 5, we used two weighting schemes (S1, S2)

which were proposed by Boyack et al [1]. And we used the dataset-1 in order to propose and evaluate 3 new weighting schemes (S3, S4 and S5) for our proposed approach. The results showed that scheme S5 performed better than the remaining 5 schemes and had an average correlation of 0.93. Since S5 performed better than the remaining 5 schemes, we preferred to use this scheme on the dataset-2.

We performed comprehensive automatic evaluation using the JSD for this approach, which showed that our proposed approach performs better than the content-based approach and the bibliographic-coupling approach. Out of 17 queries in total, the proposed approach produced better results than the other two approaches in 10 out of 17 queries in the case of top 5 rankings, in 14 out of 17 queries in the case of top 10 rankings and 16 out of 17 queries in the case of top 15 rankings. When considering all the queries, the percentage increase in accuracy of the proposed approach was 17 percent and 11 percent as compared to the bibliographic-coupling and content-based approaches respectively.

The centiles based approach uses the centile values of the in-text citations and recommends those papers whose centile distance from each other is less compared to others. This approach does not need the logical sections of the research papers to recommend papers. Unlike the DBSCAN based approach, this approach does not need to cluster the in-text citations in order to rank the bibliographically coupled papers.

Our third approach was section-based bibliographic coupling for paper recommendation. This approach arose from an intuitive sense that authors follow certain standards when they distribute the in-text citations in their papers and that in-text citation from certain sections carries more weight than the others. The initial experiments suggested that the sections Methodology, Results, Introduction and Related Work should have weights 3, 3, 2 and 1 respectively. Automatic evaluation suggested that the proposed approach performed better than the traditional bibliographic-coupling approach in all three cases i.e. top 5, top 10 and top 15 rankings. Out of a total of 17 queries, our proposed approach produced better results than the bibliographic coupling in 1 query in the case of top 5 rankings,

and in 3 queries in the case of top 10 and top 15 rankings. The figure shows that the content-based approach produced better results than our proposed approach in the case of top 5, top 10 and top 15 queries. When considering all the queries, the percentage increase in accuracy of the proposed approach was 8 percent and 3 percent as compared to the bibliographic-coupling and content-based approaches respectively.

Unlike the centiles based approach and the DBSCAN based approach, the sections based approach doesn't need to perform the normalization. It doesn't need to convert the proximities of in-text citations into centile values. The sections based approach doesn't need to perform the Min-Max normalization either. In this approach, we need to focus only the distribution of in-text citations in different logical sections of the paper.

The results for all approaches in the case of top 5, top 10 and top 15 rankings were compared and the comparison showed that using the proximity analysis for bibliographic coupling improves the accuracy of paper recommendation as compared to the traditional bibliographic coupling.

7.2 Future Work

Our comprehensive evaluation showed that using the proximity and positions of in-text citations for bibliographic coupling produces better paper recommendations as compared to the traditional bibliographic coupling approach. There are some hybrid approaches that are using the traditional bibliographic coupling. Those approaches need to be revisited now by the scientific community by replacing the standard approach of bibliographic coupling with the proposed bibliographic coupling approaches.

Another area that needs to be improved in future is the weight tuning for sections. An automatic way of assigning weights to different sections may improve the results. Neural networks can be used to assign weights automatically to different sections.

A hybrid approach needs to be proposed and developed, that could use all of the proposed approaches from this thesis and could utilize the best cases from each approach in different scenarios. Different features such as centile positions, sections and metadata can be analyzed to determine the most important features that could further improve paper recommendations.

