

CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD



**A Novel Content-based Filtering
Approach to Identify Important
Citations**

by

Muhammad Saboor Ahmed

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

in the

Faculty of Computing

Department of Computer Science

2022

A Novel Content-based Filtering Approach to Identify Important Citations

By

Muhammad Saboor Ahmed

(PC133007)

Dr. Hermann Maurer, Professor
Technical University Graz, Austria
(Foreign Evaluator 1)

Dr. Joel Rodrigues, Professor
University of Beira Covilha, Portugal
(Foreign Evaluator 2)

Dr. Abdul Basit Siddiqui
(Thesis Supervisor)

Dr. Abdul Basit Siddiqui
(Head, Department Computer Science)

Dr. Muhammad Abdul Qadir
(Dean, Faculty of Computing)

DEPARTMENT OF COMPUTER SCIENCE
CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
ISLAMABAD

2022

Copyright © 2022 by Muhammad Saboor Ahmed

All rights are reserved. No Part of the material protected by this copy right notice may be reproduced or utilized in any form or any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without the permission from the author.

DEDICATED

To

My Mother

*A strong and gentle soul who taught me to trust in Allah, believe in hard work
and that so much could be done with little.*



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 “**A Novel Content-based Filtering Approach to Identify Important Citations**” was conducted under the supervision of **Dr. Abdul Basit Siddiqui**. 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 **September 23, 2022**.

Student Name : Muhammad Saboor Ahmed
(PC133007)

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

Examination Committee :

- (a) External Examiner 1: Dr. Ayyaz Hussain
Professor
QAU, Islamabad
- (b) External Examiner 2: Dr. Waseem Shahzad
Professor
FAST-NUCES, Islamabad
- (c) Internal Examiner : Dr. Azhar Mahmood
Associate Professor
CUST, Islamabad

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

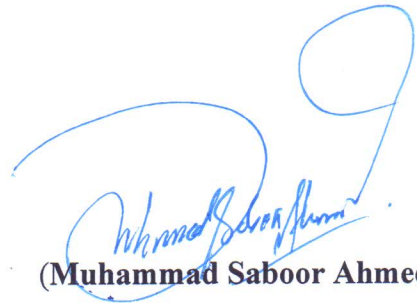
Name of HoD : Dr. Abdul Basit Siddiqui
Associate Professor
CUST, Islamabad

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

AUTHOR'S DECLARATION

I, **Muhammad Saboor Ahmed (Registration No. PC-133007)**, hereby state that my PhD thesis titled, '**A Novel Content-based Filtering Approach to Identify Important Citations**' 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.



(**Muhammad Saboor Ahmed**)

Dated: September, 2022

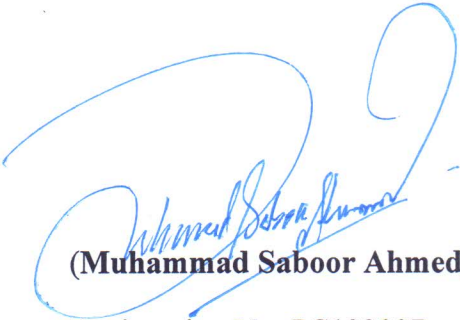
Registration No: PC133007

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled “**A Novel Content-based Filtering Approach to Identify Important Citations**” 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.



(Muhammad Saboor Ahmed)

Dated: September, 2022

Registration No: PC133007

List of Publications

It is certified that the following publication has been made out of the research work that has been carried out for this thesis:

1. M. S. Ahmed and M. T. Afzal, "Exploiting structural similarities to classify citations," *Computers, Materials & Continua*, vol. 66, no.2, pp. 11951214, 2021.

Muhammad Saboor Ahmed

(PC133007)

Acknowledgement

First and foremost, I would like to praise Allah the Almighty, the Most Gracious, and the Most Merciful for His blessing given to me during my study and in completing this thesis. May Allah's blessing goes to His final Prophet Muhammad (peace be up on him), his family and his companions.

Secondly, I would like to acknowledge and give my warmest thanks to my supervisor **Dr. Muhammad Tanvir Afzal** who made this work possible. His guidance and advice carried me through all the stages of writing my thesis. I would also like to thank to my new supervisor **Dr. Abdul Basit Siddiqui** who was appointed by the university after my previous supervisor left the university. I am thankful to him for his brilliant comments and suggestions for the completion of this thesis.

I am also thankful to my family for supporting me along my journey in life. I will forever owe my achievements to my dedicated, caring and thoughtful family. I would not be the person I am today without your help.

Muhammad Saboor Ahmed

(PC133007)

Abstract

Citations play an important role in the scientific community by assisting in measuring multifarious policies like the impact of journals, researchers, institutions, and countries. Authors cite papers for different reasons, such as extending previous work, comparing their study with the state-of-the-art, providing background of the field, etc. In recent years, researchers have tried to conceptualize all citations into two broad categories, important and non-important. Such a categorization is vital to enhance scientific output in multiple ways, for instance, (1) helping a researcher in identifying meaningful citations from a list of 100 to 1000 citations (2) enhancing the impact factor calculation mechanism by more strongly weighting important citations, and (3) improving researcher, institutional, and university rankings by only considering important citations. All of these uses depend upon correctly identifying the important citations from the list of all citations in a paper. To date, researchers have utilized many features to classify citations into these broad categories: cue phrases, in-text citation counts, and metadata features, etc. However, contemporary approaches are based on identification of in-text citation counts, mapping sections onto the Introduction, Methods, Results, and Discussion (IMRAD) structure, identifying cue phrases, etc. Accurate identification of such features is a challenging task and is normally conducted manually, with the accuracy of citation classification demonstrated in terms of these manually extracted features. This research proposes to examine the content of the cited and citing pair to identify important citing papers for each cited paper. This content similarity approach was adopted from research paper recommendation approaches. Furthermore, a novel section-based content similarity approach is also proposed. The proposed approach utilizes content of corresponding logical sections of citing and cited research articles such as Abstract, Introduction, Methodology, etc. The cosine similarity has been used to calculate the similarity scores of corresponding logical sections and different sections have been combined with average and weighted average formulas. The experiments have been performed on the comprehensive annotated dataset of Valenzuela. After comparing our results with

metadata-based and content-based contemporary approaches, the proposed approach outperformed all state-of-the-art approaches by achieving the F-Measure score of 0.75.

Contents

Author’s Declaration	v
Plagiarism Undertaking	vi
List of Publications	vii
Acknowledgement	viii
Abstract	ix
List of Figures	xiv
List of Tables	xvi
1 Introduction	1
1.1 Background Study	1
1.2 Preliminaries	5
1.2.1 Citation	5
1.2.2 Citation Analysis	5
1.2.3 Co-Citation-Analysis	6
1.2.4 Analysis of Co-citation-Proximity	7
1.2.5 Co-citation-Proximity-Analysis-Based on Byte-offset	8
1.2.6 In-text-Citation-Frequency-Analysis (ICFA)	8
1.2.7 Content Based Filtering Approach	9
1.3 Research Motivation	10
1.4 Research Questions	11
1.5 Research Problem	12
1.6 Research Scope	12
1.7 Applications of the Proposed Solution	13
1.8 Research Approach	13
1.9 Research Contribution	15
1.10 Thesis Outline	15
2 Literature Review	17
2.1 Relevant Research Paper Identification	17

2.1.1	Stereotyping	18
2.1.2	Co-occurrence Recommendation	20
2.1.3	Collaborative Filtering	21
2.1.4	Content Based Approaches	23
2.2	Citation Based Approaches	26
2.3	Citation Count	28
2.4	Citation Behavior	29
2.5	Quantitative Citations	29
2.6	Citation Classification	30
2.7	Automated Citation Classification	30
2.8	Hybrid Approaches	32
2.9	Critical Analysis	33
2.10	Conclusion	38
3	An Innovative Approach to Identify Important Citations	40
3.1	Benchmark Datasets	41
3.2	PDF to Text Conversion	43
3.3	Pre-Processing	44
3.3.1	Stop-Words Removal	45
3.3.2	Stemming	45
3.4	Content-Based Approach	45
3.5	Section-wise Content Similarity Approach	46
3.6	Section-Wise Research Paper Ranking	47
3.7	Techniques for Combining Similarity Scores of Logical Section	49
3.7.1	Average	51
3.7.2	Weighted Average	52
3.8	Evaluation Parameters	53
3.9	Conclusion	54
4	Results and Findings	55
4.1	Results and Evaluation of the Proposed Approaches	55
4.1.1	Evaluation of Proposed Six Rankings	56
4.1.1.1	Full Content Similarity-based Ranking	56
4.1.1.2	Abstract-section Similarity-based Ranking	57
4.1.1.3	Introduction-Section Similarity-Based Ranking	57
4.1.1.4	Literature Section Similarity-based Ranking	60
4.1.1.5	Methodology-Section Similarity-Based Ranking	63
4.1.1.6	Results-Section Similarity-Based Ranking	63
4.1.1.7	Conclusion of Six Similarity-Based Ranking	65
4.2	Combination of Double Parameters using Average Results	67
4.2.1	AVR(Abtract-Methodology-Sections Similarity-Based Ranking)	67
4.2.2	AVR(Abtract-Results-Sections Similarity-Based Ranking)	68

4.2.3	AVR(Methodology-Results-Sections Similarity-Based Ranking)	70
4.3	Combination of Double Parameters using Weighted Average Results	72
4.3.1	Wt.AVR(Abstract-Methodology-Sections Similarity-Based Ranking)	72
4.3.2	Wt.AVR(Abstract-Results-Sections Similarity-Based Ranking)	74
4.3.3	Wt.AVR(Methodology-Results-Sections Similarity-Based Ranking)	76
4.4	Double Ranking approaches Conclusion	77
4.5	Combination of Triple Parameters using Average Results	79
4.5.1	AVR(Abstract-Methodology-Results-Sections Similarity-Based Ranking)	79
4.6	Combination of Triple Parameters using Weighted Average Results	81
4.6.1	Wt.AVR(Abstract-Methodology-Results-Sections Similarity-Based Ranking)	81
4.7	Triple Ranking approaches Conclusion	82
4.8	Comparison to State-of-the-art Approaches	83
4.9	Comparisons of n-gram Features	87
5	Discussion	89
5.1	Conclusion	89
5.2	Limitations	93
5.3	Future Work	93
	Bibliography	94

List of Figures

1.1	Analysis of Citation between Cited and Citing Documents	6
1.2	Analysis of Citation between Cited and Citing Documents	7
1.3	Co-citation Proximity Analyses	8
1.4	Co-citation-Proximity-Analysis-Based on Byte-offset	9
1.5	In-text-Citation-Frequency-Analysis	9
1.6	Research Methodology	14
3.1	Proposed Methodology	42
3.2	Logical Sections in PDF File	43
3.3	Logical Section Abstract	44
3.4	Logical Section Introduction	44
4.1	Abstract vs. Full Content Top 3	58
4.2	Abstract vs. full Content Top 5	58
4.3	Introduction vs. Full Content Top 3	59
4.4	Introduction vs. Full Content Top 5	59
4.5	Literature vs. Full Content Top 3	61
4.6	Literature vs. Full Content Top 5	61
4.7	Methodology vs. Full Content Top 3	62
4.8	Methodology vs. Full Content Top 5	62
4.9	Results vs. Full Content Top 3	64
4.10	Results vs. Full Content Top 5	64
4.11	Comparison of the Six Proposed Rankings	66
4.12	Avg(Abstract+Methodology) Vs All Content Top 3	68
4.13	Avg (abstract + Methodology) Vs All Content Top 5	68
4.14	Avg(Abstract+Results) Vs All Content Top 3	69
4.15	Avg(Abstract+Results) Vs All Content Top 5	70
4.16	Avg(Methodology+Results) Vs All Content Top 3	71
4.17	Avg(Methodology+Results) Vs All Content Top 5	72
4.18	W.Avg(Abstract+Methodology) Vs All Content Top 3	73
4.19	W.Avg(Abstract+Methodology) Vs All Content Top 5	74
4.20	W.Avg(Abstract+Results) Vs All Content Top 3	75
4.21	W.Avg(Abstract+Results) Vs All Content Top 5	75
4.22	W.Avg(Methodology+Results) Vs All Content Top 3	77
4.23	W.Avg(Methodology+Results) Vs All Content Top 5	78
4.24	Avg(Abstract+Methodology+Results) Vs All Content top 3	80
4.25	Avg(Abstract+Methodology+Results) Vs All Content Top 5	80

4.26	W-Avg(Abstract+Methodology+Results) Vs All Content top 3 . . .	82
4.27	W.Avg(Abstract+Methodology+Results) Vs All Content	82
4.28	Comparison the precision of the proposed approaches with state-of-the-art rankings	83
4.29	Comparing the recall of the proposed approaches with state-of-the-art rankings	84

List of Tables

2.1	Critical Analysis of Different Approaches Containing Parameters, Contribution, and Claims	35
3.1	Benchmark Dataset	41
3.2	Combination of Two Sections by Using Average	51
3.3	Combination of Two Sections by Using Weighted Average	53
4.1	All Possible Combinations	66
4.2	Average Scores of Double Parameters Vs All Content.	78
4.3	Weighted Average Scores of Double Parameters Vs All Content.	78

Chapter 1

Introduction

The chapter's content is organized as follows: it provides background information as well as the key terminology used in this study to identify important and non-important citations. These are followed by a rationale for doing this research. Following the motivation section, a problem statement and study objectives were developed as a result of the critical and exhaustive literature assessment. Finally, the chapter closes with the technique used to perform this study as well as the thesis outline, which is presented at the end of this chapter.

1.1 Background Study

Researchers in all disciplines build upon the foundations laid by former researchers. This notion is succinctly summed up in Zimans statement that a scientific paper does not stand alone; it is embedded in the literature of a subject [1]. Research in the same field is interlinked, which means that existing research must always be brought in relation to former researches. New findings must be written up in the form of a scientific research paper. This research paper is then shared with other researchers so that the research process can be validated and can be continued. Therefore, while writing research findings, scholars acknowledge the scientific support they have received from former work. These acknowledgements are found in the reference section and termed as citations. Ziman [1] and Narin [2] highlighted the true strength of analyzing citations can aid in producing and authenticating

different research studies. They argue that the popularity and significance of a scientific work is expressed through the frequency with which it is cited. The citations are considered an important tool for assessing the academic and scientific strength of institutions and individuals. They can also be used to investigate authors or institutions reputations within the overall scientific community [1, 2].

The utility of citation-based measures is multifaceted. They are used to decide award nominees such as the Nobel prize [3] as well as research funding [4]. They can also be used to evaluate peer judgments [5] rank researchers [6, 7] and countries [8]. In the late 1960s, Garfield, the founder of Thomson ISI, defined a number of reasons for citations [9, 10]. This definition offers numerous opportunities to critically investigate citation behaviour [11, 12].

Although citations are included to achieve specific objectives, citation count approaches [13] have never tried to distinguish between these objectives. Consequently, such approaches fail to maintain a balance between the act of citation itself and the purpose for which a citation is made. Instead, they blindly consider all citations equal. This discrepancy has led to achieve research in this area [14, 15]. A detailed examination of citation counts was carried out by [16]. They concluded that citation count-based measures have inherited problems for example, they shift focus from quality to quantity.

Researchers have developed recommendations for improving the quality and reducing the emphasis on quantity in citation counts [17, 18]. Generally, researchers believe that the reasons for citations must be critically considered in order to acknowledge the quality of different scholars work [19]. Is it possible to differentiate between various reasons for citations? Existing citation annotation approaches proceed manually. The manual approaches rely upon interviewing the citer. Usually, authors are interviewed to share the reason for citing a particular piece of work on two different occasions: after the publication process is over and while writing the article [20, 21]. Finney [22] argues that the citation classification process can be automated. Confirming this argument, later researchers took steps to classify citations into various categories [19, 23]. However, while this idea made a significant contribution, it also brought a discouraging element to the fore, as

citations were classified based on several ambiguous reasons. As a result of this ambiguity, the major limitation of a simple citation count approach was not effectively addressed. Presently, two major types of citations have been identified: important and non-important classes [14, 15, 24].

What do we mean by important and non-important classes? Generally, during the process of writing a paper, only a few citations in the reference list have a significant impact on the citing study. This impact needs to be precisely described. Zhu et al. [15] has provided a solution by arguing that an influential research study convinces the research community to adopt or extend the presented idea [15]. To establish a clear distinction between important and non-important citations, we need to examine contemporary citation classification mechanisms. Garzone et al. [23] extends the work of Finney [22] by implementing her suggestion of associating cue words with citation function and using citation location in the classification algorithm [23]. Both Valenzuela et al. [14] and Zhu et al. [15] argue that the citation relations discussed by Finney [22] and Garzone et al. [23] are important. In contrast, Garzone et al. [23] also cite several other studies as background information, such as the citation categories introduced by Garfield [10].

Based on the aforementioned discussion, the studies [14, 15] classify citations into two major categories. The first category of citations aims to provide background knowledge, which forms the foundation of the proposed study. Researchers such as Zhu et al. [15] have termed this category as non-influential and incidental, whereas Valenzuela et al. [14] have termed it as non-important and incidental. We use the term non-important for this category. The second category of citations seeks to extend or apply the cited work. This category is termed as influential by Zhu et al. [15] and important by Valenzuela et al. [14]. We use the term important for this category.

Researchers have recently proposed [24, 25, 26] different features and strategies to identify the important categories. For example, Valenzuela et al. [14] evaluated 12 features and concluded that in-text citation count was the most accurate feature, with a precision of 0.65. However, identifying citation tags from research papers is a challenge [25]. The Valenzuelas approach [14] was further extended recently by

Nazir et al. [26] wherein in-text citation counts within different logical sections of the paper (Introduction, Related Work, Methodology, and Results) were examined. This approach has achieved a precision of 0.84. However, there are also two major issues with this approach: (1) accurately identifying logical sections and mapping section headings onto the logical sections, and (2) accurately identifying in-text citations [25]. The best-known approach for mapping section headings onto section categories has an accuracy of 78% [27].

Another binary citation classification approach presented by Qayyum et al. [24] has achieved a precision of 0.72 by examining metadata and cue phrases. However, this approach again involves the construction of cue phrases and identification of in-text citation frequencies. All of these recent approaches have certain limitations resulting from their reliance on the accurate identification of the following parameters: in-text citation counts, an updated dictionary of cue phrases, in-text citation extraction from sentences, and mapping section headings to logical sections. The extraction accuracy of each of the above parameters is around 70% [25, 27]. However, the above approaches extract these parameters in a semi-automatic way, which has been demonstrated to be accurate when the parameters are readily and accurately available.

This critical discussion highlights the need for an approach that does not involve such a complex extraction of parameters, which is often inaccurate. Examination of the relevant related literature shows that a content-based approaches were successfully employed by 55% of around 200 papers applied during the last 16 years in the domain of document recommendation systems [28]. This motivated us to evaluate the suitability of the content-based approach for identifying important citations. Moreover, in addition to evaluating the existing content-based approach, this thesis further proposes a novel section-wise content based approach.

The results indicate significant precision and recall values without any manual identification of complex parameters. The study's in-depth analysis of [24, 25, 28] papers complete content and content within different sections suggests that content-related similarities in the abstracts of the cited and citing papers be used to classify the citing paper as an important/non-important citation for the cited

paper. Therefore, the proposed approach has a great potential to be applied in citation indexes and open new horizons for future researches in citation classification.

1.2 Preliminaries

1.2.1 Citation

The citation indicates the relationship that exists between the research papers that are citing and those that are cited. A citation is an abbreviated alphanumeric word that appears in the body text of research papers that cite other research documents. This word implies to a string named as reference which can be found in the section called bibliographic of the given article, and its purpose is to describe the relevance of the current scientific work with the other scholarly articles. In most cases, preparing a citation involves both the in-text citation anchor (such as "Liu 2014") and the reference strings in the document. Citations make it possible for writers to link to prior works in a systematic and extremely well-organized way. In the next sections, various types of citation analysis will be discussed.

1.2.2 Citation Analysis

Citation reference strings are first examined exclusively in the bibliography section of the referencing research documents [29]. The importance of the citations has not been considered in the body of the citing research document. This approach of citation-analysis is often referred to as a direct citation, in which the citing article directly cited the explicitly referred research article. For example, there are four research articles named as A, B, C, D, which are published in the years 2000, 2003 2006 and 2008 respectively. The research article A, is also called as cited document and the articles B, C and D are named as citing documents as shown in the following Figure 1.1. This is because the citing documents B, C, and D cited document A in their bibliographic sections. Hence, the citation count for document A is 3, because it is cited by three different documents B, C, D.

Therefore, the citation count is often calculated on the basis of citation analysis. Citation count increases with the passage of time, that is why, it is considered as dynamic measurement.

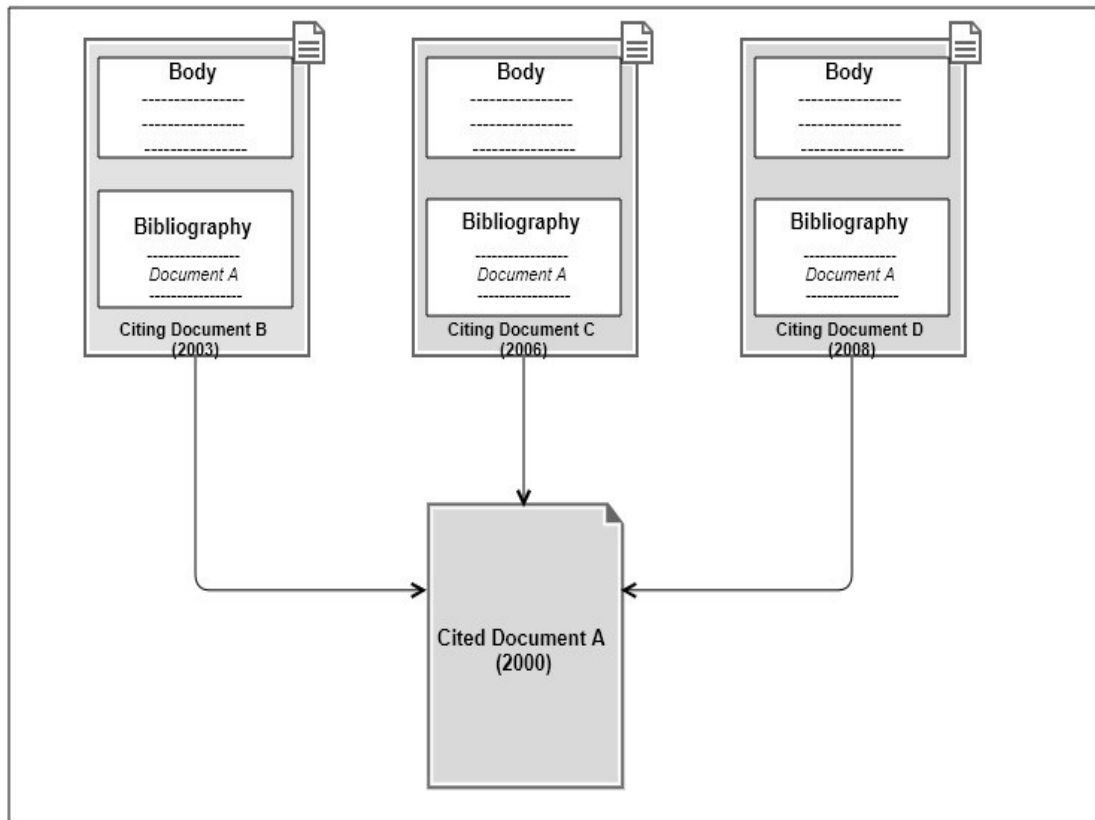


FIGURE 1.1: Analysis of Citation between Cited and Citing Documents

1.2.3 Co-Citation-Analysis

In co-citation [30], two research articles are considered to be similar if they are cited in the bibliographical portion of one or more citing articles. These articles are also referred to as cited and citing documents. **Figure 1.2** illustrates the concept clearly.

There are five research articles or documents named as A, B, C, D, and E. In the bibliography section of the citing papers A, B, and C, the two cited documents D and E are cited jointly. The co-citation weight of the two supplied co-cited

publications (D and E) is 3. This calculation displays the highest co-citation strength. In traditional co-citation analysis, the substance of the citing article is not evaluated for the research paper suggestion.

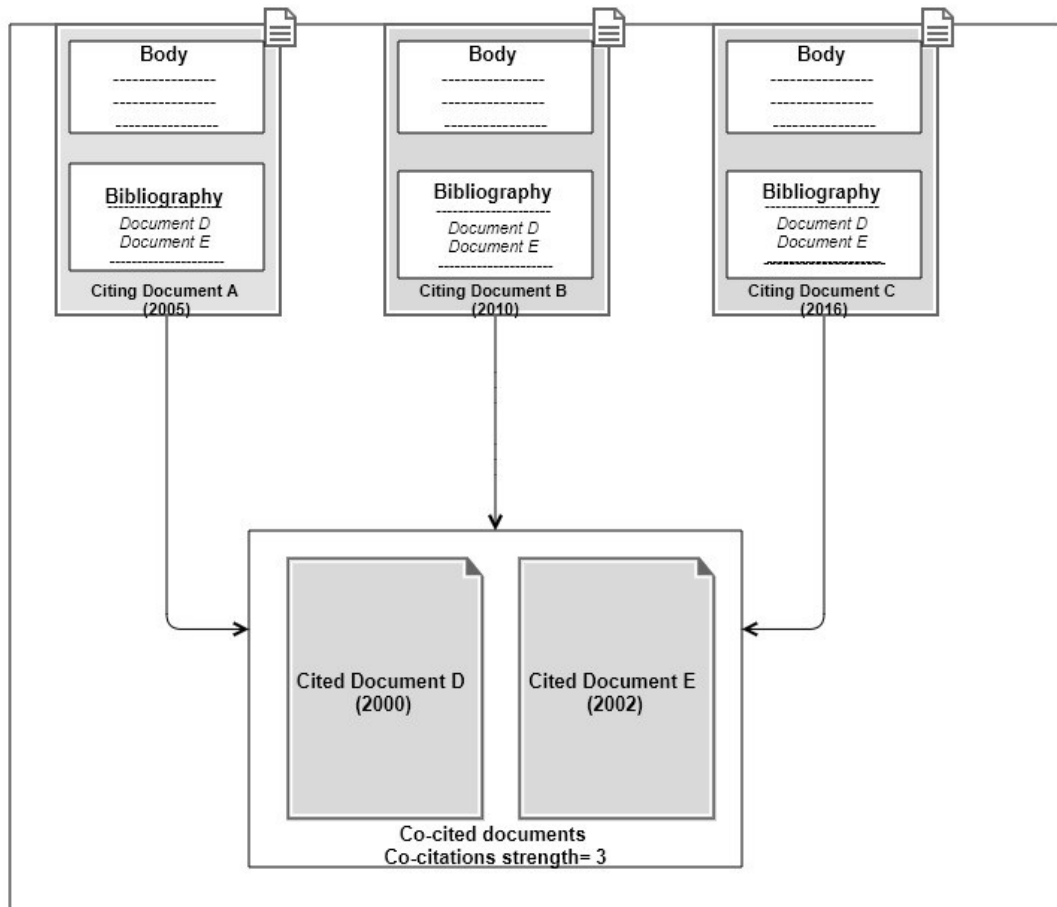


FIGURE 1.2: Analysis of Citation between Cited and Citing Documents

1.2.4 Analysis of Co-citation-Proximity

Co-citation-proximity-analysis [31] is an improved variant of the co-citation analysis. The distance or similarity between the citations shall be measured in the full text of the citing document. The two citations are supposed to be related with each other, if they occurred simultaneously within the full-text document. The CPI (Citation-Proximity-Index) calculation is used to determine the similarity between the two co-cited articles. The probability of similarity between two co-cited articles will be higher if they occurred in the same given sentence say ($CPI = 100\%$ or 1) than the probability of similarity in the same paragraph ($CPI = 1/4 = 0.25\%$).

For example, in Figure 1.3 paper B and C are more strongly linked with each other as they appeared in the same sentence.

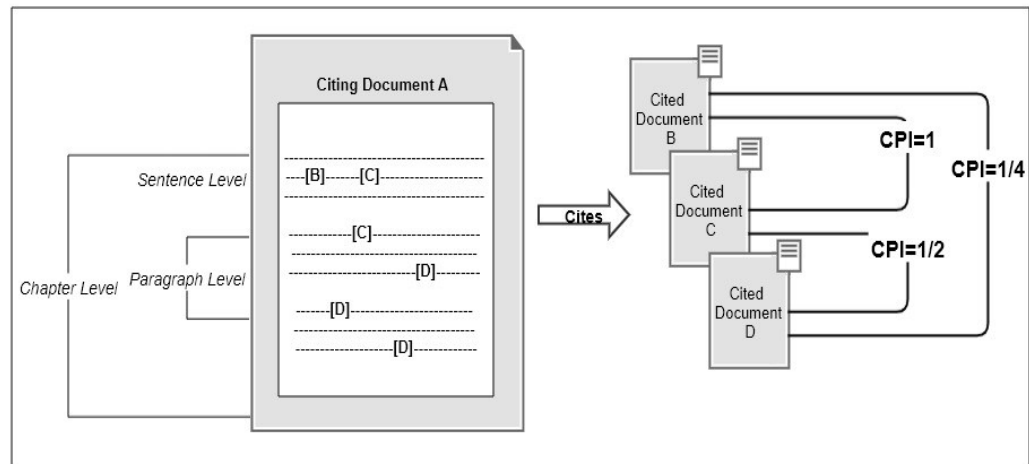


FIGURE 1.3: Co-citation Proximity Analyses

1.2.5 Co-citation-Proximity-Analysis-Based on Byte-offset

Boyack et al [32] performed a co-citation-proximity-study-based on byte offset in a full-text article. The citations are analyzed in separate sizes of byte blocks such as 375, 1500 and 6000 with allocated weights 3, 2 and 1 respectively. For example, the referenced documents B, C, D, E and F are shown in the full text of document A in Figure 1.4. Four pairs of referenced documents have been provided here, such as (B, C), (B, D), (B, E) and (B, F). Citations B, C in pairs (B, C) inside the same bracket, a weight of 4, while citation pairs i.e. (B, D), (B, E) and (B, F) of 375, 1500 and 6000 bytes are given a weight of 3, 2, and 1 respectively. Citation pairs which are more than 6000 bytes apart are given a zero weight.

1.2.6 In-text-Citation-Frequency-Analysis (ICFA)

Initially, Gipp et al. [33] introduces a new measurement in-text citation frequency. Later on shahid et al. [34] used this measurement to identify the relationship of citations across the sections of the citing articles. In-text-Citation-frequency-analysis is the frequency that determines the total number of occurrence of a

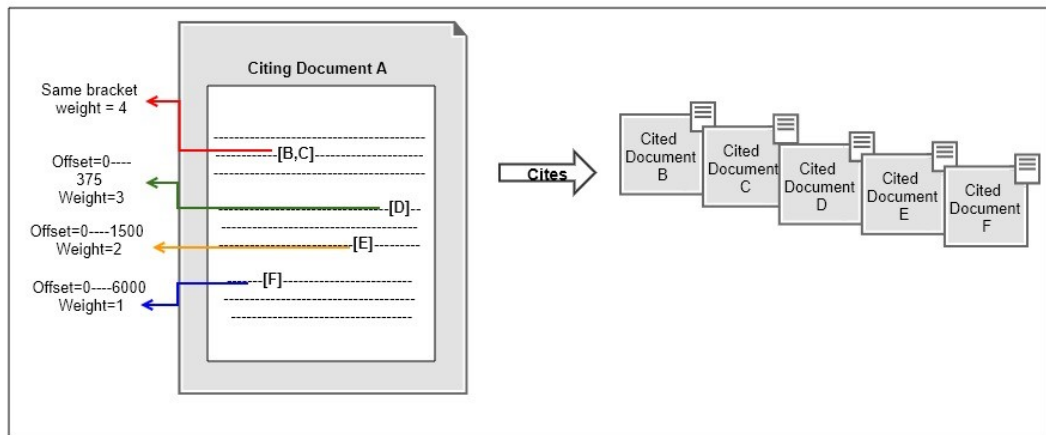


FIGURE 1.4: Co-citation-Proximity-Analysis-Based on Byte-offset

cited document within the citing documents. In Figure 1.5 there are three cited documents B, C, and D, in which B has maximum frequency equals to 4.

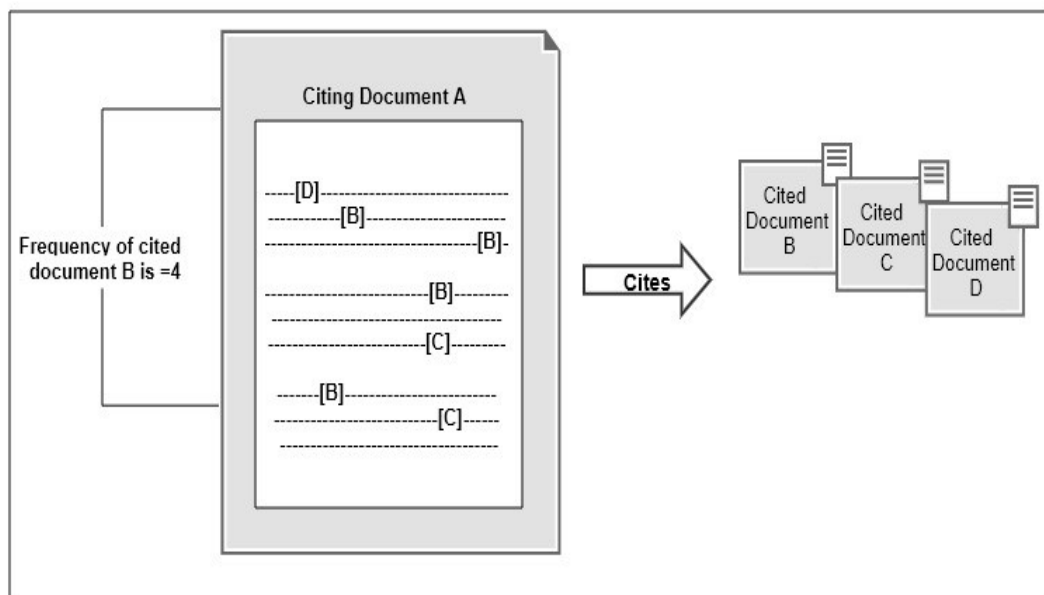


FIGURE 1.5: In-text-Citation-Frequency-Analysis

1.2.7 Content Based Filtering Approach

Content-based Filtering approach exploits contents of research papers to identify similarity and makes decisions by comparing similarities in features. This approach is commonly utilized by the recommender systems [28], to recommend research

articles to users based on information gathered about the user [35]. User modeling is an important process in this technique [36, 37]. The items with which the user has interacted can be used to determine the user's interest i-e downloading [35], authoring [38, 39], tagging [40, 41] and browsing [42, 43, 44]. The user model is made up of item features. There are two types of item features textual and non-textual features. Textual features can be single words, phrases, or n-grams. But the non-textual features include writing style [72, 73], information layout [74, 75] and XML tags [76]. The most descriptive features are used to create items and user models; weights are typically assigned to these features and stored as vectors.

1.3 Research Motivation

We have seen from the above-mentioned literature that there are two major techniques used by Qayyum & Afzal [24] and Valenzuela [14] for the extraction of important citations. Qayyum & Afzal [24] criticises Valenzuela [14] and points out that it depends mostly on the content of research articles that are not publicly available in major journals such as ACM, Elsevier, IEEE, Springer, etc. And they also used 12 different features to achieve 0.65 accuracy. However, Qayyum & Afzal [24] used content-based metadata features (i.e. abstract, references and cue words) and evaluated their approach using unigram, bigram, and trigram. The best combination achieved an accuracy of 0.72 by combining all the features.

Both of these approaches have given importance to metadata and in-text citation counts and their positions, but research papers are written with the help of domain-specific terms and knowledge, and the research gap we have identified is that no one has tried to compare the important terms represented in different corresponding logical sections of the research papers.

This is also supported by another factor of Valenzuela [14] important citations definition. They described that cited papers have important citations with citing papers if cited papers extend or adapt the presented idea of the citing papers. It is more probable that both papers might use similar vocabulary and terms as they belong or they are closely working on the same topic or extending once

work in another work. Therefore, it is more likely that both articles will belong to a common domain and that the comparison paper will use domain-specific terminology and information in the Abstract, Introduction, Methods, Literature Review and Result section. We suggest that no researcher used the domain-specific terminology of the accompanying conceptual parts to classify significant citations.

1.4 Research Questions

Based on the critical analysis presented in the previous section, this thesis evaluates the role of section-wise content similarity for the identification of important citations. The main objectives of this work are to enquire, search, identify and evaluate the role of sections to identify the important citations. In the end, we will be able to conclude that cited paper has an important citation with citing paper if the contents and vocabulary terms used in those papers corresponding logical sections are similar. The Important and Non-Important citations are describing as below [14, 15, 24]:

Important:

The citations in which authors extend or adapt the proposed technique of cited article.

Non-Important:

The citations in which authors just write the background information or some theory portion of the proposed technique of the cited article.

Based on the above discussion this research thesis devised the following research questions.

RQ-1: What is the role of section-wise content similarity for the identification of important citations?

RQ-2: Which independent logical section is performing better in identification

of important citations?

RQ-3: Which combination of logical sections is performing better in identification of important citations?

1.5 Research Problem

To classify a citing paper into one of the two categories such as: Important and Non-important, researchers have utilized different features like: metadata, in-text citation frequency, and cue-phrases etc. However, such features do not capture the relevance of citing and cited papers because the papers present their ideas using the content not using the above mentioned features. Therefore, such relevance may be measured using the content of both citing and cited papers. This thesis poses a research question Whether content based approach could be used to classify citations. Subsequently, this thesis explores all possibilities of employing content based approach for the citation classification task. The proposed approach works into two dimensions such as: (a) considering the complete content for citation classification. (b) Dividing the content into logical sections (Abstract, Introduction, Literature Review, Methodology, and Results) and comprehensively answers the raised question at two levels: (1) identification of the best performing independent section and (2) identification of best combination of sections. Based on the learning from (1) and (2), a novel citation classification approach has been proposed.

1.6 Research Scope

This thesis scope is to exploit the paper-citations pairs to quantify and classify the citations into just two classes as Important and Non-Important (Incidental) citations. The scope of this study is also limited to the available dataset provided by Valenzuela [14]. This annotated dataset has been used for experimentations that contain 465 tuples of root paper and cited paper from ACL anthology. The ACL anthology is a digital archive of research papers in computational linguistics and

a citation network containing only those papers and citations which are published in the ACL anthology itself.

1.7 Applications of the Proposed Solution

This research work will benefit scientific society in a number of areas, some of those most important areas are listed below e.g: Researchers and Authors Ranking, the available systems compute such ranking by using all of the citations in equal capacity. However, important citation should be given more weightage for computation of such ranking systems. Therefore, the following areas can get indirect benefit from our proposed approach and can re-evaluate their proposed rankings by giving due weightage to important citations.

- 1) Researchers and Authors Ranking
- 2) Educational Institutions Ranking
- 3) Countries Ranking
- 4) Journals Ranking
- 5) Peer judgments
- 6) Research funds allocation
- 7) Impact factor calculation of Journals
- 8) Impact factor calculation of researchers
- 9) Researchers Nobel prizes and awards allocation

1.8 Research Approach

The approach suggested by Kumar [45] was used to conduct this research, with minor adjustments to meet the needs of this study. The actions performed during this research are outlined below, and a mapping between these activities and Kumar's model is shown in Figure 1.6. The given approach is based on three main parts (phases). These parts are then subdivided into eight rules/steps.

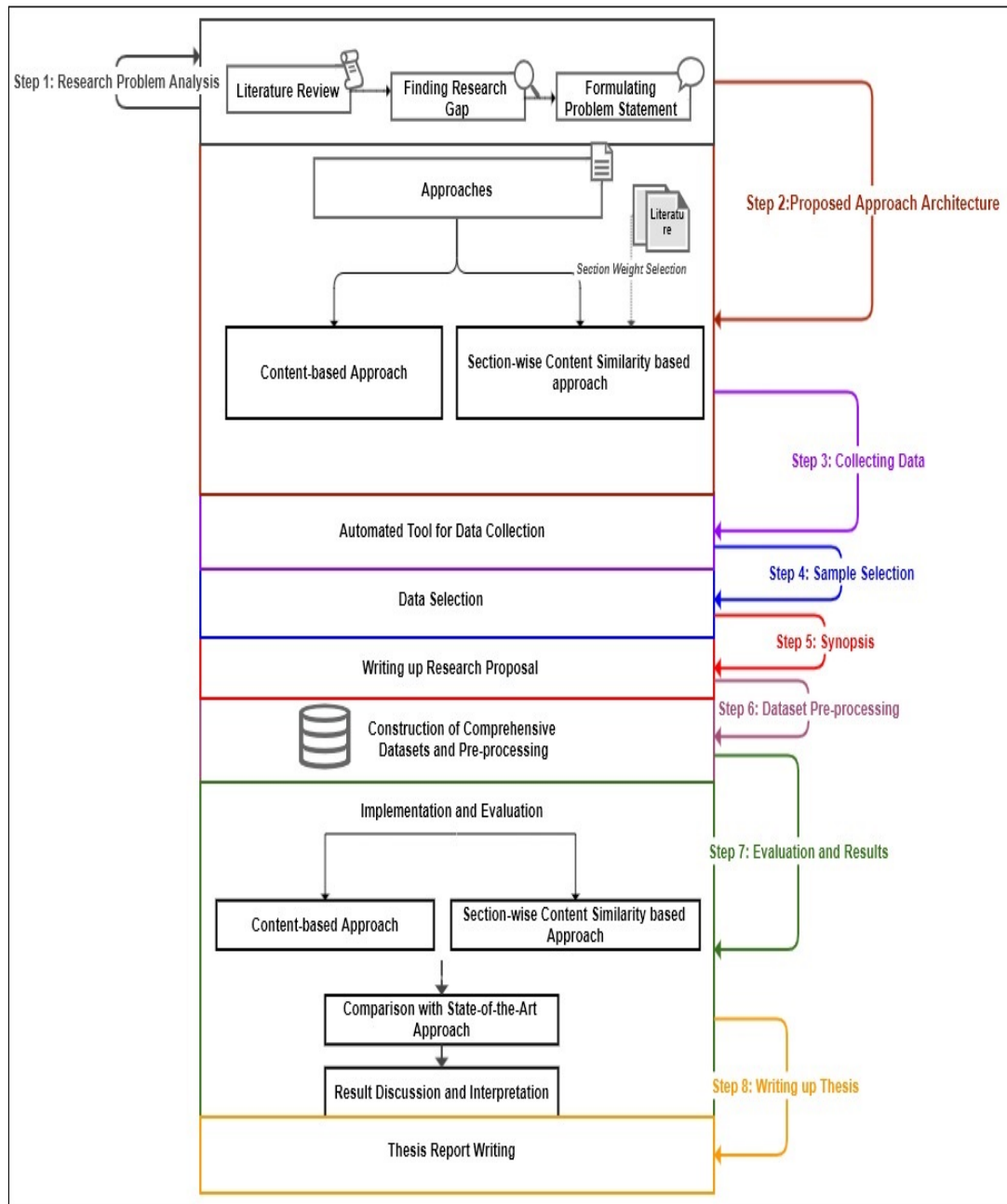


FIGURE 1.6: Research Methodology

Phase I: Choosing What to Investigate

Rule 1: Problem of Research: This stage includes three tasks. (1) Review of the literature; (2) identification of research gaps; and (3) development of research problems.

Phase II: Planning to Conduct Study and Research

Rule 2: Proposed Technique Design: We first presented the unique Section Wise Content Similarity strategy based on rule 1, and then created the technique for carrying out the proposed approach.

Rule 3: Method of Collecting Data: In the stage 3, we have developed the automated tool to collect research papers.

Rule 4: Selecting Sample: In this stage, we randomly picked a sample of research articles from the collection of articles obtained in rule 3.

Rule 5: Synopsis: Following the initial experiment in this study, we created the basic document based on initial experiments.

Phase III: Research study implementation.

Rule 6: Pre-processing Dataset : This stage is used to produce the comprehensive datasets of semi-structured research articles documents, as well as any necessary preprocessing.

Rule 7: Results and Evaluation: In this stage, the suggested approach's results will be reviewed and discussed in comparison to state-of-the-art techniques.

Rule 8: Thesis: This is the final phase in our research approach, and we have completed the thesis document.

1.9 Research Contribution

The main contributions of this research work are (1) Identification of singular features of research articles, (2) Identification of combination of features of research articles and (3) section wise content similarity approach.

1.10 Thesis Outline

This dissertation comprises of five chapters. Introduction and literature review of the proposed work is discussed in chapters 1 and 2 respectively. In Chapter 3, the building blocks of proposed approach, as well as the main contributions of

this research, have been elaborated. The results of research contributions are discussed comprehensively in the Chapter 4. The last chapter concludes the research presented in this thesis. Furthermore, limitations and future work have also been presented, for the scientific community to quickly identify the current research challenges and research dimensions in this area.

Chapter 2

Literature Review

In this chapter, a comprehensive literature survey is elaborated to understand the scope and importance of the current research work, which contributes in the following manners (1) Identification of singular features (2) Identification of combination of features and (3) Section wise content Similarity. The detailed literature review written in the upcoming sections provides support to all the contributions done by this research work. It describes the recommender systems approaches, along with the citation based state-of-the-art approaches.