Bibliography

- [1] K. W. Boyack, H. Small, and R. Klavans, “Improving the accuracy of co-citation clustering using full text,” *Journal of the Association for Information Science and Technology*, vol. 64, no. 9, pp. 1759–1767, 2013.
- [2] M. Khabsa and C. L. Giles, “The number of scholarly documents on the public web,” *PloS one*, vol. 9, no. 5, p. e93949, 2014.
- [3] M. Hristakeva, D. Kershaw, M. Rossetti, P. Knoth, B. Pettit, S. Vargas, and K. Jack, “Building recommender systems for scholarly information,” in *Proceedings of the 1st Workshop on Scholarly Web Mining*. ACM, 2017, pp. 25–32.
- [4] M. Amami, G. Pasi, F. Stella, and R. Faiz, “An lda-based approach to scientific paper recommendation,” in *International Conference on Applications of Natural Language to Information Systems*. Springer, 2016, pp. 200–210.
- [5] N. Ratprasartporn and G. Ozsoyoglu, “Finding related papers in literature digital libraries,” *Research and Advanced Technology for Digital Libraries*, pp. 271–284, 2007.
- [6] X.-Y. Liu and B.-C. Chien, “Applying citation network analysis on recommendation of research paper collection,” in *Proceedings of the 4th Multidisciplinary International Social Networks Conference on ZZZ*. ACM, 2017, p. 30.
- [7] H. Sahijwani and S. Dasgupta, “User profile based research paper recommendation,” *arXiv preprint arXiv:1704.07757*, 2017.

- [8] J. Beel, B. Gipp, S. Langer, and C. Breitinger, “paper recommender systems: a literature survey,” *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, 2016.
- [9] B. Gipp and J. Beel, “Citation proximity analysis (cpa): a new approach for identifying related work based on co-citation analysis,” in *ISSI09: 12th International Conference on Scientometrics and Informetrics*, 2009, pp. 571–575.
- [10] T. Kuhn, M. Perc, and D. Helbing, “Inheritance patterns in citation networks reveal scientific memes,” *Physical Review X*, vol. 4, no. 4, p. 041036, 2014.
- [11] M. Perc, “Zipfs law and log-normal distributions in measures of scientific output across fields and institutions: 40 years of slovenias research as an example,” *Journal of Informetrics*, vol. 4, no. 3, pp. 358–364, 2010.
- [12] H. Small, “Co-citation in the scientific literature: A new measure of the relationship between two documents,” *Journal of the Association for Information Science and Technology*, vol. 24, no. 4, pp. 265–269, 1973.
- [13] M. M. Kessler, “Bibliographic coupling between scientific papers,” *Journal of the Association for Information Science and Technology*, vol. 14, no. 1, pp. 10–25, 1963.
- [14] A. Shahid, M. Afzal, and M. Qadir, “Discovering semantic relatedness between scientific articles through citation frequency,” in *Workshop on Text mining for Scholarly Communications and Repositories, Australian Journal of Basic Applied Sciences*, vol. 5, 2011, pp. 1599–1604.
- [15] R. Kumar, “Research methodology: A step-by-step guide for beginners. london: Sage publication ltd,” 2011.
- [16] A. E. Jinha, “Article 50 million: an estimate of the number of scholarly articles in existence,” *Learned Publishing*, vol. 23, no. 3, pp. 258–263, 2010.

- [17] M. Biba and E. Gjati, “Boosting text classification through stemming of composite words,” in *Recent Advances in Intelligent Informatics*. Springer, 2014, pp. 185–194.
- [18] M. T. Afzal, *Context aware information discovery for scholarly e-community*. na, 2010.
- [19] S. Doerfel, R. Jäschke, A. Hotho, and G. Stumme, “Leveraging publication metadata and social data into folkrank for scientific publication recommendation,” in *Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web*. ACM, 2012, pp. 9–16.
- [20] M. T. Afzal, N. Kulathuramaiyer, and H. A. Maurer, “Creating links into the future.” *J. UCS*, vol. 13, no. 9, pp. 1234–1245, 2007.
- [21] E. Garfield *et al.*, “Citation analysis as a tool in journal evaluation.” American Association for the Advancement of Science, 1972.
- [22] Y. Ding, G. Zhang, T. Chambers, M. Song, X. Wang, and C. Zhai, “Content-based citation analysis: The next generation of citation analysis,” *Journal of the Association for Information Science and Technology*, vol. 65, no. 9, pp. 1820–1833, 2014.
- [23] S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl, “On the recommending of citations for research papers,” in *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. ACM, 2002, pp. 116–125.
- [24] J. Lee, K. Lee, and J. G. Kim, “Personalized academic research paper recommendation system,” *arXiv preprint arXiv:1304.5457*, 2013.
- [25] K. Sugiyama and M.-Y. Kan, “Exploiting potential citation papers in scholarly paper recommendation,” in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2013, pp. 153–162.