2.1 Relevant Research Paper Identification

Although citation classification is a different research task than finding a relevant research paper. Citation classification attempts to identify the relevance between two papers (cited and citing), and classify citing paper into different categories such as: Important and Non-important in the context of the cited paper. However, finding relevant papers, attempts to find a relevant paper for the focused paper from a pool of the papers. Therefore, this chapter highlights the important techniques proposed in the area of finding relevant research papers. Based on this critically analysis, best practices and approaches will be adapted from this closely related domain for the task of citation classification. In the last decade, the researchers have been investing a tremendous amount of effort on research paper recommender system for recommending the most relevant research papers from

the huge bulk of research articles. There have been 216 approaches proposed in the last 16 years for the recommendation of research papers [28]. But none of the given techniques is considered to be the best for the recommendation of research literature. One of the biggest limitation in all of the given technique is the lack of evaluation criteria. Beel et.al [28] investigated that it is very difficult to recommend the best approach because all the given approaches do not follow all the important and relevant aspects to recommend the important research articles. In the course of our study, we have gone through approximately 80 research papers to analyze this hot issue for recommending the most important research paper. This study led us to the conclusion that we can classify the recommender system approaches into the following classes: stereotyping [46], co-occurrence based [42, 43, 44] Collaborative Filtering [47] Content based [28] and citation based [30, 48]. In the following text, these different techniques have been discussed briefly followed by the research gap identified by this thesis.

2.1.1 Stereotyping

Stereotyping is considered to be the most primitive modeling technique in the recommendation class. Initially it was presented by Rich in the recommender system developed by Grundy [46]. This recommender system suggests novels (work of fiction) to the users based on their interest. Rich was impressed by the stereotyping that was used in psychology. The stereotyping was used by psychologist to judge peoples behavior immediately by observing few characteristics. These were called as facets by Rich which means collection of characteristics. For example one might be considered as judge then he or she probably be forty years old, well educated, honest, reasonably established, fairly affluent and well respected in the society. The aforementioned points are characteristics or facet of a judge. The researchers for the recommendation systems also took the motivation from the stereotyping to recommend the relevant research papers. For example if a person provided a search information retrieval in the recommender system. Probably he or she will be beginner or expert. If a person is a beginner in this field then characteristic or facets for the beginners are as follows, some of which can even be

extracted automatically, he/she has not studied preliminary courses in the area; casual reader, performing breadth wise searching. On the basis of above mentioned characteristic, we can assume that his or her intention will be looking for very basic information or material about information retrieval and the search result will be filtered accordingly. However if the same search query is given by some expert, then the characteristics or facets for the expert will be: consistent and focused reader in this particular area, good concepts in relevant field, knowledgeable person, already studied preliminaries etc.

The search result will be filtered on the basis of characteristics defined for the expert, maybe he or she will be looking for more advanced articles or research topics in the same field. Here the characteristic of user that is collection of facets will define the class of a user either he/she is a beginner or expert. This procedure is called as classification of data. Consequently the search result will be filtered on the basis of two classes that are beginner and expert in this case.

In the field of recommender systems Beel et.al [28] remained the prominent researchers for using stereotyping [49, 50, 51]. They considered their users of Docear are only students and researchers. Hence the most relevant books and papers are recommended by recommender system to the researcher and students. Beel.et.al uses stereotyping as a backup system, when all the approaches failed to produce appropriate outcome. They have observed that stereotyping approaches perform on the average 4% with click-through rates (CTR) whereas content based approaches perform 6% on the average with CTRs.

In stereotyping there are certain limitations: firstly, they have pigeonhole users (The pigeonhole principle states that if n items are put into m containers, with $n > m$, then at least one container must contain more than one item). Secondly, all the persons may not be categorized as beginners or expert because in some cases may be an expert person wants to look into the very basic information. This does not fit well in case of all persons. In addition to this, stereotyping performs manual classification for each single facet, which is highly laborious work. Because of this the number of items becomes limited.

2.1.2 Co-occurrence Recommendation

The important aspect of co-occurrence recommendation is to emphasize the relatedness of the papers instead of similarity. The definition of similarity in document recommender system is to calculate that how many common features lie within the two papers. On the other hand, relatedness does not depend on the feature of two items instead it is focused on how closely the two items are coupled together. For example, two documents A and B are considered to be similar if both have similar features (words). Whereas there is no similarity between the papers and pen but both are considered equally important and relevant because for writing a letter both are needed and hence they are related to each other. Consequently, co-occurrence recommends more serendipitous suggestions, which means that a researcher may find some interesting discovery while searching for something else, and in this way they are equal to collaborative filtration. Furthermore the computation complexity is low and no access to content is required. To ensure the privacy of users anonymous recommendation is generated.

In this approach, the recommender system considers those items as potential candidate that are co-occurred frequently in some source documents. The very first application of Co-occurrence was presented by Small et al [30] in 1973. They introduced the technique known as co-citation analysis. Small proposed that if the two papers are co-cited frequently then these two papers are considered to be more relevant. Later the same concept was implemented by many scientists. The well-known example, which uses this concept, is Amazons. They also analyzed that when items are bought with the product. The recommender system considered those items as frequent items for that particular product. Whenever the product is browsed again by any user. The recommender system will recommend the items that are related with the product. Hence this also ensured the concept of relatedness.

In case of research papers, the recommender system recommends those papers to the users that are co-viewed frequently by the user with the browsed paper [42, 43, 44]. To determine the relatedness between the papers, an extended for of co-citation is known as Citation Proximity Analysis (CPA) used [31]. It enhances

the precision by considering the position of citation within the text. The cited papers are believed to be more relevant if the proximity of references is closer enough within the paper. The key benefits of Citation Proximity Analysis are: (1) Identification of relevant section within the research article and (2) greater precision as compared to co-citation analysis, keyword based approaches and bibliographic coupling.

The significance of co-occurrence is not suitable all the time. For instance on arXiv.org, 66 percent papers of total volume do not have co-citations. Whereas those papers that contain citations are not more than one or two [52, 53]. Furthermore, the recommendation can only recommend papers if they co-occurred at least once with another paper.

2.1.3 Collaborative Filtering

Goldberg et al [47] proposed the idea of collaborative filtering in 1992. In this work it was proposed that if we involve peoples in the filtration process then the information filtration will become more effective [47]. The idea that is recognized today of collaborative filtration is coined two years later by Resnick et al [54]. Their idea was that users like what like-minded users like, where two users were considered like-minded when they ranked the same products as well. When like-minded users were recognized, products that one user ranked favorably were suggested to the other user, and the other way around. In comparison to CBF, CF offers three advantages. First, Collaborative filtering does not depend on contents, [55, 56, 57, 58]. Second, because people do the scores, CF performs quality of evaluations [59]. Finally, CF gives serendipitous suggestions because suggestions are not depending on product likeness but on user likeness [57, 60, 61].

Yang et al [62] was expecting that the user will rate the document, but the sluggish behavior of the users extremely affect the rating. Naak et al [63] experienced the same issue and created artificial rating for their assessment. This demonstrates one of the main problems. Collaborative Filtering needs user contribution, but it is observed that the participant motivation to take part is too low. This issue is known as the cold-start issue, which may occur in three situations such as: new

areas or professions, new products, new users [56]. In a new community, no users have ranked products, so no suggestions can be made and as a result, the inspiration for users to rate products is low. If products is new and has not been ranked yet by at least one user, it cannot be suggested. If a new user rates few or no products, the system cannot find like-minded users and therefore cannot offer suggestions.

To get rid from the cold-start issue, implicit scores may be deduced from the communications between users and products. Yang et al [62] deduced implicit scores from the number of webpages the users read. The more webpages users study, the more the users were believed to like the documents. Pennock et al [35] considered communications, such as downloading a document, including it to user profile, modifying document details, and viewing its bibliography as beneficial ballots. McNee et al [61] believed that an authors citations indicate a positive trend towards the document. They postulated that when two writers cite the same document, they are like-minded. Similarly, if a user flows or points out a document the citation of the mentioned document should be liked by the user.

Generally, reliable citation information is not widely available. Therefore, access to the papers submissions is needed to build a citation network, but this process is even more fault-prone than word extraction in CBF. In CBF, the written text of the documents must be obtained, and maybe the sections such as the title or abstract must be recognized. For citation-based CF, the written text must also be extracted but in the written text, the bibliography and its individual references must be recognized, such as title and writers etc [64].

A general issue of collaborative filtration in the area of research-paper recommender techniques is sparsity. Vellino et al [65] compared the implied scores on Mendeley and Blockbuster online (movies), and discovered that sparsity on Blockbuster online was three order of magnitude lower than on Mendeley [40]. This is due to the different ratio of users and products. In domain like film suggestions, there are generally few products and many users. For example, the film recommender Movie Lens has 65,000 users and 5,000 films [58]. Generally, many users viewed the same films. Therefore, like-minded users are available for most users

and suggestions can be given effectively. In the same way, most films have been viewed by at least some users and hence most films can be suggested. The situation is different in the area of document analysis. There are generally few users but an incredible number of documents and very few users have ranked the same document. Finding like-minded users is often not possible. In addition, many documents are not ranked by any users and therefore cannot be suggested.

There are further aspects associated with Collaborative Filtering. Processing time for CF tends to be higher than for content-based filtration [56]. Collaborative filtration is generally less scalable and needs more off-line computation than CBF [66]. Torres et al [55] note that collaborative filtration makes similar users and Sundar et al [67] noticed that collaborative filtration dictates views. Lops et al [68] makes the critique that collaborative filtration techniques are black boxes that cannot describe why products is suggested except that other users liked it. Adjustment is also a problem: since collaborative filtration is dependent on user views, blackguards might try to operate scores to advertise their products so they are suggested more often [69, 70, 71].

2.1.4 Content Based Approaches

Content based techniques are most dominated techniques in the era of research paper recommendation approaches [28]. Interaction between user and items are normally developed through downloading [35], adding social tags [40, 41], reading [62], browsing papers [42, 43, 44], authoring [38, 39], having papers in personal databank. In this technique, the user modeling is an important process. The interest of user can be deducted from the items with which user has interacted. Items are normally textual for example email [36], webpage [37] etc. The interaction of the user is defined by the actions he has taken during the search process i.e downloading [35], authoring [38, 39], tagging [40, 41] and browsing [42, 43, 44].

The user model consists of items features. There are two type of item features textual features and non-textural features. The textual features are single words or they may be phrases or n-grams. But the non-textual features are writing

style [72, 73], information layout [74, 75] and XML tags [76]. The most descriptive features are used to develop items and user models commonly weights are assigned to these features and stored as vectors. The user models are designed by using these item features. Therefore to recommend research articles to the users the features involved in the modeling process are compared with the recommendation candidates using cosine similarity and the vector space model.

The discriminative power of words is different in different fields of research articles [77]. The word lies in title are more influential as compared to the word that lies in the body text. Nascimento et al [78] performed experiment and found that the weighted terms in title are three times stronger than the body text. Whereas the words lies in the abstracts are twice stronger than the body text [78]. The techniques used for weighing the terms are arbitrarily selected. These techniques are not selected on the basis of practical evidence.

The techniques extracted words from header [79], title [80, 81, 82], abstract [83, 84, 85] keywords [83, 86, 40], introduction [40], foreword [40], bibliography [87], and body text [78, 82, 88]. The words can be extracted from different other sources such as ACM classification tree and DMOZ [89, 90] categories and citation context [84, 40, 88]. Many techniques utilizes plain text as feature and some other technique uses n-grams that is combination of two words or a single word that appeared as social tag [40, 41] on CiteULike [85] (topic) and concept.

The Vector Space Model (VSM) is the most dominating model used in this domain. There are 9 approaches out of 14 approaches those who have utilized Vector Space Model for storing features and developing user models. Other techniques uses graph based approaches to represent and store their users. The graph based approaches model their users as a list of topics. These topics are allocated by applying machine learning techniques or by applying ACM category hierarchies.

In the content based approaches, there is a need of weighing scheme. TF-IDF is the most popular weighing scheme. Nearly seventy percent of the approaches use this scheme. There are other weighing schemes as well such as plain term frequency and the technique that are known as phrase depth and life span. Content based

filtration has many advantages over the stereotyping. CBF perform user-based personalization to recommend the most relevant document for each individual user rather than being limited by stereotype. In content based filtration approaches, user model is built automatically. Therefore it requires less up-front classification work.

Nascimento et al [78] proposed a technique for the recommendation of research articles. The recommender system designed by Nascimento et al [78] is based on content-based approach. This technique takes research paper contents (words) as an input, then generates keywords and several queries to perform its task of recommending research papers. The generated keywords and queries are then submitted to different digital libraries. These digital libraries contain the source articles and as a result of submission (keywords and queries) the given libraries generate a candidate set of source articles which are freely available. The articles are then ranked by applying a content-based algorithm. For the recommendation of most relevant papers. The given technique uses title and metadata of abstract.

Ferrara [91] proposed a research article recommendation technique build on content-based algorithm. In this approach they build the profile of users and documents by extracting the key phrases. This is done by KPEM (Key Phrase Extraction Module). The creation of users profile is carried out by utilizing tagging mechanism which was assigned earlier. Users profile and documents are considered to be the basis for the recommendation to research articles. In order to accomplish this task uni-gram, bi-gram and tri-grams are used. To produce the experimental results and evaluation, ACL anthology reference dataset is used. This dataset is freely available. The given dataset contains 597 research articles written by 28 researchers. The experimental results obtained from the given dataset shows that tri-gram performs better than uni-gram and bi-gram.

Pruitikanee et al [92] presented a fuzzy clustering based technique for the recommendation of research article. There are four basic steps to accomplish the task of this proposed technique. In the first step input is given in the form of query as a result of which it returns all the research papers which contain at least one keyword in the query. In the second step fuzzy clustering is utilized to classify

all the documents according to the similar topics and interest. In the third step, user interaction with huge databases is minimized by computing the representative research papers. In the last step, a classical rank model (i.e. page rank) is utilized to rank the representative research articles.

2.2 Citation Based Approaches

Kessler et al [48] presented one of the oldest citation based approach known as: Bibliographic Coupling for finding related research articles. They proposed that two research articles A1 and A2 are bibliographically coupled if they have cited the same research document. In the course of experiments, Kessler discovered that if two documents have more than one common paper cited, then there is a close association between the two papers. The experimental evaluation of recommended approach is carried out by selecting 8,521 research articles from 795 different resources having 137,000 references.

The benefit of this strategy is that the citations of documents are easily accessible. However, the restriction of this strategy is that the citations are not very easily mined, due to the different types of citations formats [93, 94]. Furthermore, the association between two documents remains static, because the references of published documents are not modified in future.

Bichteler et al [95] suggested a hybrid technique for relevant document recommendation. They suggested technique by combining co-citation and bibliographic technique to find the relevant documents. They performed a user study on 1,712 documents. Their outcomes revealed that using both the bibliographic coupling of documents and their co-citations i.e. cited and citing documents, give better outcomes in comparison with using only the co-citations or bibliographic coupling.

The primary restriction of this strategy is that it doesn't take into concern, the real contents of the targeted documents to evaluate the relevance. Gori et al suggested a Google Page Rank methodology, to rank the research documents [28]. This strategy is dependent on the Random Walk Approach. Gori et al determine

two properties which are used in this strategy. These properties are attenuation and Propagation. Propagation indicates that if a document is linked to an excellent document, then this document itself is an excellent one too. Attenuation means that an excellent document should also distribute its impact to other documents to which it is linked. But this decreases the significance of the document itself. To control this, a decay factor is presented. They think that a person has a partial document with some bibliography already published. This is used as an input. An undirected graph of cited/cited-by is used to present the citations. From this graph, a connectivity matrix is built and gradually a stochastic and correlation matrix is produced. These matrices help in discovering the link between documents. Authors used a specific web crawler to gather dataset from ACM. Both off-line and online assessment techniques were used, which revealed a 100% ranking of most relevant documents. The restriction of this strategy is that it has not been examined for the negative documents. What would be the results if completely irrelevant documents are given as an input?

Strohman et al [96] suggested a technique for ranking the relevant papers by supposing that a user has submitted his incomplete or unpublished papers as a query to the recommender system. They applied different similarity measures to find the relevant articles such as: featured based, text based, and citation based similarity. They have used distinct features such as: publication year, text similarity, co-citation coupling, same author, Katz and citation count. Their strategy consists of two steps. (1) the user submit a query, and the recommender system returns 100 most similar documents out of over 1 million documents. (2) All documents cited by any of these 100 documents are also added to the same list. Now the resultant list contains almost 1000-3000 documents. The above mentioned features (publication year, text similarity, co-citation coupling, same author, Katz and citation count) are then used to rank the documents. The documents scores are computed by combining the features in a weighted linear model. The writers examined this procedure on Rexa database. They discovered that Katz feature is very important. If the Katz feature is removed then the performance falls down to 50%. Their methodology outperformed the text based similarity. Krapivin et al [97] suggested a technique for finding the relevant documents and to position

(ranking) suggested a technique for finding the relevant documents and to position (ranking) the relevant documents by using the famous Web Page Rank algorithm [97]. Web Page Rank algorithm is used by search engines to position the web documents. The critical point in this algorithm is that it uses the in-bound links to determine the importance and position of a document.

The greater the quantity of in-bound links to a paper ABC, and higher the page position of the source page (of that link to ABC), the greater will be the position of ABC. In page position, the outgoing links also matter. The greater the quantity of outgoing links is, the lower is the position of the given papers. This may be an issue in case of scientific papers, since the survey documents normally have a lot of outgoing links citing other documents. And this large amount of outgoing links should not decrease the position of the document.

To eliminate this issue, Krapivin [97] suggested the new strategy called Focused Page Rank. Also, the recently published documents normally dont have a lot of incoming links, but that doesnt mean that they are not associated with a given document. Krapivin [97] analyzed their strategy by exploring 266,788 documents, published in ACM conferences and journals. Krapivin [97] found that implementing the Focused page Rank on documents gives almost the same results as those of citation count and are much better than implementing the simple Web Page Rank algorithm.

2.3 Citation Count

The citation count is utilized to conduct various types of bibliometric analyses with multidimensional utilities. Such analyses have been used to build indexing systems [79, 98], and formulate various academic policies and present awards such as Noble prizes [3].

These analyses have also been used to rank researchers [6, 7] and countries [8]. However, researchers believe that all citations cannot be considered equal, and each citation should be treated according to its true standing [1, 12, 14, 15, 19, 99].

2.4 Citation Behavior

A specific study might be cited for myriad reasons. Garfield [10] was the first researcher to analyze citation behavior [10]. He identified 15 citation reasons by examining various factors such as the citations location in the paper and scrutinizing the differences and patterns. Some of the reasons include (1) acknowledging the contributions of predecessors, (2) highlighting fundamental contextual details, (3) extending existing work and targeting expanded objective(s), etc. Later, Lipetz et al [100] identified various classes of citations.

However, while both studies appropriately conceptualized the notion of citation reasons, no statistical measures were introduced [12]. Nevertheless, despite this shortcoming, these studies attracted enormous attention from the research community. Consequently, many empirical investigations have been carried out to identify citation reasons. Subsequently, other studies have attempted to capture actual citation behavior [11, 101]. However, common to all of these approaches is the treatment of all citations as being on the same level of importance.

2.5 Quantitative Citations

According to Zhu [15], studies involving straightforward quantitative citation analysis can be enhanced by eliminating incidental citations from the citation count. Additionally, maintaining a list of only important citations can be of substantial help for scholars seeking to identify influential studies on a specific topic. Until the mid-1990s, citation reasons were manually identified.

For example, a general trend at that time was to interview authors during the process of writing an article or after their proposed article had been formally published, requesting them to describe the specific reasons for citing particular works [20, 21]. However, differentiating scholars citing behavior using cognitive approaches seemed rather impractical. Therefore, researchers have realized the need for an automated system to identify and classify citation reasons.

2.6 Citation Classification

Finney [22] demonstrated that the citation classification process can be automated. She created a citation function on an experimental basis. Later, she posited a relationship between cue-words and citation location and combined it with the citation function. Though her approach was not fully automatic, however, it underscored the probability of developing a fully automatic citation classification mechanism in the future [23]. However, other researchers were a bit reluctant to acknowledge Finneys contribution because it was a doctoral thesis rather than a formal publication [12].

2.7 Automated Citation Classification

Drawing inspiration from Finneys approach, Garzone and Mercer [23] took their place among the trendsetters by creating an automated citation classification system. The authors argued that Finneys approach had several limitations, which they addressed by creating 35 categories to classify citations. They were able to successfully solve the classification task by introducing 14 parsing rules and 195 lexical matching rules. The dataset comprised 9 biochemistry and 11 physics articles. The system found to be stable. It produced average results on unseen articles and appreciative results on previously seen articles. Though the system produced encouraging results, there was also one concern: due to many classes, the system was unable to neatly distinguish between divergent classes. Pham and Hoffman [102] classified citations from 482 citation contexts into four categories. They employed the ripple-down rules (RDR) hierarchy using cue-phrases.

Zhu et al. [15] characterized academic influence as a reference that serves as a source for extracting an idea, problem, method, or experiment. They generated a total of 3143 paper-reference pairs from 100 papers extracted from the Association of Computational and Linguistics (ACL) anthology. These pairs were annotated by the authors of the citing papers. In contrast, Valenzuela et al. [14] introduced a supervised classification approach to identify important and non-important citations. The authors have extracted 465 paper-citation pairs from the

ACL anthology. Two domain experts annotated these pairs as either important and non-important citations. Inter-annotator agreement between two annotators was 93.9%. Twelve features were used to classify the citations into important and non-important classes. These features include the total number of direct citations, the number of direct citations per section, the total number of indirect citations, the number of indirect citations per section, author overlap, etc. Two different classifiers, Random Forest and Support Vector Machine (SVM), were used to train these features. Both models attained 0.90 recall and 0.65 precision. Zhus approach [15] was criticized by Valenzuela et al. [14]. The latter claimed that biased annotation cannot be ruled out if citations are coded by the citing authors.

Another approach was proposed by Qayyum and Afzal [24]. They utilized metadata and cue-phrase to discover important citations. However, a limitation of this approach is that cue-phrases are identified from the papers content and thus need to be updated for different datasets and domains. There is a need for a domain expert to manually identify cue-phrases for each domain and keep them updated. Nazir et al [26] extended Valenzuelas approach by identifying suitable weights for in-text citation frequencies in different sections. This approach has outperformed the previous approaches. However, a critical examination highlights limitations regarding accurately mapping section headings onto logical sections and accurately identifying in-text citations. Although there are many approaches to identifying important citations [14, 24, 26], in order to practically apply those approaches, there is a need to accurately identify the following information: (1) accurately identify in-text citations [14, 26], (2) accurately map section headings onto logical sections [26], and (3) create an updated accurate list of cue terms [24]. Existing approaches have reported precisions up to 0.84. However, these approaches depend on the accurate extraction of the above parameters and either ignore inaccurate results and correct the missing values manually to demonstrate the power of these parameters. The automatic extraction of such parameters is still a challenge [25, 27]. This motivates us to fill this research gap by creating a novel approach that does not require these parameters to be extracted. A critical examination of the related domain of research paper recommendations motivated us to use the content of the cited and citing papers. A survey paper by Beel

[28] indicates that more than 55% of more than 200 articles on research paper recommendation in the last two decades used a content-based filtering approach.

2.8 Hybrid Approaches

Singla et al [103] suggested a hybrid strategy for finding the relevant documents. They used a hybrid of citation and content based techniques. In this strategy, the writers first of all, created an overview for the documents using the keywords, which are extracted from the title and the sources of the documents. 15% phrases of total summary are from the abstract, and the rest of the 35%, 20% and 30% are from introduction, relevant work and methodology segments respectively. The review of cited by documents is also designed in the same manner. The summaries are then in comparison using some language be evaluated. This generates a similarity ranking. The document position is then acquired by splitting this similarity ranking with the number of documents that cited the document.

The writers used this procedure on 10 documents and determined that there is important enhancement in the ranking. To get better ideas into the performance or outcomes of this procedure, a better analysis on a larger dataset needs to be conducted.

Reyhani et al [104] suggested a methodology known as SimCC to find the similarity between two documents. Similarity between two documents is regarded as a contribution score of the cited document into the citing document. According to this strategy, the quantity of citation (alone) obtained by a document doesn't illustrate the association between documents truly. Citation and content both need to be used, simultaneously. SimCC ranking of a cited document is measured for each phrase. To determine this, contribution score is included to the relevance score. Relevance score is the TF/IDF value of the term. To determine contribution score, a complicated recursive procedure is implemented off-line. The writers in comparison the outcomes of SimCC with cosine, Cube, BM25 and Kullback Leibler Distance and discovered that there was a 60% enhancement. Nassiri et al [105] suggested a strategy based on citation network. They suggested a strategy

known as Normalized Similarity Index (NSI) to evaluate the similarity between two documents. In NSI, three types of citation relationship are taken into consideration. These includes longitudinal coupling, co-citations, bibliographic coupling. Indirect citation between two documents is referred as longitudinal coupling, i.e. the two documents are linked through some other inter-linked document. They measured NSI for five different citation networks. The results are compared with the peer reviews. There was a high correlation between the two outcomes. In comparison with combined linkage (CL) and weighted direct citation (WDC) of those 5 networks. NSI produces much better results.

2.9 Critical Analysis

In background study, we have thoroughly examined the state-of-the-art approaches that are utilized for the citation classification. During the course of our study, we have discovered that citation classification methodologies rely heavily upon in-text citation count, Cue phrases, and the content of the research article. Table 2.1 shows a summary sketch and overview of various methodologies, along with their outcomes, features and limitations.

According to the preceding background study, Valenzuela [14] and Qayyum & Afzal [24] presented two basic methodologies to classify citations into (1) important and (2) non-important categories. Qayyum & Afzal [24] criticize Valenzuela's [14] approach, claiming that it is based mostly on the content of research publications that aren't publically accessible by major journals such as Elsevier, ACM, Springer, IEEE, etc and that they also utilized 12 different features to get a 0.65 accuracy. Qayyum & Afzal [24] combined metadata-based features (i.e. title similarity, author overlap, references) with the contents feature (i.e. abstract and cue words) and evaluated their methodology by using unigram, bigram, and trigram.

Both of these methodologies have given importance to the position of citation, in-text citation counts and metadata. It is obvious that the research papers have been written by considering knowledge and domain-specific terminologies; as a result, we have concluded that no one (i.e authors and researchers) has attempted

to compare the contents of corresponding logical sections of research publications. In order to identify important and non-important citations.

In the field of recommender systems Beel et.al [28] remained the prominent researchers for using stereotyping [49, 50, 51]. They considered their users of Docear are only students and researchers. Hence the most relevant books and papers are recommended by recommender system to the researcher and students. Beel.et.al uses stereotyping as a backup system, when all the approaches failed to produce appropriate outcome.

They have observed that stereotyping approaches perform on the average 4% with click-through rates (CTR) whereas content based approaches perform 6% on the average with CTRs. In stereotyping there are certain limitations: firstly, they have pigeonhole users (The pigeonhole principle states that if n items are put into m containers, with $n > m$, then at least one container must contain more than one item).

Secondly, all the persons may not be categorized as beginners or expert because in some cases may be an expert person wants to look into the very basic information. This does not fit well in case of all persons. In addition to this, stereotyping performs manual classification for each single facet, which is highly laborious work. Because of this the number of items becomes limited.

In case of research papers, the recommender system recommends those papers to the users that are co-viewed frequently by the user with the browsed paper [42, 43, 44]. To determine the relatedness between the papers, an extended for of co-citation is known as Citation Proximity Analysis (CPA) used [31]. It enhances the precision by considering the position of citation within the text. The cited papers are believed to be more relevant if the proximity of references is closer enough within the paper.

The key benefits of Citation Proximity Analysis are: (1) Identification of relevant section within the research article and (2) greater precision as compared to co-citation analysis, keyword based approaches and bibliographic coupling. (3) It enhances the precision by considering the position of citation within the text.

TABLE 2.1: Critical Analysis of Different Approaches Containing Parameters, Contribution, and Claims

Ref	Parameters	Contribution	Claims
[51]	Collection of characteristic / facets for beginners or expert users.	The search result will be filtered on the basis of two classes that are beginner and expert.	<p>In stereotyping there are certain limitations: firstly, they have pigeon-hole users. Secondly, all the persons may not be categorized as beginners or expert because in some cases may be an expert person wants to look into the very basic information.</p> <p>This does not fit well in case of all persons. In addition to this, stereotyping performs manual classification for each single facet, which is highly laborious work. Because of this the number of items becomes limited.</p>
[52]	The important aspect of co-occurrence recommendation is to emphasis the relatedness of the papers instead of similarity. The relatedness does not depend on the feature of two items instead it is focused on how closely the two items are coupled together.	In case of research papers, the recommender system recommends those papers to the users that are co-viewed frequently by the user with the browsed paper, to determine the relatedness between the papers.	The significance of co-occurrence is not suitable all the time. For instance on arXiv.org, 66 percent papers of total volume do not have co-citations. Whereas those papers that contain citations are not more than one or two [59][60]. Furthermore, the recommendation can only recommend papers if they co-occurred at least once with another paper.

Continued Table 2.1: Critical Analysis of Different Approaches Containing Parameters, Contribution, and Claims

Ref	Parameters	Contribution	Claims
[38]	Bibliographic Coupling for finding related research articles.	They proposed that two research articles A1 and A2 are bibliographically coupled if they have cited the same research document.	In the course of experiments, Kessler discovered that if two documents have more than one common paper cited, then there is a close association between the two papers. The experimental evaluation of recommended approach is carried out by selecting 8,521 research articles from 795 different resources having 137,000 references.
[41]	They suggested technique by combining co-citation and bibliographic technique to find the relevant documents.	They performed a user study on 1,712 documents.	Their outcomes revealed that using both the bibliographic coupling of documents and their co-citations i.e. cited and citing documents, give better outcomes in comparison with using only the co-citations or bibliographic coupling.
[25]	They used Cue Phrases and Cue words as features	The system generated a better result on seen research articles and average results on unseen research articles.	With the change in the domain the new extensive list of cue-words and cue-phrases needs to be developed. Construction of 195 lexical matching rules and 14 parsing rules needs expert human-level knowledge. Ten (10) citation categories and their further division into 35 categories can cause confliction with each other.

Continued Table 2.1: Critical Analysis of Different Approaches Containing Parameters, Contribution, and Claims

Ref	Parameters	Contribution	Claims
[21]	Cue-phrases and cue-words	The system gained F-measure of 0.71	With the change in the domain the new extensive list of cue-words and cue-phrases needs to be developed and there is need to recreate list for every new dataset. Citations are annotated where manually selected words appear and not annotated where manually selected words dont appear e.g better.
[17]	In-text citations count, count based features, similarity-based features, context based features, position based features.	In-text citations count precision is 0.35 and it out-classed other features	It ignores important cue phrases which occur immediately before and after the in-text citations.
[16]	In-text citations count, similarity between abstracts.	The F-measure score is 0.65 and in-text citation count feature outperformed all other 11 features with the precision of 0.37.	It ignores important cure phrases which occur immediately before and after the in-text citations. The list of keywords needed to be updated for every new dataset.
[26]	They used 5 features for classification which includes title similarity, author overlap, references, abstract and cue terms	The system gained 0.68 precision value just by depending on freely available metadata.	They used metadata for important citations extraction that was not domain-specific.

2.10 Conclusion

This chapter has presented a detailed comprehensive review of state-of-the-art and contemporary approaches for the task of citation classification and research paper recommendations. The approaches proposed in the citation classification task has been classified into broad level categories such as: Citation Count, Automated Citation Classification, Quantitative Citations etc. Furthermore, this thesis has identified a closely relevant research area known as: research paper recommendations. Researchers in this research task identify relevant research papers for a focused research paper. Researchers have contributed in this area in the last two decades. Their best approaches and learning from the last 20 years have been applied in this research thesis for the task of citation classification which is more focused on finding a relevant research paper for the cited paper from the list of citing papers. The literature found for the task of Research paper recommender systems has been critically classified into following categories: Stereotyping, metadata, content based filtering, and hybrid approaches etc.

All papers from the literature were critically reviewed and this chapter has concluded the following gap in the literature of citation classification. Researchers have focused on metadata, in-text citation frequency, and cue-phrases to identify the relevance between citing and cited paper. However, a paper is actually represented by the content it contains. Authors of the paper explains their ideas to the scientific community using the content of the research papers. Such content has not been used by the research community working on the Citation Classification Task. This might be due to the fact that content was not openly available before the implementation of open access research publications which now have been adapted by many publishers.

Furthermore, it has been identified that for the task of Research paper recommender systems in the last 20 years, scientific community has applied content based approaches most of the time (55% of the time), to get best accuracies. It has a clear evidence that if the content of citing and cited papers may be utilized for the task of citation classification, the probability of achieving good accuracy is

high. Off course, such approach can only be applied when the complete content is available. Therefore, this thesis explore different possibilities of finding similarities between the content of citing and cited papers such as: (1) complete content, (2) section wise content with average weights, (3) section wise content with weighted average.

Chapter 3

An Innovative Approach to Identify Important Citations

This research proposes a comprehensive methodology to identify important citing papers for a cited paper by using the content of the pair (cited and citing paper). The content-based approach has been successfully applied in the last two decades for relevant research paper recommendations [27, 106]. Taking inspiration from this research, this study evaluates two types of content-based comparisons between the citing and cited paper pairs. In the literature, the documents entire content has been used to identify relevant papers. However, in this study, we not only adapt the standard content-based approach for the task of important citation identification, but also propose a novel approach termed as section-wise content-based similarity.

Figure: 3.1 depicts the complete methodology proposed in this study. In the first step of Figure 3.1, a benchmark dataset is selected which provides input for both approaches (content-based approach and section-wise content-based approach). The left side of **Figure 3.1** presents methodological steps of the section-wise content-based approach, which produces a similarity score for each section. The right side of **Figure 3.1** depicts the content-based approach, which produces an overall content similarity score between the two papers. Both approaches produce top-recommended papers. Thus, **Figure 3.1** depicts a state-of-the-art evaluation and comparison strategy. The following sections elaborate on each step of this process in detail.

3.1 Benchmark Datasets

The dataset selected to perform the experiments is the benchmark dataset developed by Valenzuela et al. [14]. This benchmark dataset is freely available online. The dataset belongs to the field of information systems and encompasses 465 annotated paper-citation pairs collected from the Association of Computational and Linguistics (ACL) anthology. The ACL anthology is a digital archive of research papers in computational linguistics and a citation network containing only those papers and citations which are published in the ACL anthology itself. **Table 3.1** provides a clear description of the dataset. The first column represents the two domain experts who annotated the dataset, denoted A and B in the Annotator column. The second column contains the source paper ID from the ACL anthology. The third column contains the IDs of the citing papers for the source paper.

The fourth column Follow-up contains the score assigned by the annotators (i.e. 0 for incidental and 1 for important paper-citation pairs). The dataset also contains Portable Document Format (PDF) files, which were converted into text files to extract the full content and sections of the papers.

TABLE 3.1: Benchmark Dataset

Annotator	Paper	Cited by	Follow-up
A	A00-1043	C00-2140	0
A	A00-1043	P02-1057	0
A	A97-1011	W09-1118	1
A	A97-1011	A00-2017	1
B	P05-1045	C10-1083	1
B	P05-1045	C10-1087	0
B	P05-1045	C10-1105	1
B	P05-1045	C10-1131	1

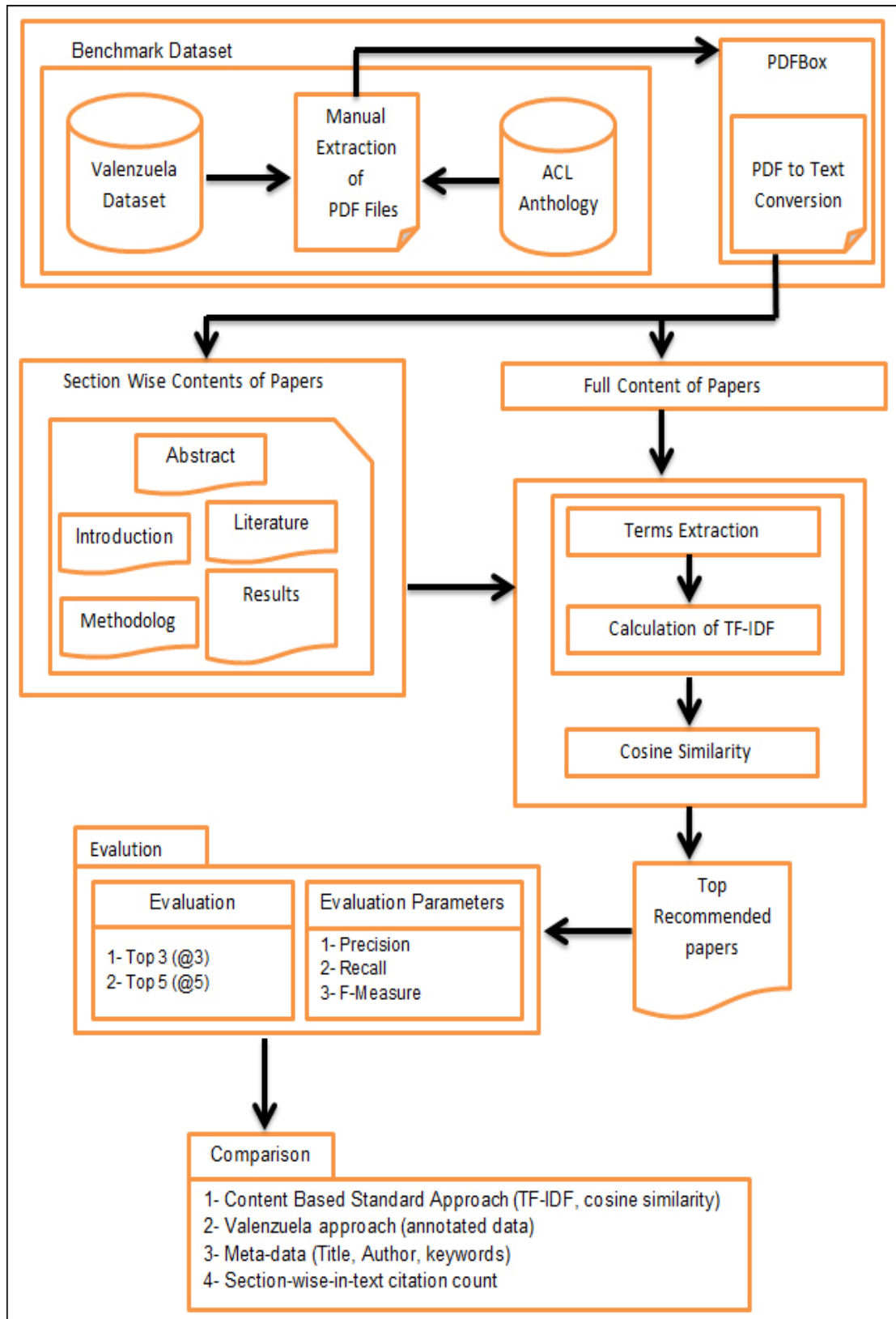


FIGURE 3.1: Proposed Methodology

3.2 PDF to Text Conversion

The PDF research articles are available on ACL anthology. These articles contain all the logical sections of a research paper. After this, we used the PDFBox tool that converts PDFs into text files. These files contain the combination of all sections (e.g. Abstract, Introduction, Literature Review, Methodology, Results) of research articles as shown in **Figure 3.2**. The PDF file in **Figure 3.2** contains different headings (i.e. Abstract, Introduction, etc.). These headings are treated as different logical sections. We extracted these logical sections manually from this PDF.

Then we generated five text folders. Each folder contains one section of all papers. For example, one folder of Abstract contains all the Abstract sections of all papers available in the benchmark dataset. Similarly, five folders were created for Abstract, Introduction, Literature Review, Methodology, Results as shown in **Figure 3.3** and **Figure 3.4**. All of the required five logical sections of research articles based on headings have been extracted on a similar pattern as explained above.

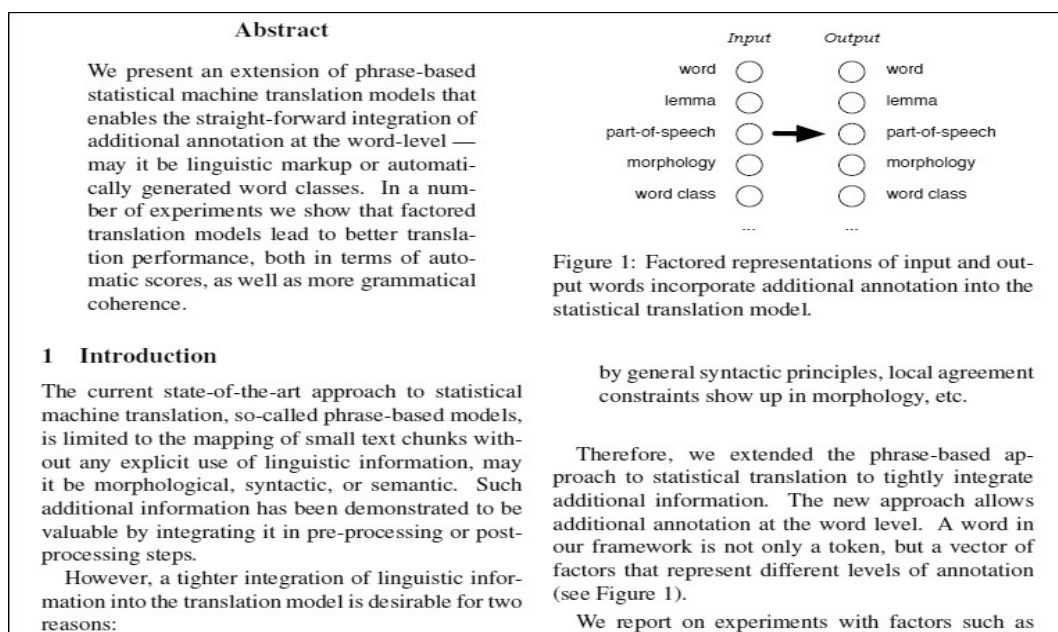


FIGURE 3.2: Logical Sections in PDF File

Abstract

We present an extension of phrase-based statistical machine translation models that enables the straightforward integration of additional annotation at the word-level may it be linguistic markup or automatically generated word classes. In a number of experiments we show that factored translation models lead to better translation performance, both in terms of automatic scores, as well as more grammatical coherence.

FIGURE 3.3: Logical Section Abstract

Introduction

The current state-of-the-art approach to statistical machine translation, so-called phrase-based models, is limited to the mapping of small text chunks without any explicit use of linguistic information, may it be morphological, syntactic, or semantic, such additional information has been demonstrated to be valuable by integrating it in pre-processing or post-processing steps. However, a tighter integration of linguistic information into the translation model is desirable for two reasons.

FIGURE 3.4: Logical Section Introduction

3.3 Pre-Processing

Pre-processing is required to reduce noise from the dataset and to improve the quality of achieved results that have been produced by the proposed methodology. In order to achieve such an outcome, the pre-processing phase is split into two parts.

In first part, we have used a process to remove the stop words from the contents of research articles and in the second part, we have utilized the stemming procedure to transform all remaining terms into their base words or root words for experimentation. The following is a step-by-step summary of these two approaches:

3.3.1 Stop-Words Removal

The words like the, in, an, a, and, as, at, be, by, for, is, are, which, from, that, has, he, is, its, of, on, to, was, were, will, with occur frequently in English language sentences. So, there is a need to remove these stop words from the contents of different logical sections to obtain rare terms from the content. For removing the stop words, we picked the Onix Text Retrieval Toolkit Stop word List for removing stop words.

3.3.2 Stemming

The stemming procedure converts terms/words from research articles' contents into their root words. For example, if two terms such as retrieval in the cited paper and retrieved in the citing paper are syntactically matched without the stemming process, these two words will not match. The Porter Stemming Algorithm is used to transform terms/words into their root words [107]. Words like retrieval, retrieved, and retrieves, for example, will be converted to their base word retriev. The full logical sections of the research articles were subjected to this algorithm.

3.4 Content-Based Approach

The content-based approach is the most dominant method for the task of relevant paper recommendation [27, 106]. In this study, we propose using the content-based approach to find important citing papers for each cited paper. The implementation steps for the content-based Similarity approach are shown on the right side of **Figure. 3.1**. This study employs Lucene indexing. The Apache Lucene application programming interface (API) is considered the standard software for term indexing. It is widely used by researchers for indexing and finding content similarities [15]. For the extraction of important terms, the papers full content are provided to the Apache Lucene API. Apache Lucene API indexes all terms within the content. Subsequently, the term frequency inverse document frequency (TF-IDF) scheme is used to extract important terms from the indexed terms. The term extractor

TF-IDF can be mathematically defined as given in equation 3.1. This equation is implemented for all citing and cited papers in the dataset. The basic idea of the TF-IDF technique is elaborated with the following example. For instance, the term T1 frequently occurs in document D1, but T1 is not found frequently in the other documents D2 to Dn. Thus, the conclusion is reached that term T1 is the most important term for document D1. Conversely, if any term T2 frequently exists in all documents D1 to Dn, it means that term T2 is not important at all to distinguish documents.

$$tf - Idf(t, d, D) = tf(t, d) * ldf(t, D) \quad (3.1)$$

The next step is to measure the similarity of the papers content. For this purpose, the cosine similarity technique is used. Equation 3.2 shows the mathematical model for cosine vector similarity computation. The important terms extracted from document D1 are presented as vector A and the important terms from document D2 as vector B. The rule of 30 [108] is applied to build the vector A and B, which means top 30% terms are selected to form the term vectors.

$$Content - Similarity = \frac{A.B}{|A||B|} \quad (3.2)$$

Cosine similarity was computed for each document compared to all other documents. The generated similarity scores lie between 0 and 1. All text files received a similarity score using the cosine technique. After calculating the cosine similarity scores, the results were sorted in descending order to obtain a ranked list of the top 3 (T@3) and top 5 (T@5) recommended papers.