- [26] B. Kaya, “User profile based paper recommendation system,” *International Journal of Intelligent Systems and Applications in Engineering*, vol. 6, no. 2, pp. 151–157, 2018.
- [27] H. A. M. Hassan, “Personalized research paper recommendation using deep learning,” in *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. ACM, 2017, pp. 327–330.
- [28] K. Haruna and M. A. Ismail, “An ontological framework for research paper recommendation,” *International Journal of Soft Computing*, vol. 11, no. 2, pp. 96–99, 2016.
- [29] M. Amami, R. Faiz, F. Stella, and G. Pasi, “A graph based approach to scientific paper recommendation,” in *Proceedings of the International Conference on Web Intelligence*. ACM, 2017, pp. 777–782.
- [30] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, “Using collaborative filtering to weave an information tapestry,” *Communications of the ACM*, vol. 35, no. 12, pp. 61–70, 1992.
- [31] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, “Eigentaste: A constant time collaborative filtering algorithm,” *Information Retrieval*, vol. 4, no. 2, pp. 133–151, 2001.
- [32] P. Hongyan, L. Hongfei, and Z. Jing, “Collaborative filtering algorithm based on matrix partition and interest variance [j],” *Journal of the China Society for Scientific and Technical Information*, vol. 1, p. 008, 2006.
- [33] K. D. Bollacker, S. Lawrence, and C. L. Giles, “Discovering relevant scientific literature on the web,” *IEEE Intelligent Systems and their Applications*, vol. 15, no. 2, pp. 42–47, 2000.
- [34] S. Bethard and D. Jurafsky, “Who should i cite: learning literature search models from citation behavior,” in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 609–618.

- [35] I. Nassiri, A. Masoudi-Nejad, M. Jalili, and A. Moeini, “Normalized similarity index: An adjusted index to prioritize article citations,” *Journal of Informetrics*, vol. 7, no. 1, pp. 91–98, 2013.
- [36] M. Krapivin and M. Marchese, “Focused page rank in scientific papers ranking.” in *ICADL*. Springer, 2008, pp. 144–153.
- [37] M. Gori and A. Pucci, “Research paper recommender systems: A random-walk based approach,” in *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*. IEEE, 2006, pp. 778–781.
- [38] K. El-Arini and C. Guestrin, “Beyond keyword search: discovering relevant scientific literature,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 439–447.
- [39] S. Singla, N. Duhan, and U. Kalkal, “A novel approach for document ranking in digital libraries using extractive summarization,” *International Journal of Computer Applications*, vol. 74, no. 18, 2013.
- [40] T. Strohman, W. B. Croft, and D. Jensen, “Recommending citations for academic papers,” in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 705–706.
- [41] M. Reyhani, S.-C. Lee, S.-W. Kim, and D.-J. Kim, “On exploiting content and citations together to compute similarity of scientific papers,” 2013.
- [42] J. Bichteler and E. A. Eaton, “The combined use of bibliographic coupling and cocitation for document retrieval,” *Journal of the Association for Information Science and Technology*, vol. 31, no. 4, pp. 278–282, 1980.
- [43] 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*. ACM, 2011, pp. 297–306.

- [44] J. Beel, S. Langer, M. Genzmehr, and A. Nürnberg, “Introducing docear’s research paper recommender system,” in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries.* ACM, 2013, pp. 459–460.
- [45] A. Daud, A. M. A. R. Shaikh, and A. H. Rajpar, “Scientific reference mining using semantic information through topic modeling,” *Research Journal of Engineering & Technology*, vol. 28, no. 2, pp. 253–262, 2009.
- [46] F. Ferrara, N. Pudota, and C. Tasso, “A keyphrase-based paper recommender system.” in *IRCDL.* Springer, 2011, pp. 14–25.
- [47] S. Pruitikanee, L. Di Jorio, A. Laurent, and M. Sala, “Paper recommendation system: A global and soft approach,” in *FUTURE COMPUTING’2012: Fourth International Conference on Future Computational Technologies and Applications*, 2012, p. 7.
- [48] G. Linden, B. Smith, and J. York, “Amazon. com recommendations: Item-to-item collaborative filtering,” *IEEE Internet computing*, vol. 7, no. 1, pp. 76–80, 2003.
- [49] N. Agarwal, E. Haque, H. Liu, and L. Parsons, “A subspace clustering framework for research group collaboration,” 2009.
- [50] T. Bogers and A. Van den Bosch, “Recommending scientific articles using citeulike,” in *Proceedings of the 2008 ACM conference on Recommender systems.* ACM, 2008, pp. 287–290.
- [51] S. Pohl, F. Radlinski, and T. Joachims, “Recommending related papers based on digital library access records,” in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries.* ACM, 2007, pp. 417–418.
- [52] M. Baez, D. Mirylenka, and C. Parra, “Understanding and supporting search for scholarly knowledge,” *Proceeding of the 7th European Computer Science Summit*, pp. 1–8, 2011.

- [53] K. Chandrasekaran, S. Gauch, P. Lakkaraju, and H. Luong, “Concept-based document recommendations for citeseer authors,” in *Adaptive hypermedia and adaptive web-based systems*. Springer, 2008, pp. 83–92.
- [54] W. Choochaiwattana, “Usage of tagging for research paper recommendation,” in *Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on*, vol. 2. IEEE, 2010, pp. V2–439.
- [55] R. Dong, L. Tokarchuk, and A. Ma, “Digging friendship: paper recommendation in social network,” in *Proceedings of Networking & Electronic Commerce Research Conference (NAEC 2009)*, 2009, pp. 21–28.
- [56] G. Geisler, D. McArthur, and S. Giersch, “Developing recommendation services for a digital library with uncertain and changing data,” in *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2001, pp. 199–200.
- [57] A. Kodakateri Pudhiyaveetil, S. Gauch, H. Luong, and J. Eno, “Conceptual recommender system for citeseerx,” in *Proceedings of the third ACM conference on Recommender systems*. ACM, 2009, pp. 241–244.
- [58] T. Theeramunkong and K. Sriphaew, “Discovery of relations among scientific articles using association rule mining,” in *Proceedings of the 2007 NSTDA Annual Conference Science (Science and Technology for National Productivity and Happiness), Thailand Science Park, Pathumthani, Thailand*, 2007.
- [59] S. C. Cazella and L. O. C. Alvares, “Combining data mining technique and users relevance opinion to build an efficient recommender system,” *Revista Tecnologia da Informação, UCB*, vol. 4, no. 2, 2005.
- [60] J. Beel, S. Langer, and M. Genzmehr, “Sponsored vs. organic (research paper) recommendations and the impact of labeling,” in *International Conference on Theory and Practice of Digital Libraries*. Springer, 2013, pp. 391–395.

- [61] J. Beel, S. Langer, M. Genzmehr, and A. Nürnberg, “Persistence in recommender systems: giving the same recommendations to the same users multiple times,” in *International Conference on Theory and Practice of Digital Libraries*. Springer, 2013, pp. 386–390.
- [62] J. Beel, S. Langer, A. Nürnberg, and M. Genzmehr, “The impact of demographics (age and gender) and other user-characteristics on evaluating recommender systems,” in *International Conference on Theory and Practice of Digital Libraries*. Springer, 2013, pp. 396–400.
- [63] J. Bollen and L. Rocha, “An adaptive systems approach to the implementation and evaluation of digital library recommendation systems,” *Research and Advanced Technology for Digital Libraries*, pp. 356–359, 2000.
- [64] R. Torres, S. M. McNee, M. Abel, J. A. Konstan, and J. Riedl, “Enhancing digital libraries with techlens,” in *Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on*. IEEE, 2004, pp. 228–236.
- [65] C. Wang and D. M. Blei, “Collaborative topic modeling for recommending scientific articles,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 448–456.
- [66] M. D. Ekstrand, P. Kannan, J. A. Stemper, J. T. Butler, J. A. Konstan, and J. T. Riedl, “Automatically building research reading lists,” in *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 2010, pp. 159–166.
- [67] Y. Wang, H. Zhang, Y. Li, D. Wang, Y. Ma, T. Zhou, and J. Lu, “A data cleaning method for citeseer dataset,” in *International Conference on Web Information Systems Engineering*. Springer, 2016, pp. 35–49.
- [68] A. Constantin, S. Pettifer, and A. Voronkov, “Pdflx: fully-automated pdf-to-xml conversion of scientific literature,” in *Proceedings of the 2013 ACM symposium on Document engineering*. ACM, 2013, pp. 177–180.