3.5 Section-wise Content Similarity Approach

The implementation steps for the section-wise content similarity approach are shown on the left side of Figure 3.1. The Apache Lucene API was again used to index the terms. The similarities between corresponding sections of the papers were identified. To extract the important terms, the content of corresponding

paper sections were provided to the Apache Lucene API. Apache Lucene API indexed all terms in the section. Then, the TF-IDF technique was used to identify the most important indexed terms. The term extractor TF-IDF can be mathematically defined as given in equation 3.3. This equation was implemented for all corresponding sections of the citing and cited papers in the dataset.

$$tf - ldf(t, s, S) = tf(t, s) * ldf(t, S) \quad (3.3)$$

In equation 3.3, t represents the important terms in section s whereas s represents the section content of the cited paper and S the content of the corresponding section of all citing paper. The section-wise similarity technique is used to measure the similarity between corresponding sections. The mathematical model for section-wise similarity is given in equation 3.4. The vector $V1s$ refers to the extracted important terms for section s of cited paper $P1$, while the vector $V2s$ refers to the important terms for the corresponding section s of citing paper $P2$. The rule of 30 [108] is applied to build the vector $V1$ and $V2$, which means top 30% terms are selected to form the term vectors.

$$Section - wise_{similarity} = \frac{V1s.V2s}{|V1s||V2s|} \quad (3.4)$$

The section-wise similarity was computed for each section and was compared to all the corresponding sections. The generated section-wise similarity scores lies between 0 and 1. All text files received a similarity score using the section-wise similarity technique. After obtaining the similarity scores, the results were sorted in descending order to create a ranked list of the top 3 and top 5 recommended papers.

3.6 Section-Wise Research Paper Ranking

In every research article, there are five logical sections, which are as follows: an abstract, an introduction, a literature review section, a methodology section, and a section on the results. Each research article can be represented in the form

of vector. To further grasp it, let's look at an example. If we have two research articles, represented as D1 and D2 respectively. Then D1 is a cited article donated by vector A and D2 is a citing article donated by vector B. Now we want to compare vector A with vector B in order to determine how similar they are. To determine the degree of similarity between vector A and vector B, the cosine similarity technique is applied. Similarly, the cosine similarity of each logical section of a research article is calculated and compared to the cosine similarity of all research article sections included in the dataset. All research articles were given a similarity score, which was then sorted in descending order in the form of a rank list to show how similar they were. In the end, we combined similarity scores from different logical sections of research articles using average (mean) and weighted average methodologies, which will be elaborated in the following section.

Algorithm 1 Extraction of logical section for identification of important citations

Input: Research documents in PDF

Output: Ranking of logical sections of research documents

```

1:  Cited-Document ← Assign All cited documents
2:  Citing-Document ← Assign All citing document Against cited documents
3:  for x in cited-documents do
4:      for each section in x do    \ \ Read all sections of document x
5:          for each citing in citing-documents[x] do \ \ Citing document
6:              for each section in citing do
7:                  if x[section] == Citing-section
8:                      Text[I]=Cosine-Similarity(x[section],Citing section)
9:                      Score[section]←Cosine-Similarity(x[section],Citing section)
10:                 End for Loop
11:            End for Loop
12:        Sort-Score←Sort-in-Desending(Test[I])
13:        Average[section]←(Average[section]+F-measure(Sort-Score))
14:    End for Loop
15: End for Loop

```

The worst case analysis of the algorithm given above:

$$\begin{aligned}
 &= O(m+n) * T_{Ft, d_i} \log((m+n)/DFt) + O(m) * O(ms_i) * O(n) * O(ns_j) \\
 &= O(m+n) * T_{Ft, d_i} \log((m+n)/DFt) + O(m) * O(1) * O(n) * O(1) \\
 &= O(m+n) * T_{Ft, d_i} \log((m+n)/DFt) + O(m) * O(n) \\
 &= O(m+n) * T_{Ft, d_i} \log((m+n)/DFt) + O(m*n) \\
 &= \mathbf{O(m*n)}
 \end{aligned}$$

Where

m is no. of cited papers

n is no. of citing papers or $m < n$

ms_i is no. of sections in i^{th} cited paper.

ns_j is no. of section in j^{th} citing paper.

$m+n$ is the total no. of articles in the dataset.

TF_{t,d_i} is the frequency of a specific term t in document d_i

DF_t is the no. of documents which include a specific term t

The logical sections in research papers vary from 5 to 6. For example, each of the research articles given in the dataset used in this research has 5 logical sections and it is a constant number of computations.

3.7 Techniques for Combining Similarity Scores of Logical Section

The similarity scores of corresponding logical sections were computed for example Abstract of citing paper and Abstract of cited paper. Similarly, Introduction Vs Introduction, Literature Review Vs Literature Review, Methodology Vs Methodology and Results Vs Results. We have further joined sections in all possible combinations illustrated below.

- 1) Abstract + Introduction Vs Abstract + Introduction
- 2) Abstract + Literature Review Vs Abstract + Literature Review
- 3) Abstract + Methodology Vs Abstract + Methodology
- 4) Abstract + Results Vs Abstract + Results
- 5) Introduction + Literature Review Vs Introduction + Literature Review
- 6) Introduction + Methodology Vs Introduction + Methodology

- 7) Introduction + Results Vs Introduction + Results
- 8) Literature Review + Methodology Vs Literature Review + Methodology
- 9) Literature Review + Results Vs Literature Review + Results
- 10) Methodology + Results Vs Methodology + Results
- 11) Abstract + Introduction + Literature Review Vs Abstract + Introduction + Literature Review
- 12) Abstract + Introduction + Methodology Vs Abstract + Introduction + Methodology.
- 13) Abstract + Introduction + Results Vs Abstract + Introduction + Results
- 14) Abstract + Literature Review + Methodology Vs Abstract + Literature Review + Methodology
- 15) Abstract + Literature Review + Results Vs Abstract + Literature Review+ Results
- 16) Abstract + Methodology + Results Vs Abstract + Methodology + Results
- 17) Introduction + Literature Review + Methodology Vs Introduction + Literature Review + Methodology
- 18) Introduction + Literature Review + Results Vs Introduction + Literature Review + Results
- 19) Literature Review + Methodology + Results Vs Literature Review + Methodology+ Results
- 20) Abstract + Introduction + Literature Review + Methodology vs Abstract+ Introduction + Literature Review + Methodology
- 21) Abstract + Introduction + Literature Review + Results vs Abstract + Introduction+ Literature Review + Results
- 22) Introduction + Literature Review + Methodology + Results vs Introduction+ Literature Review + Methodology + Results
- 23) Abstract + Introduction + Literature Review + Methodology + Results vs Abstract + Introduction + Literature Review + Methodology + Results

Every possible combination of logical sections that are mentioned above has been combined with one another using the Average and Weighted Average techniques, which are briefly discussed in the following sections 3.7.1, and 3.7.2.

3.7.1 Average

An average, also known as an arithmetic mean, is calculated by adding the number of values in a cluster and dividing them with the total number of values in the cluster. This research utilizes the average technique to combine the similarity scores of different logical sections of research articles, as discussed in section 3.6 of this chapter. The mathematical representation of arithmetic mean is given by the formula shown in equation 3.5.

TABLE 3.2: Combination of Two Sections by Using Average

C	D	E	F	G	H	I
Cited	Citing	Abstract	Methodology	Result	Average	Average
A97	P98	0.1184	0.2701	0.2312	0.1942	0.1748
A97	A00	0.1684	0.2338	0.1327	0.2011	0.1505
A97	E12	0.0606	0.1869	0.1765	0.1238	0.1185
A97	P01	0.0260	0.2266	0.1996	0.1263	0.1128
A97	W06	0.0446	0.2232	0.1362	0.1339	0.0904
A97	W04	0.0451	0.2019	0.1178	0.1235	0.0815
A97	W09	0.0227	0.1597	0.1255	0.0192	0.0741
A97	C00	0.0164	0.1525	0.7370	0.0845	0.0450

$$Average = \frac{\sum_{i=1}^n X_i}{n} \quad (3.5)$$

Where $X_i = X_1 + X_2 + X_3 \dots X_n$ denotes the similarity scores of different sections of research articles, and n denotes the total number of sections.

Example: To combine the similarity scores of two or more sections, we used the average technique. **Table 3.2** shows how the similarity scores from the Abstract and Methodology sections are combined in column H, and the similarity scores from the Abstract and Results sections are combined in column I using the average

formula. Similarly, the average technique is used to combine different sections of research articles, as discussed in section 3.6.

3.7.2 Weighted Average

Another experiment was conducted in which the logical sections were given different weights based on their importance. The weighted average technique was used in this research to combine the similarity scores of different logical sections, as discussed in section 3.6 of this chapter.

The formula shown in equation 3.6 was used in this technique. Every text file with logical sections is given a weight based on the number of words counted in the file using an online wordcounter.

$$WeightedAverage = \frac{(x_1 * \frac{1}{count(s_1)})(x_2 * \frac{1}{count(s_2)}) + (x_n * \frac{1}{count(s_n)})}{Sum(\frac{1}{count(s_1)}, \frac{1}{count(s_2)}, \dots, \frac{1}{count(s_n)})} \quad (3.6)$$

Where x_1 , x_2 , and x_n represent the similarity scores of the Abstract, Methodology, and Result sections of research articles, respectively, and $count(s_1)$, $count(s_2)$, ..., $count(s_n)$ represents the total number of words in the Abstract, Methodology, and Result sections.

Example To combine the similarity scores of different logical sections of research articles, the weighted average technique is used. Abstract section similarity score and weights (i.e. $count(abstract)$) are presented in Column F and Column G, respectively, as shown in **Figure 3.6**.

The score and weights for the Methodology section (i.e. $count(methodology)$) are displayed in columns I and J, respectively. In the last column (i.e. Column L) of **Table 3.3**. The weighted average of the Abstract and Methodology sections is calculated using the formula shown in equation 3.6. In this study, the weighted average formula is used to combine the similarity scores of various sections of research articles discussed in section 3.7.

TABLE 3.3: Combination of Two Sections by Using Weighted Average

D	E	F	G	H	I	J	K	L
Cited	Citing	Abstract	Cnt	1/G	Method	Cnt	1/J	WAvg
A97	P98	0.1184	84	0.0119	0.2701	748	0.0013	0.0286
A97	A00	0.1684	59	0.0169	0.2338	1078	0.0009	0.0149
A97	E12	0.0606	124	0.0080	0.1869	1689	0.0005	0.0132
A97	P01	0.0260	179	0.0055	0.2266	426	0.0023	0.0671
A97	W06	0.0446	110	0.0090	0.2232	1179	0.0008	0.0194
A97	W04	0.0451	71	0.0140	0.2019	1534	0.0006	0.0095
A97	W09	0.0227	84	0.0119	0.1597	410	0.0024	0.0274
A97	C00	0.0164	95	0.0105	0.1525	630	0.0015	0.0201

3.8 Evaluation Parameters

The proposed approach was evaluated using standard evaluation parameters used by state-of-the-art approaches, namely by (1) Valenzuela et al. [14], (2) Qayyum and Afzal [24], and Nazir et al. [26]. The evaluation parameters used by state-of-the-art approaches are precision, recall, and F-measure. The definition of each parameter is given below:

The formula to calculate the precision is shown in equation 3.7. The dataset contains citations classified as important and non-important. Precision in identifying important citation using the proposed technique is defined as the ratio of citations correctly classified as important citations to the total number of citations classified by the technique as important citations.

$$Precision = \frac{\text{Correctly classified as important citations}}{\text{Total number of classified as important citations}} \quad (3.7)$$

The formula to calculate the recall is depicted in equation 3.8. Recall in identifying important citations using the proposed technique is defined as the ratio of citations correctly classified as important citations to the total number of citations which are in actual fact important citations.

$$Recall = \frac{\text{correctly classified as important citations}}{\text{Actual "important citation"}} \quad (3.8)$$

The F-measure is the harmonic mean of precision and recall and is calculated as shown in equation 3.9.

$$F - measure = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad (3.9)$$

The precision, recall, and F-measure for each cited paper were calculated against all of its citing papers by considering the classification presented in Valenzuela et al. [14]. Subsequently, the average precision, recall, and F-measure were calculated for the full dataset. Then, the precision, recall, and F-measures for the state-of-the-art approaches were taken from the original published papers [14, 23, 25], all of which worked on the same dataset.

3.9 Conclusion

Therefore, this research applies two types of content-based filtering methods. Firstly, the complete content of both the citing and cited paper is used to categorize the citing paper as an important/non-important citation for the cited paper. Furthermore, a novel section-based approach to citation classification is proposed. This chapter has proposed a comprehensive methodology adapted from other domains where content based filtering approach has been successfully employed. This chapter presented comprehensive methodology to exploit the complete content, section-wise content, and section-weights. Furthermore, the evaluation parameters have been discussed and the baseline approaches have been highlighted. The results will be discussed in the chapter 4.

Chapter 4

Results and Findings

This thesis has proposed two approaches for the task of the citation classification. This chapter presents the analysis of results for both of the proposed approaches. Furthermore, comparisons have been made with the state-of-the-art recent approaches which have worked on the same dataset. The objectives of this chapter are twofold: Firstly, we sought to identify the applicability of using content to identify important citations. This refers to both approaches, i.e. using the full content and evaluating individual sections. Secondly, we sought to compare the results with existing state-of-the-art approaches proposed by Qayyum and Afzal [24], Valenzuela et al. [14], and Nazir et al [26]. Section 4.1 presents the results and evaluation of the two proposed approaches, whereas Section 4.2 compares the best results from the proposed approach with contemporary approaches.

4.1 Results and Evaluation of the Proposed Approaches

Our first approach was to adapt a content-based filtering technique for citation classification. The second approach was to apply the same content-based technique to individual paper sections, namely the abstract, introduction, literature review, methodology, and results sections. Subsequently, we discuss which section plays the best role in classifying citations into two classes, important and non-important. Specifically, the following six similarity-based rankings were computed.

1. Full content similarity-based ranking
2. Abstract-section similarity-based ranking
3. Introduction-section similarity-based ranking
4. Literature section similarity-based ranking
5. Methodology-section similarity-based ranking
6. Results-section similarity-based ranking

In the next section, the section-based similarity rankings (No. 2 to No. 6 above) will be compared to the full-content-based ranking (No. 1 above).

4.1.1 Evaluation of Proposed Six Rankings

This section presents the results for all six similarity rankings proposed in this research and listed above. The first ranking is a full-content similarity-based ranking.

4.1.1.1 Full Content Similarity-based Ranking

In this ranking, the full content of the cited document is taken and compared to the full content of the citing documents in the list. Similarity scores are calculated for comparison purposes. Afterwards, the similarity scores are sorted in descending order to rank the top 3 as well as top 5 citing documents for each cited paper. Then precision, recall, and F-measure scores are calculated. In the end, a cumulative F-measure for the top 3 and top 5 documents is calculated and compared to the F-measure for the top 3 and top 5 documents identified using the other similarity rankings listed above. The cumulative F-measure for the full-content similarity-based ranking was 0.63 for the top 3 documents and 0.65 for the top 5 documents, respectively. This is a significant result obtained by solely examining the content of research papers. It will be compared with existing state-of-the-art approaches in the next section.

4.1.1.2 Abstract-section Similarity-based Ranking

The second ranking was produced by computing the similarity between the abstracts of the cited and citing documents. The cumulative F-measures for the abstract-section similarity-based ranking were 0.70 for the top 3 documents and 0.69 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. Comparisons between the abstract section similarity and full-content similarity are shown in **Figure 4.1** and 4.2 for the top 3 and top 5 ranked documents, respectively. The cited paper number is listed on the x-axis and the F-measure score on the y-axis. The red line shows the F-measure for each cited paper using the full-content approach, whereas the blue line shows the abstract section-based F-measure. For most of the cited papers, the red line and blue line follow the same path, meaning that both approaches produced the same results in these cases. It is also clear from **Figure 4.1** and **4.2** that when the results of abstract-based similarity and content-based similarity differ, abstract-based similarity produces more accurate results. This result clearly shows that abstract-section similarity-based ranking outperforms full-content similarity-based ranking. This is because the abstract is a concisely written paper section of just a few hundred words in which the author has to explain the whole idea of the research paper, including motivation, research gap, state-of-the-art, research question, methodology, results, and comparisons.

Thus, the abstract has more descriptive power regarding the context and contribution of a research paper. In contrast, the full content of a paper encompasses many different sections, including the introduction, literature review, etc., which might not be feasible to compare and might not deliver such strong results. Research papers abstracts are normally available for free for both citing papers as well as cited papers.

4.1.1.3 Introduction-Section Similarity-Based Ranking

The third proposed ranking involves the introduction sections of the cited and citing papers.

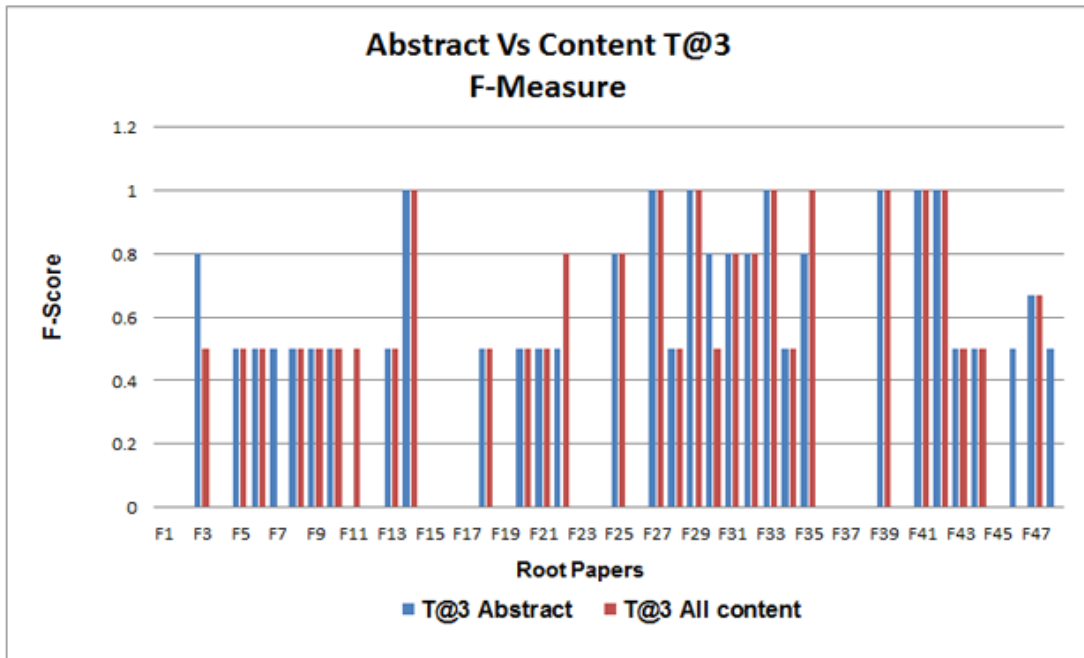


FIGURE 4.1: Abstract vs. Full Content Top 3

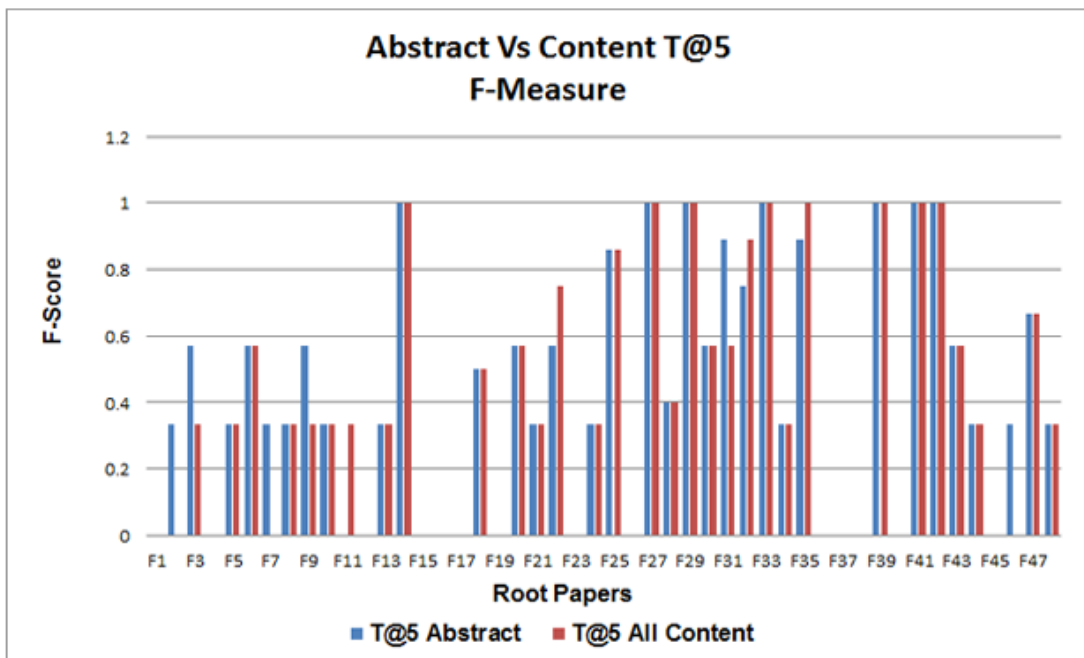


FIGURE 4.2: Abstract vs. full Content Top 5

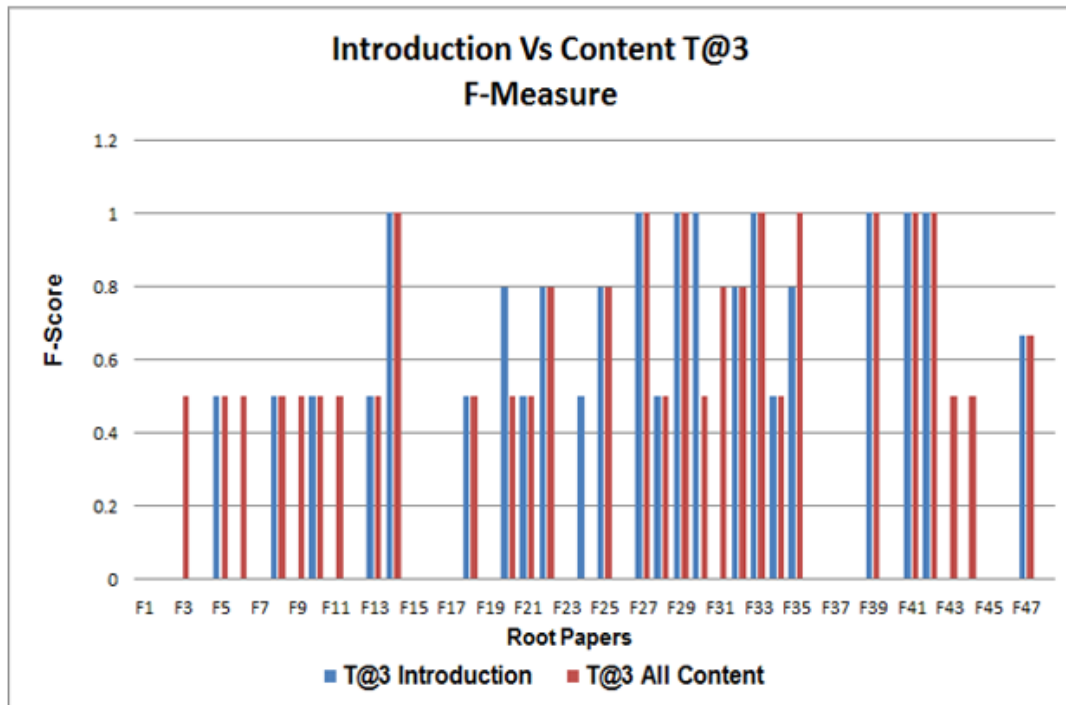


FIGURE 4.3: Introduction vs. Full Content Top 3

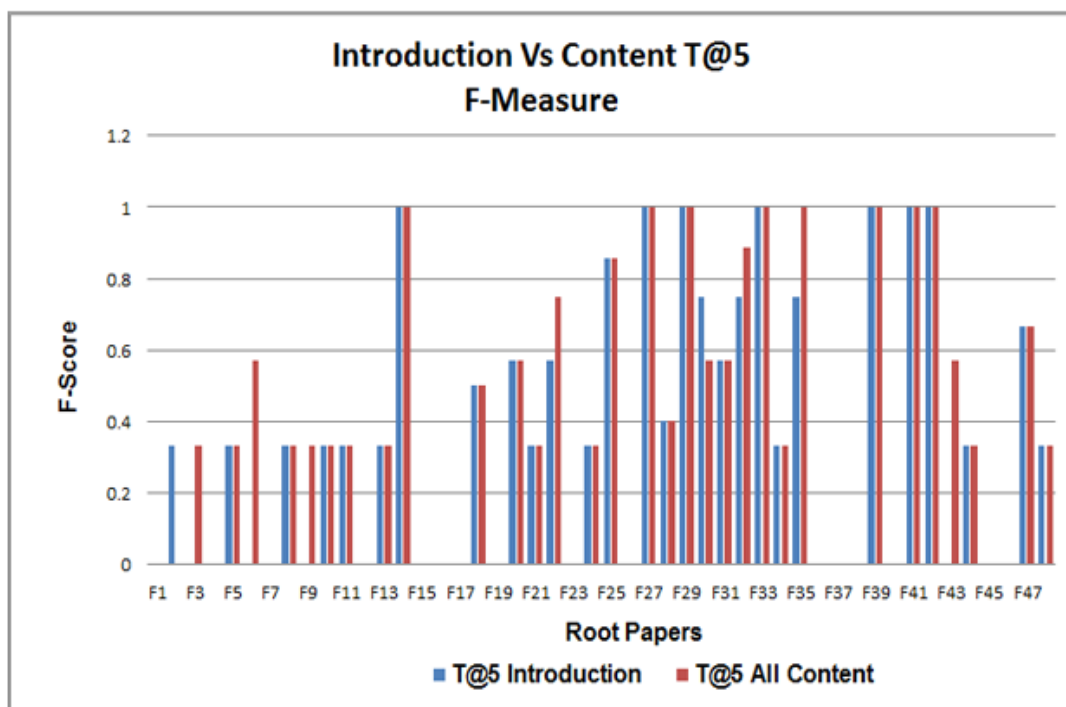


FIGURE 4.4: Introduction vs. Full Content Top 5

The cumulative F-measures for the introduction-section similarity-based ranking were 0.57 for the top 3 documents and 0.59 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively.

This result clearly shows that full-content similarity-based ranking outperforms introduction-section similarity-based ranking. The results are shown in **Figure 4.3** and **4.4**. An in-depth analysis of the terms identified as important for the introduction sections of some of the randomly selected papers illustrates the reason for such results. The introduction section is usually an extended version of the abstract. The general flow of the introduction section is as follows: (1) background of the problem, (2) existing state-of-the-art approaches, (3) research gap, (4) methodology, and (5) results and comparisons. Accordingly, most of the content in papers introduction sections tends to be very similar, leading towards the citing paper to be considered an important citation for the cited paper. Moreover, this section is typically not very long.

4.1.1.4 Literature Section Similarity-based Ranking

The fourth ranking was produced by examining the content of the literature review sections of the cited and citing paper pairs. The cumulative F-measures for the literature section similarity-based ranking were 0.59 for the top 3 documents and 0.62 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. The results are shown in **Figure 4.5** and **4.6**.

This result clearly shows that full-content similarity-based ranking outperforms literature-section similarity-based ranking. This is because the literature review section contains very generic terms to explain others work. Every author has a unique way of writing the literature review section by explaining existing approaches in the respective research areas and critical analyzing the literature. Therefore, the literature review section is not significant for identifying contextual similarities between cited and citing pairs.

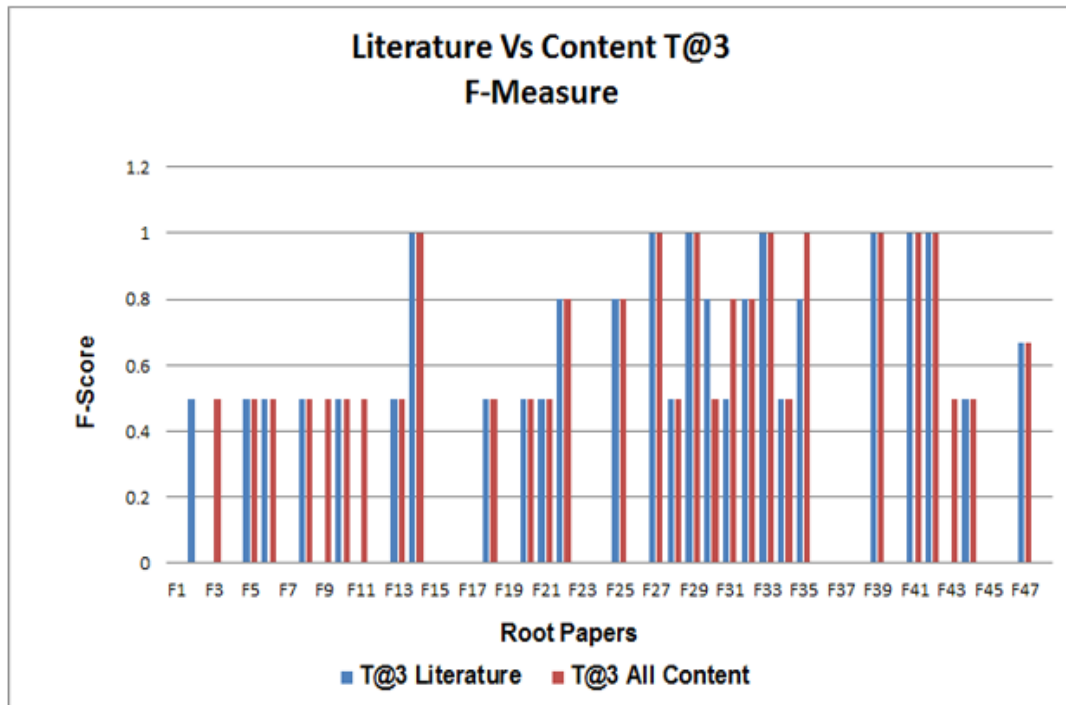


FIGURE 4.5: Literature vs. Full Content Top 3

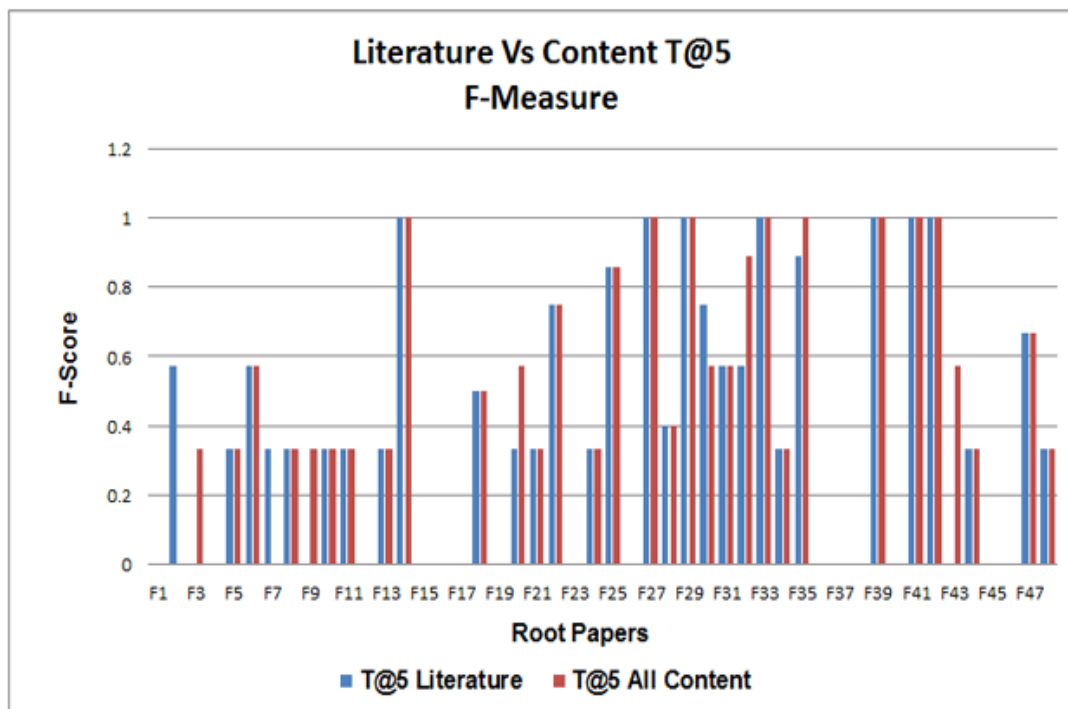


FIGURE 4.6: Literature vs. Full Content Top 5

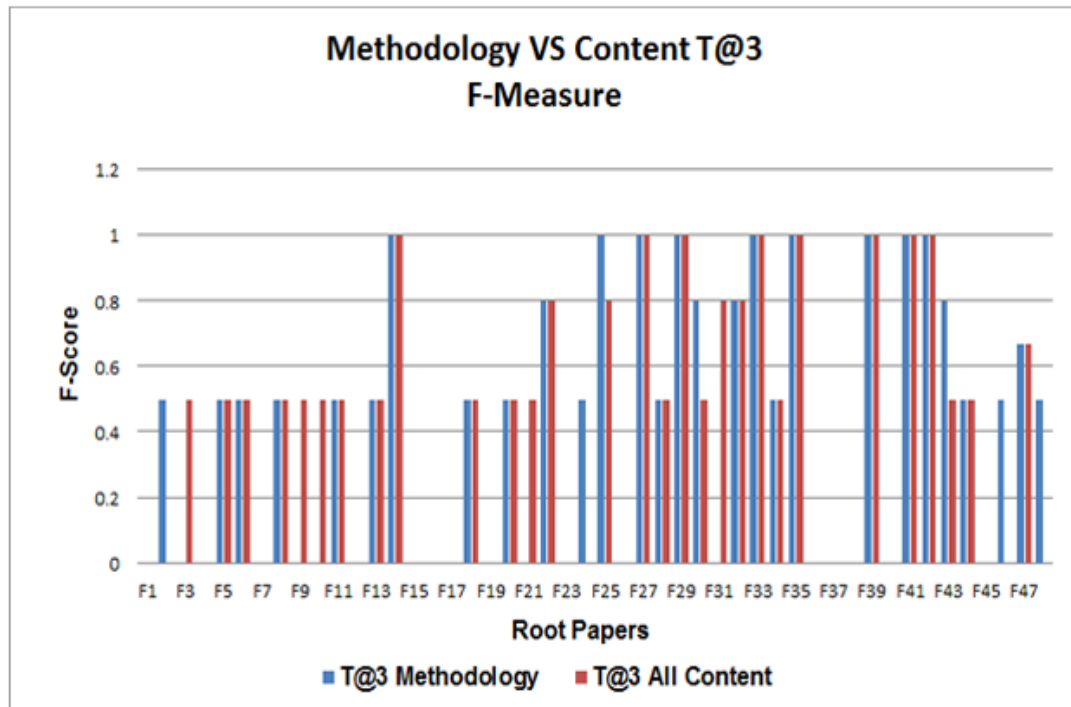


FIGURE 4.7: Methodology vs. Full Content Top 3

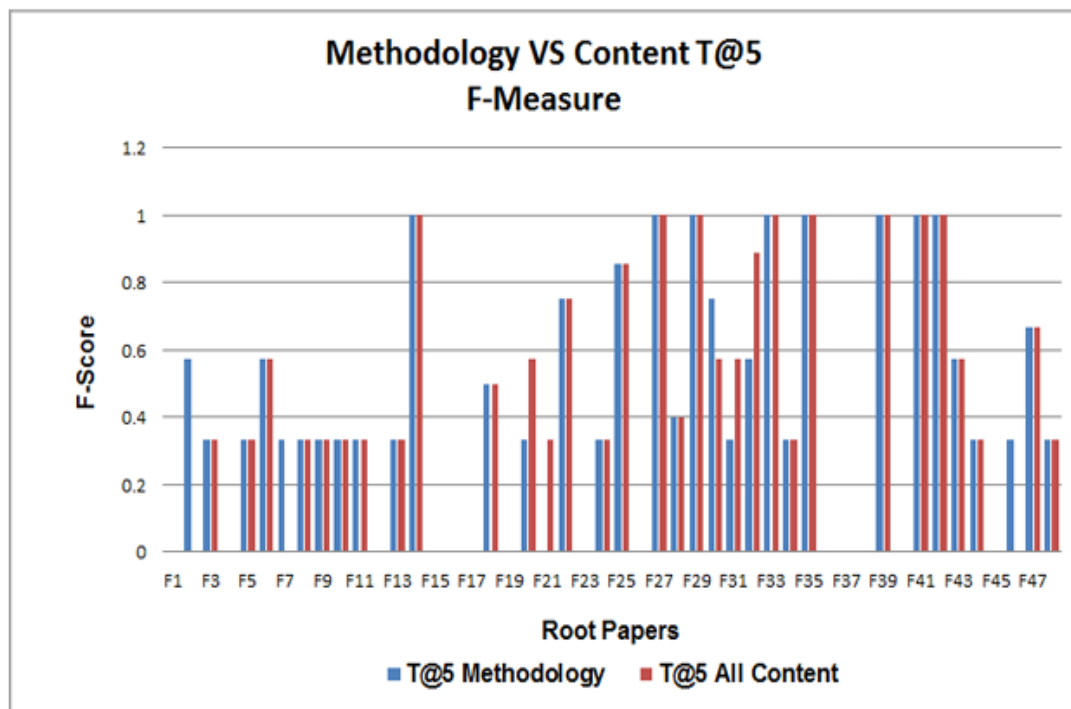


FIGURE 4.8: Methodology vs. Full Content Top 5

4.1.1.5 Methodology-Section Similarity-Based Ranking

The fifth ranking was achieved by examining the content of the methodology sections of both cited and citing papers. The cumulative F-measures for the methodology-section similarity-based ranking were 0.72 for the top 3 and 0.66 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. The results are represented in **Figure. 4.7** and **4.8**. This results clearly show that methodology-section similarity-based ranking outperforms full-content similarity-based ranking. The results demonstrate the expressive power of the cited papers methodology section to identify important citations from the list of citing papers. The methodology section presents the study conceptualization of both papers, which involves the use of similar domain-related terms. Nevertheless, while the results for the methodology section were good, the results for the abstract were even better.

4.1.1.6 Results-Section Similarity-Based Ranking

The sixth and final ranking proposed in this research is depicted in **Figure. 4.9** and **4.10**. This ranking was achieved by comparing the content of the results sections of both papers. The cumulative F-measures for the results-section similarity-based ranking were 0.64 for the top 3 and 0.63 for the top 5 documents, respectively.

Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. This result clearly shows that results-section similarity-based ranking and full-content similarity-based ranking are approximately equal. The results section is also an important section of a research paper, as it reports the important findings of both cited and citing papers. The results section similarity was found to be significant when: (1) both papers address the same topic and use a common vocabulary of terms for the specific domain, (2) both papers use the same dataset, (3) both papers apply similar evaluation metrics, and (4) both papers compare their results with the same/similar research papers.

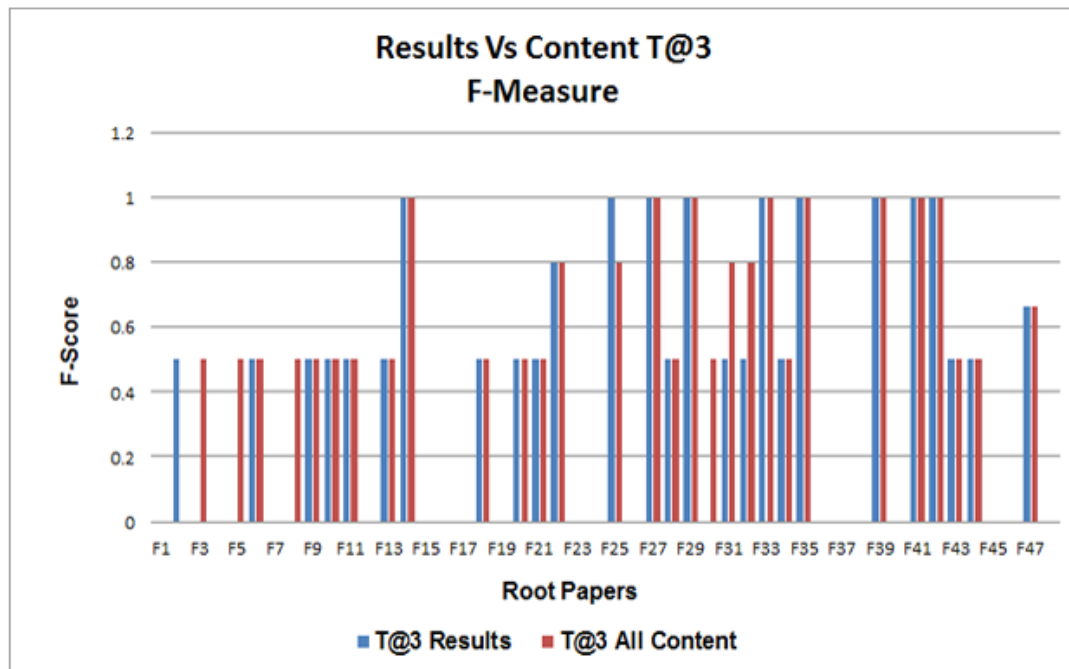


FIGURE 4.9: Results vs. Full Content Top 3

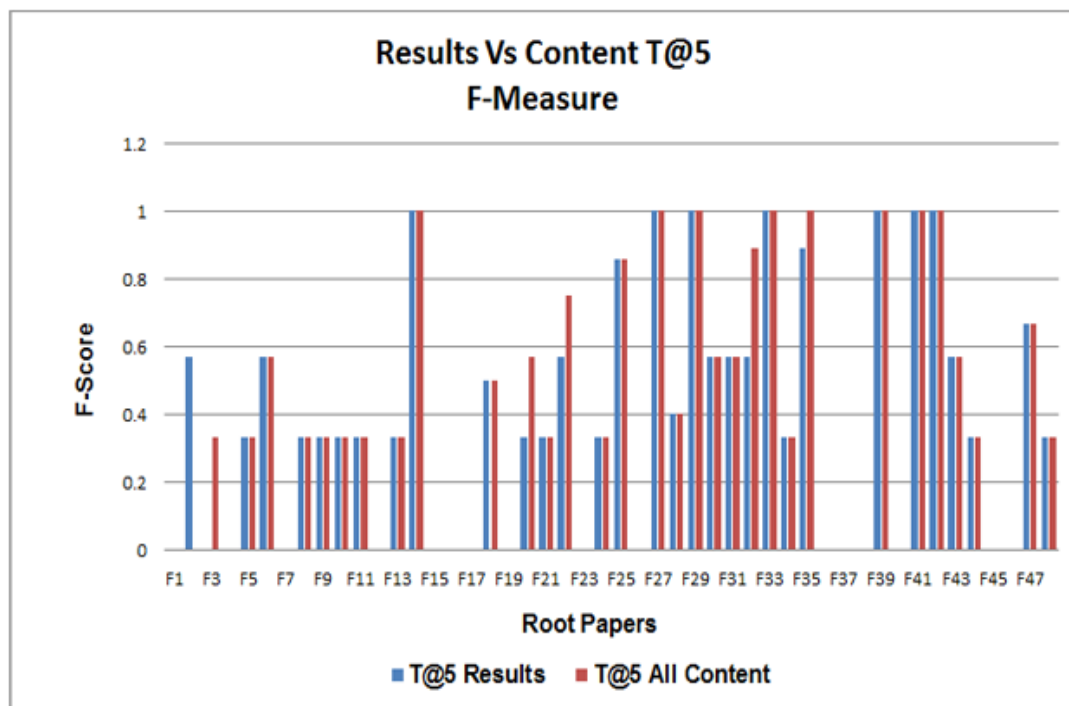


FIGURE 4.10: Results vs. Full Content Top 5

4.1.1.7 Conclusion of Six Similarity-Based Ranking

This section has reported the results of all six proposed rankings. The full-content-based ranking was adapted from the domain of identifying relevant research papers [28, 109]. Furthermore, this research proposed the novel approach of section-based similarity. Five further rankings could be calculated when applying this proposed approach. The content of the five major sections of research papers, namely abstract, introduction, literature review, methodology, and results, were systematically compared.