- [69] B. Cronin, *The citation process: The role and significance of citations in scientific communication.* T. Graham London, 1984.
- [70] Y. Ding, X. Liu, C. Guo, and B. Cronin, “The distribution of references across texts: Some implications for citation analysis,” *Journal of Informetrics*, vol. 7, no. 3, pp. 583–592, 2013.
- [71] B. Golshan, T. Lappas, and E. Terzi, “Sofia search: a tool for automating related-work search,” in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data.* ACM, 2012, pp. 621–624.
- [72] T. Hengl and M. Gould, “Rules of thumb for writing research articles,” *Enschede, September*, 2002.
- [73] S. Teufel, “Citations and sentiment,” in *Workshop on Text mining for Scholarly Communications and Repositories, University of Manchester, UK*, 2009.
- [74] R. Habib and M. T. Afzal, “Paper recommendation using citation proximity in bibliographic coupling,” *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 25, no. 4, pp. 2708–2718, 2017.
- [75] G. Holmes, A. Donkin, and I. H. Witten, “Weka: A machine learning workbench,” in *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on.* IEEE, 1994, pp. 357–361.
- [76] C. Spearman, “The proof and measurement of association between two things,” *The American journal of psychology*, vol. 15, no. 1, pp. 72–101, 1904.
- [77] P. Cremonesi, F. Garzotto, and R. Turrin, “Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study,” *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 2, p. 11, 2012.

- [78] P. Cremonesi, F. Garzotto, S. Negro, A. V. Papadopoulos, and R. Turrin, “Looking for good recommendations: A comparative evaluation of recommender systems,” in *IFIP Conference on Human-Computer Interaction*. Springer, 2011, pp. 152–168.
- [79] M. M. Mukaka, “A guide to appropriate use of correlation coefficient in medical research,” *Malawi Medical Journal*, vol. 24, no. 3, pp. 69–71, 2012.
- [80] H. G. Small, “Structural dynamics of scientific literature,” *International Classification*, vol. 3, no. 2, pp. 67–74, 1976.
- [81] S. Liu and C. Chen, “The differences between latent topics in abstracts and citation contexts of citing papers,” *Journal of the American Society for Information Science and Technology*, vol. 64, no. 3, pp. 627–639, 2013.
- [82] M. Bertin, I. Atanassova, V. Lariviere, and Y. Gingras, “The distribution of references in scientific papers: An analysis of the imrad structure,” in *Proceedings of the 14th ISSI Conference*, vol. 591, 2013, p. 603.
- [83] H. Voos and K. S. Dagaev, “Are all citations equal? or, did we op. cit. your idem?.” *Journal of Academic Librarianship*, vol. 1, no. 6, pp. 19–21, 1976.
- [84] K. McCain and K. Turner, “Citation context analysis and aging patterns of journal articles in molecular genetics,” *Scientometrics*, vol. 17, no. 1-2, pp. 127–163, 1989.
- [85] S. Maričić, J. Spaventi, L. Pavrić, and G. Pifat-Mrzljak, “Citation context versus the frequency counts of citation histories,” *Journal of the Association for Information Science and Technology*, vol. 49, no. 6, pp. 530–540, 1998.
- [86] J. Hauke and T. Kossowski, “Comparison of values of pearson’s and spearman’s correlation coefficients on the same sets of data,” *Quaestiones geographicae*, vol. 30, no. 2, p. 87, 2011.