The results indicate that examining content alone makes it possible to identify important citations for the cited paper from the list of citing papers. Furthermore, the abstract, methodology, and results sections achieved similar or better results compared to the full content of research papers. Specifically, the abstract section outperformed the full content of research papers. Therefore, we conclude that papers abstracts should be used to compute content-based similarity for two reasons: (1) better performance compared to the full-content approach, and (2) abstracts are normally freely available.

Figure 4.1-4.10 present the results of the five proposed rankings based on comparing individual paper sections. **Figure 4.11** compares the results of six proposed ranking approaches. The complete content achieved the highest precision of 0.68; however, the abstract alone achieved a very close precision score of 0.66. Furthermore, the abstract alone was able to achieve a very high recall value of 0.94, second only to the methodology section. Due to this high recall, the abstract outperformed all other approaches in terms of F-measures. Therefore, out of the five proposed section-wise rankings, the abstract section was selected to be compared to existing state-of-the-art approaches in the following sections.

During the experimentation, two single similarity-based rankings, Introduction-section and Literature-section, have produced lower F-measure values than the Full-Content similarity-based ranking. These sections normally carry larger portion of the paper than the other sections. Therefore, authors of papers have to write the broad level introduction and background of the research area. Due

to large coverage of terms, contents written in these sections remained dissimilar even when two papers were relevant. As a result, these two rankings (Introduction-section and Literature-section) were skipped for further testing. For subsequent experimentation, we simply utilized three rankings (Abstract-section, Methodology-section, and Result-section) in all conceivable combinations. Table 4.1 shows all of the possible combinations of Abstract-section, methodology-section, and Result section. To create double and triple similarity-based ranking parameters, each parameter is joined with all other parameters in all possible ways using the Average and Weighted-average approaches.

TABLE 4.1: All Possible Combinations

Average	
1	Abstract+Methodology Vs Abstract+Methodology
2	Abstract+Results Vs Abstract+Results
3	Methodology+Results Vs Methodology+Results
4	Abstract+Methodology+Results Vs Abstract+Methodology+Results
Weighted Average	
5	Abstract+Methodology Vs Abstract+Methodology
6	Abstract+Results Vs Abstract+Results
7	Methodology+Results Vs Methodology+Results
8	Abstract+Methodology+Results Vs Abstract+Methodology+Results

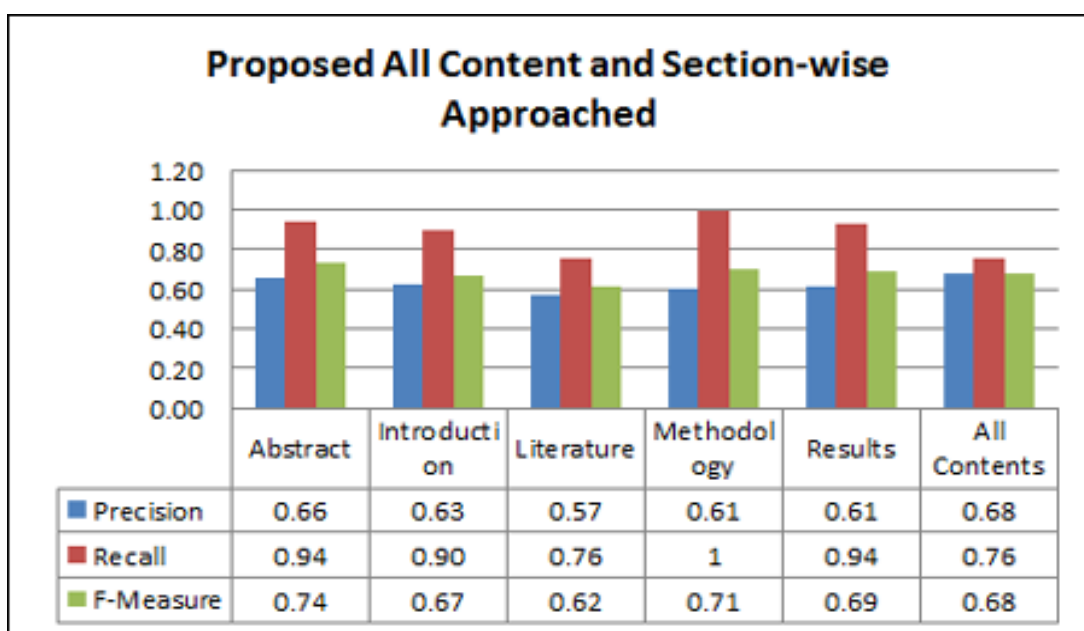


FIGURE 4.11: Comparison of the Six Proposed Rankings

4.2 Combination of Double Parameters using Average Results

4.2.1 AVR(Abtract-Methodology-Sections Similarity-Based Ranking)

In this ranking technique, the contents of abstract and methodological sections of cited document was compared with the contents of corresponding logical sections (i-e abstract and methodology) of all citing documents. Then the similarity score was computed between the logical sections of cited and citing documents. The combination of similarity scores of two logical sections (abstract and methodology) was done by utilizing the average formula depicted in equation 3.5. The computed score was then arranged in descending order. This score was utilized to rank the top 3 and top 5 documents. The F-measure for each cited document was computed individually. The cumulative F-measures for the AVR(abstract-methodology-section similarity-based ranking) were 0.74 for the top 3 documents and 0.70 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. Comparisons between the AVR(abstract-methodology-section similarity-based ranking) and full-content similarity are shown in **Figure 4.12** and **4.13** for the top 3 and top 5 ranked documents, respectively.

The cited paper number is listed on the x-axis and the F-measure score on the y-axis. The red line shows the F-measure for each cited paper using the full-content approach, whereas the blue line shows the AVR(abstract-methodology-section similarity-based ranking) F-measure. For most of the cited papers, the red line and blue line follow the same path, meaning that both approaches produced the same results in these cases.

It is also clear from **Figure 4.12** and **4.13** that when the results of AVR(abstract-methodology-section similarity-based ranking) and content-based similarity differ, AVR(abstract-methodology-section similarity-based ranking) produces more accurate results.

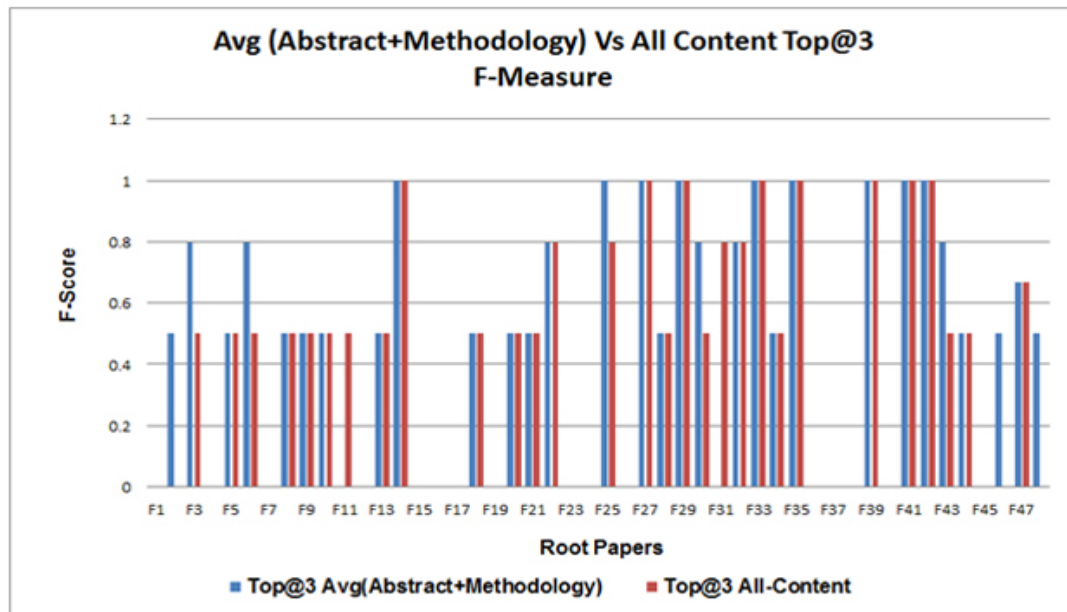


FIGURE 4.12: Avg(Abtract+Methodology) Vs All Content Top 3

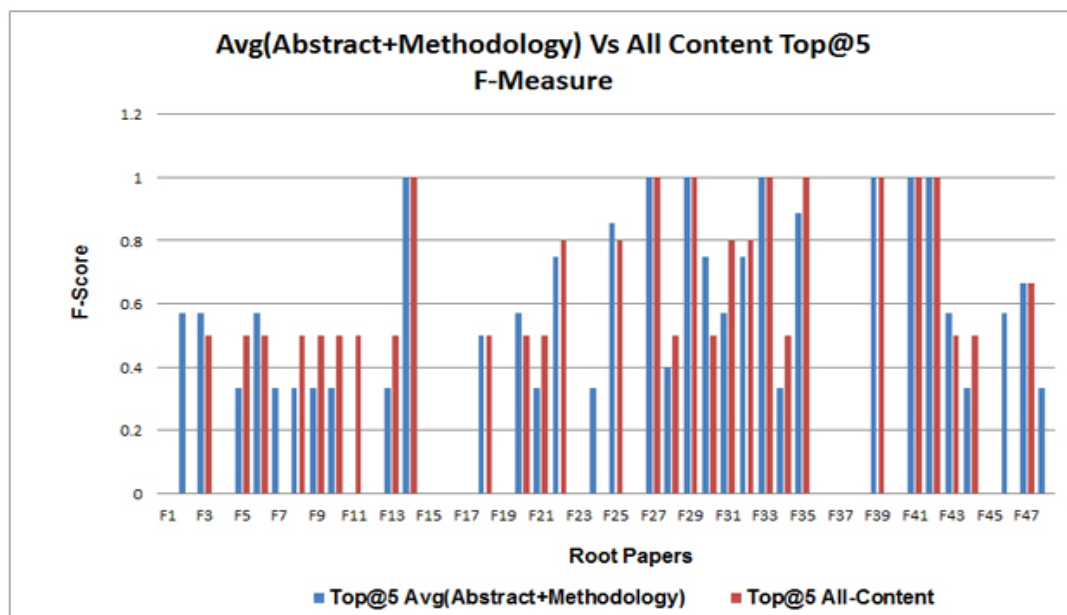


FIGURE 4.13: Avg (abstract + Methodology) Vs All Content Top 5

4.2.2 AVR(Abtract-Results-Sections Similarity-Based Ranking)

In this ranking technique, the contents of abstract and results sections of cited document was compared with the contents of corresponding logical sections (i-e

abstract and results) of all citing documents. Then the similarity score was computed between the logical sections of cited and citing documents. The combination of similarity scores of two logical sections (abstract and results) was done by utilizing the average formula depicted in equation 3.5. The computed score was then arranged in descending order. This score was utilized to rank the top 3 and top 5 documents. The F-measure for each cited document was computed individually.

The cumulative F-measures for the AVR(abstract-results-section similarity-based ranking) were 0.63 for the top 3 documents and 0.67 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. Comparisons between the AVR(abstract-results-section similarity-based ranking) and full-content similarity are shown in **Figure 4.14** and **4.15** for the top 3 and top 5 ranked documents, respectively.

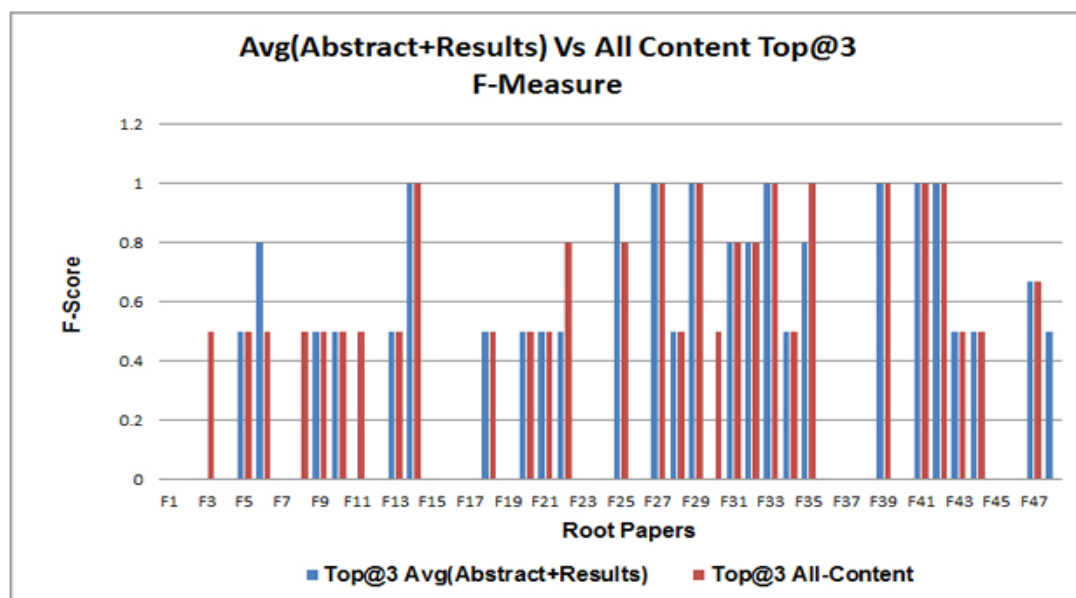


FIGURE 4.14: Avg(Abtract+Results) Vs All Content Top 3

The cited paper number is listed on the x-axis and the F-measure score on the y-axis. The red line shows the F-measure for each cited paper using the full-content approach, whereas the blue line shows the AVR(abstract-results-section similarity-based ranking) F-measure. For most of the cited papers, the red line and blue line

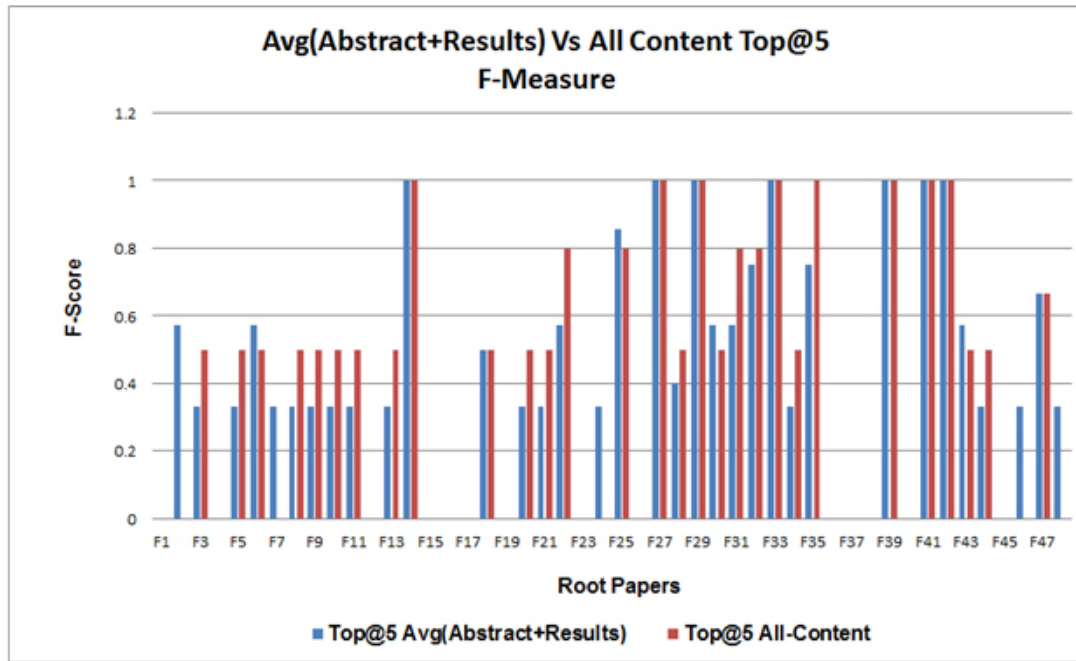


FIGURE 4.15: Avg(Abstract+Results) Vs All Content Top 5

follow the same path, meaning that both approaches produced the same results in these cases. It is also clear from **Figure 4.14** and **4.15** that the results of top 3 AVR(abstact-results-section similarity-based ranking) and content-based similarity are equal, but the results of top 5 AVR(abstact-results-section similarity-based ranking) outperforms the full-content similarity based ranking.

4.2.3 AVR(Methodology-Results-Sections Similarity-Based Ranking)

In this ranking technique, the contents of abstract and results sections of cited document was compared with the contents of corresponding logical sections (i.e methodology and results) of all citing documents. Then the similarity score was computed between the logical sections of cited and citing documents. The combination of similarity scores of two logical sections (methodology and results) was done by utilizing the average formula depicted in equation 3.5. The computed score was then arranged in descending order. This score was utilized to rank the top 3 and top 5 documents. The F-measure for each cited document was computed individually. The cumulative F-measures for the AVR(methodology-results-section

similarity-based ranking) were 0.67 for the top 3 documents and 0.67 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. Comparisons between the AVR(methodology-results-section similarity-based ranking) and full-content similarity are shown in **Figure 4.16** and **4.17** for the top 3 and top 5 ranked documents, respectively.

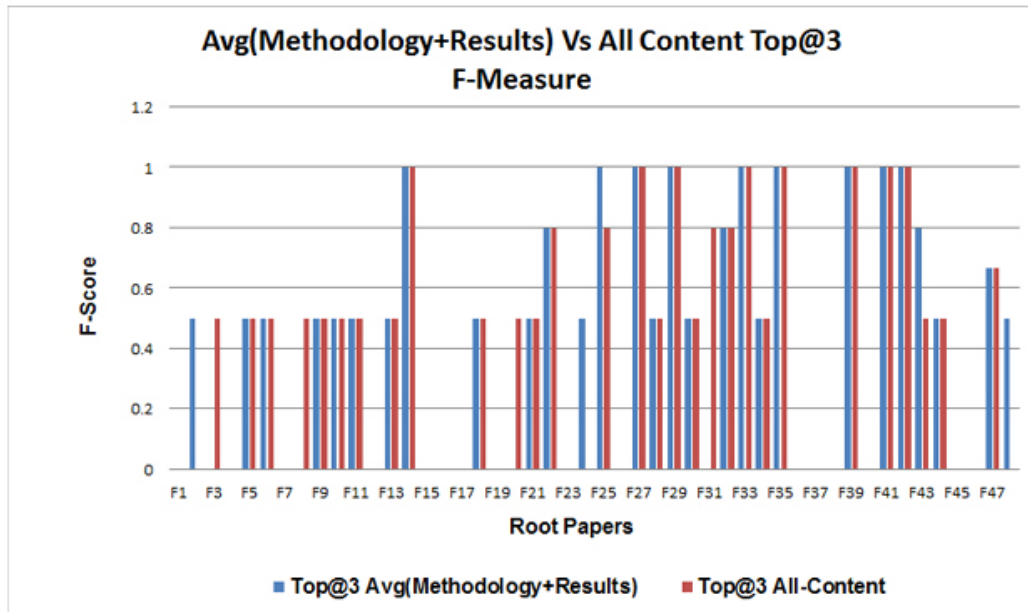


FIGURE 4.16: Avg(Methodology+Results) Vs All Content Top 3

The cited paper number is listed on the x-axis and the F-measure score on the y-axis. The red line shows the F-measure for each cited paper using the full-content approach, whereas the blue line shows the AVR(methodology-results-section similarity-based ranking) F-measure.

For most of the cited papers, the red line and blue line follow the same path, meaning that both approaches produced the same results in these cases. It is also clear from **Figure 4.16** and **4.17** that the results of top 3 AVR(methodology-results-section similarity-based ranking) outperforms against the full content-based similarity, but the results of top 5 AVR(methodology-results-section similarity-based ranking) and full-content similarity based ranking are approximately equal.

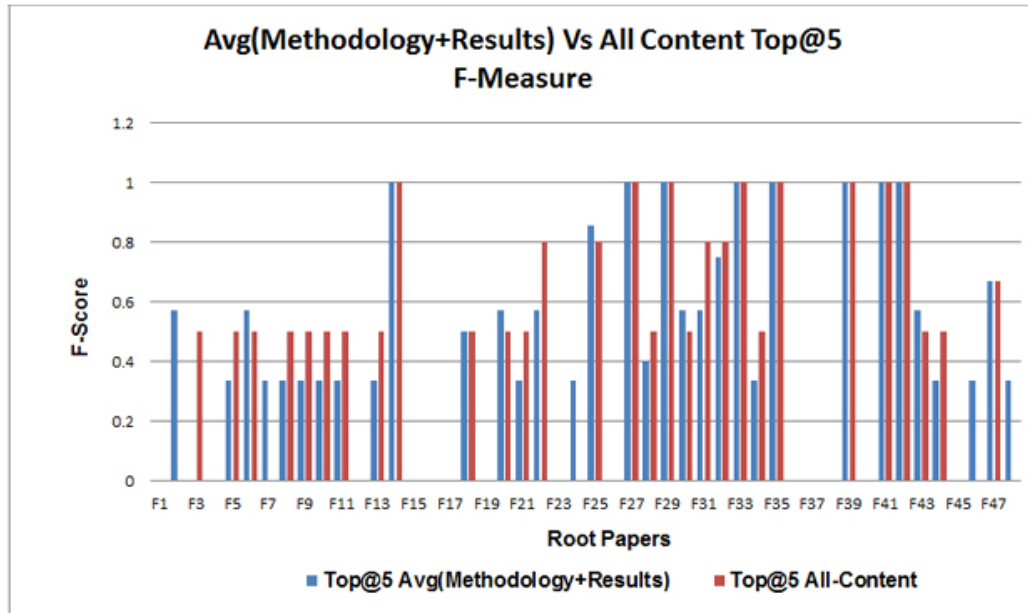


FIGURE 4.17: Avg(Methodology+Results) Vs All Content Top 5

4.3 Combination of Double Parameters using Weighted Average Results

4.3.1 Wt.AVR(Abstract-Methodology-Sections Similarity-Based Ranking)

In this ranking technique, the contents of abstract and methodological sections of cited document was compared with the contents of corresponding logical sections (i-e abstract and methodology) of all citing documents. Then the similarity score was computed between the logical sections of cited and citing documents.

The combination of similarity scores of two logical sections (abstract and methodology) was done by utilizing the weighted average formula depicted in equation 3.6. The computed score was then arranged in descending order. This score was utilized to rank the top 3 and top 5 documents. The F-measure for each cited document was computed individually. The cumulative F-measures for the Wt.AVR(abstract-methodology-section similarity-based ranking) were 0.72 for the

top 3 documents and 0.68 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. Comparisons between the Wt.AVR(abstract-methodology-section similarity-based ranking) and full-content similarity are shown in **Figure 4.18** and **4.19** for the top 3 and top 5 ranked documents, respectively.

The cited paper number is listed on the x-axis and the F-measure score on the y-axis. The red line shows the F-measure for each cited paper using the full-content approach, whereas the blue line shows the Wt.AVR(abstract-methodology-section similarity-based ranking) F-measure. For most of the cited papers, the red line and blue line follow the same path, meaning that both approaches produced the same results in these cases.

It is also clear from **Figure 4.18** and **4.19** that when the results of Wt.AVR(abstract-methodology-section similarity-based ranking) and content-based similarity differ, Wt.AVR(abstract-methodology-section similarity-based ranking) produces more accurate results.

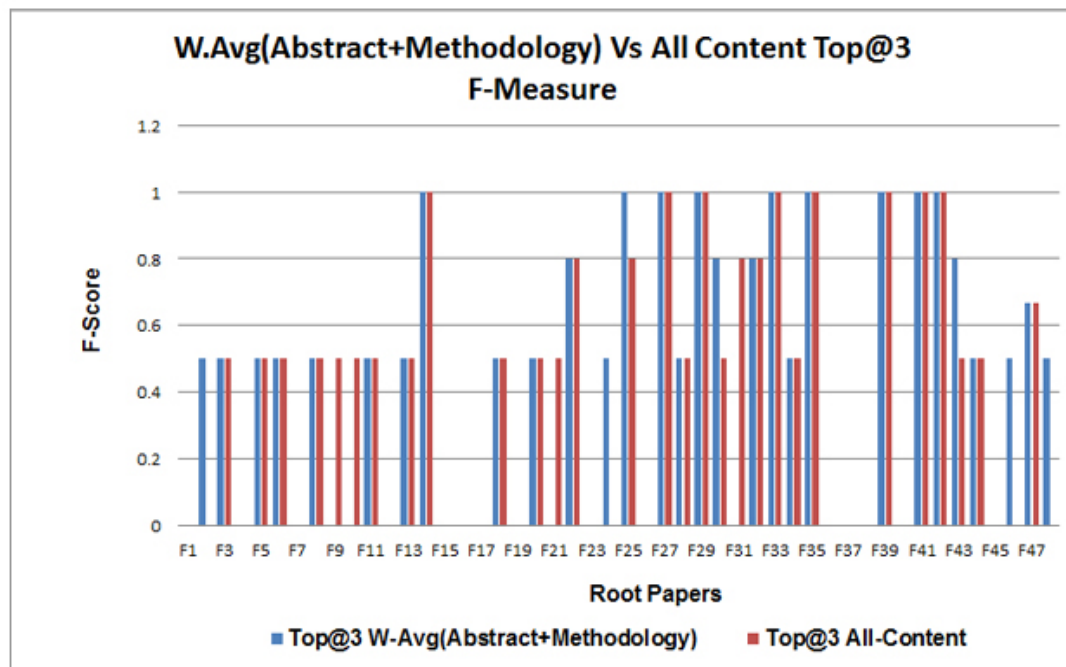


FIGURE 4.18: W.Avg(Abtract+Methodology) Vs All Content Top 3

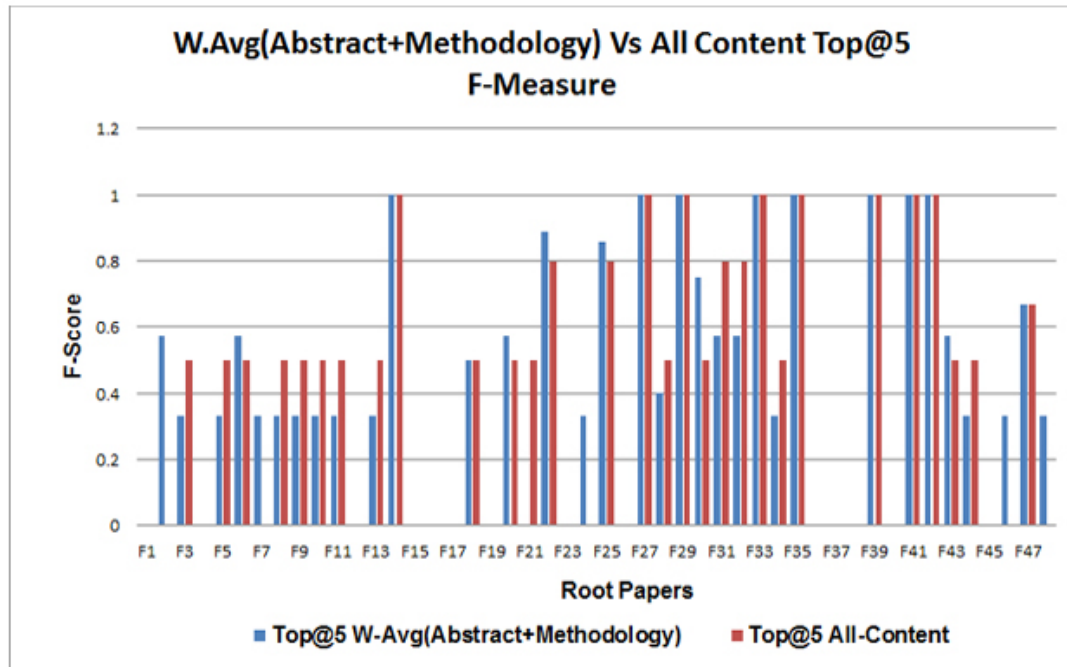


FIGURE 4.19: W.Avg(Abtract+Methodology) Vs All Content Top 5

4.3.2 Wt.AVR(Abtract-Results-Sections Similarity-Based Ranking)

In this ranking technique, the contents of abstract and results sections of cited document was compared with the contents of corresponding logical sections (i.e. abstract and results) of all citing documents. Then the similarity score was computed between the logical sections of cited and citing documents. The combination of similarity scores of two logical sections (abstract and results) was done by utilizing the weighted average formula depicted in equation 3.6. The computed score was then arranged in descending order.

This score was utilized to rank the top 3 and top 5 documents. The F-measure for each cited document was computed individually. The cumulative F-measures for the Wt.AVR(abstract-results-section similarity-based ranking) were 0.67 for the top 3 documents and 0.64 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63

and 0.65, respectively. Comparisons between the Wt.AVR(abstract-results-section similarity-based ranking) and full-content similarity are shown in **Figure 4.20** and

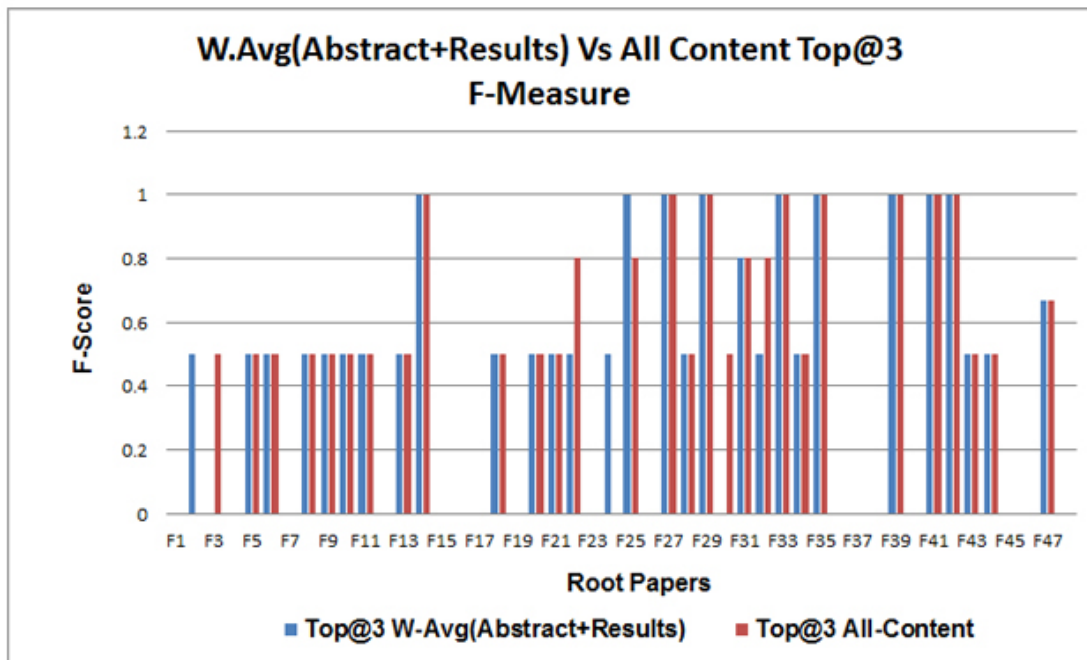


FIGURE 4.20: W.Avg(Abtract+Results) Vs All Content Top 3

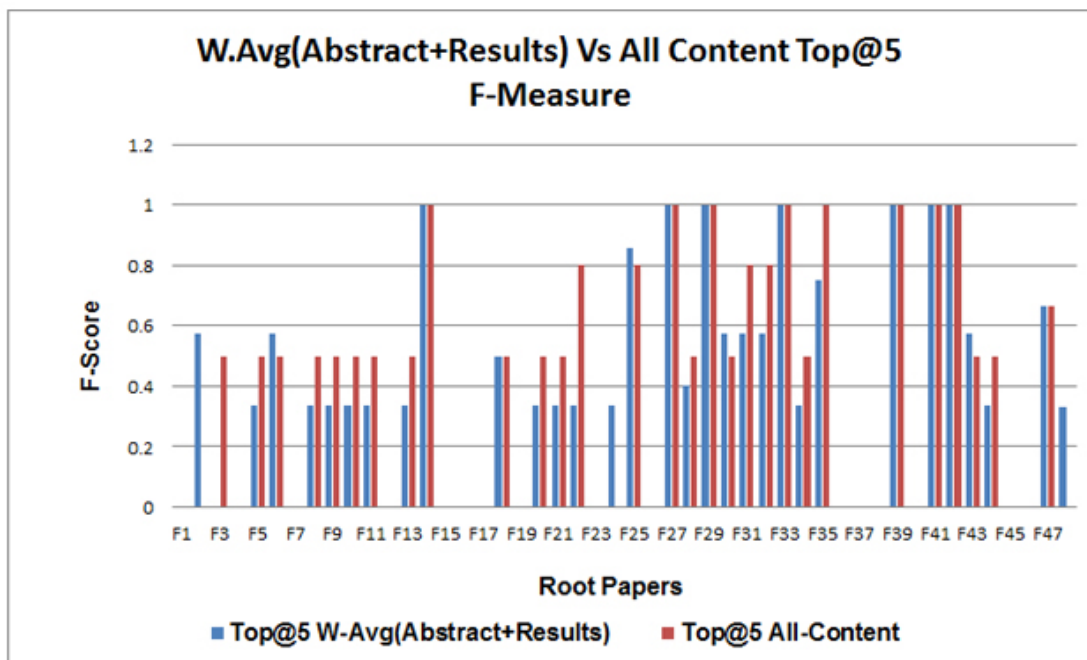


FIGURE 4.21: W.Avg(Abtract+Results) Vs All Content Top 5

4.21 for the top 3 and top 5 ranked documents, respectively. The cited paper number is listed on the x-axis and the F-measure score on the y-axis. The red line shows the F-measure for each cited paper using the full-content approach, whereas the blue line shows the Wt.AVR(abstract-results-section similarity-based ranking) F-measure. For most of the cited papers, the red line and blue line follow the same path, meaning that both approaches produced the same results in these cases.

It is also clear from **Figure 4.20** and **4.21** that the results of top 3 Wt.AVR(abstract-results-section similarity-based ranking) outperforms against content-based similarity, but the results of top 5 Wt.AVR(abstract-results-section similarity-based ranking) are approximately to the full-content similarity based ranking.

4.3.3 Wt.AVR(Methodology-Results-Sections Similarity-Based Ranking)

In this ranking technique, the contents of abstract and results sections of cited document was compared with the contents of corresponding logical sections (i.e methodology and results) of all citing documents. Then the similarity score was computed between the logical sections of cited and citing documents. The combination of similarity scores of two logical sections (methodology and results) was done by utilizing the weighted average formula depicted in equation 3.6. The computed score was then arranged in descending order. This score was utilized to rank the top 3 and top 5 documents. The F-measure for each cited document was computed individually.

The cumulative F-measures for the Wt.AVR(methodology-results-section similarity-based ranking) were 0.67 for the top 3 documents and 0.67 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. Comparisons between the Wt.AVR(methodology-results-section similarity-based ranking) and full-content similarity are shown in **Figure 4.22** and **4.23** for the top 3 and top 5 ranked

documents, respectively. The cited paper number is listed on the x-axis and the F-measure score on the y-axis. The red line shows the F-measure for each cited paper using the full-content approach, whereas the blue line shows the Wt.AVR(methodology-results-section similarity-based ranking) F-measure. For most of the cited papers, the red line and blue line follow the same path, meaning that both approaches produced the same results in these cases.

It is also clear from **Figure 4.22** and **4.23** that the results of top 3 Wt.AVR(methodology-results-section similarity-based ranking) outperforms against the full content-based similarity, but the results of top 5 Wt.AVR(methodology-results-section similarity-based ranking) and full-content similarity based ranking are approximately equal.

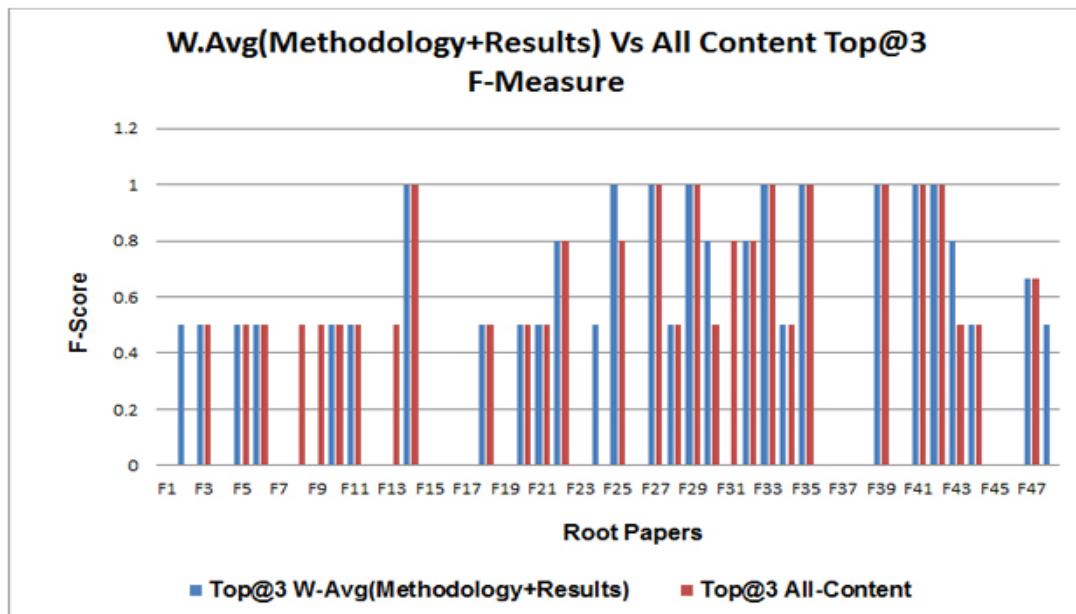


FIGURE 4.22: W.Avg(Methodology+Results) Vs All Content Top 3

4.4 Double Ranking approaches Conclusion

The Table 4.2 summarizes the results of combining two parameters with the average approach in contrast to the Full-content parameter. The findings are derived from the top three and top five ranked research papers. Table 4.3 shows the results of combining two parameters with the Weighted Average approach in comparison

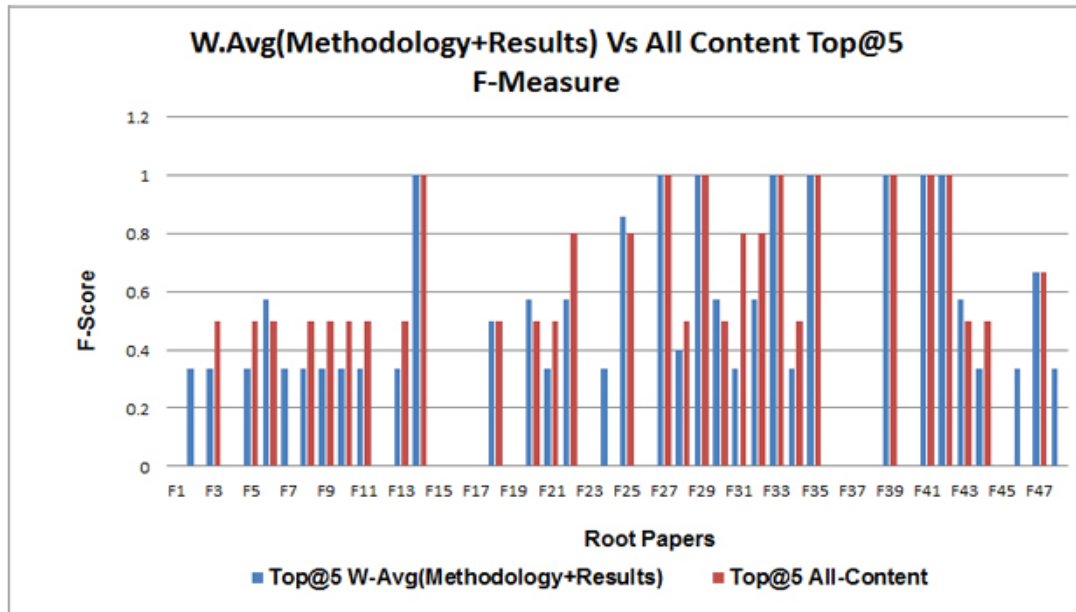


FIGURE 4.23: W.Avg(Methodology+Results) Vs All Content Top 5

TABLE 4.2: Average Scores of Double Parameters Vs All Content.

Double Parameters	R.Papers	Score	All-Content
AVR(Abstract,Methodology)	Top 3 Ranked	0.74	0.63
AVR(Abstract,Methodology)	Top 5 Ranked	0.70	0.65
AVR(Abstract,Result)	Top 3 Ranked	0.63	0.63
AVR(Abstract,Result)	Top 5 Ranked	0.67	0.65
AVR(Methodology,Result)	Top 3 Ranked	0.67	0.63
W.AVR(Methodology,Result)	Top 5 Ranked	0.67	0.65

TABLE 4.3: Weighted Average Scores of Double Parameters Vs All Content.

Double Parameters	R.Papers	Score	All-Content
W.AVR(Abstract,Methodology)	Top 3 Ranked	0.72	0.63
W.AVR(Abstract,Methodology)	Top 5 Ranked	0.68	0.65
W.AVR(Abstract,Result)	Top 3 Ranked	0.67	0.63
W.AVR(Abstract,Result)	Top 5 Ranked	0.64	0.65
W.AVR(Methodology,Result)	Top 3 Ranked	0.68	0.63
W.AVR(Methodology,Result)	Top 5 Ranked	0.65	0.65

to the Full-Content parameter. It is evident that the double parameter (Abstract + Methodology) scored higher than all other double parameters when combined with the (1)-Average approach and (2)-Weighted Average approach. This parameter obtained F-measures of 0.74 and 0.70 for the top three and top five research papers, respectively, when paired with the average approach. But when it was used with

the Weighted Average method, it gave F-measure scores of 0.72 for the top three research papers and 0.68 for the top five.

4.5 Combination of Triple Parameters using Average Results

4.5.1 AVR(Abstract-Methodology-Results-Sections Similarity-Based Ranking)

In this ranking technique, the contents of abstract and methodological and results sections of cited document was compared with the contents of corresponding logical sections (i-e abstract methodology and results) of all citing documents. Then the similarity score was computed between the logical sections of cited and citing documents. The combination of similarity scores of three logical sections (abstract methodology and results) was done by utilizing the average formula depicted in equation 3.6. The computed score was then arranged in descending order. This score was utilized to rank the top 3 and top 5 documents. The F-measure for each cited document was computed individually.

The cumulative F-measures for the AVR(abstract-methodology-results-section similarity-based ranking) were 0.75 for the top 3 documents and 0.71 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. Comparisons between the AVR(abstract-methodology-results-section similarity-based ranking) and full-content similarity are shown in **Figure 4.24** and **4.25** for the top 3 and top 5 ranked documents, respectively. The cited paper number is listed on the x-axis and the F-measure score on the y-axis. The red line shows the F-measure for each cited paper using the full-content approach, whereas the blue line shows the AVR(abstract-methodology-results-section similarity-based ranking) F-measure. For most of the cited papers, the red line and blue line follow the same path, meaning that both approaches produced the same results in

these cases. It is also clear from **Figure 4.24** and **4.25** that when the results of AVR(abstract-methodology-results-section similarity-based ranking) and content-based similarity differ, AVR(abstract-methodology-results-section similarity-based ranking) produces more accurate results.

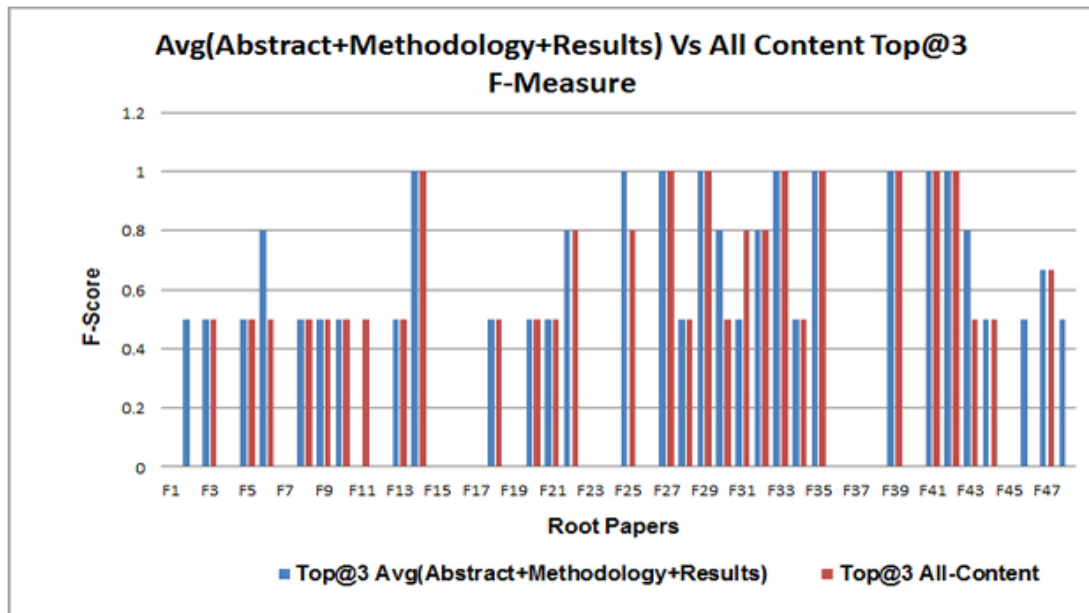


FIGURE 4.24: Avg(Abtract+Methodology+Results) Vs All Content top 3

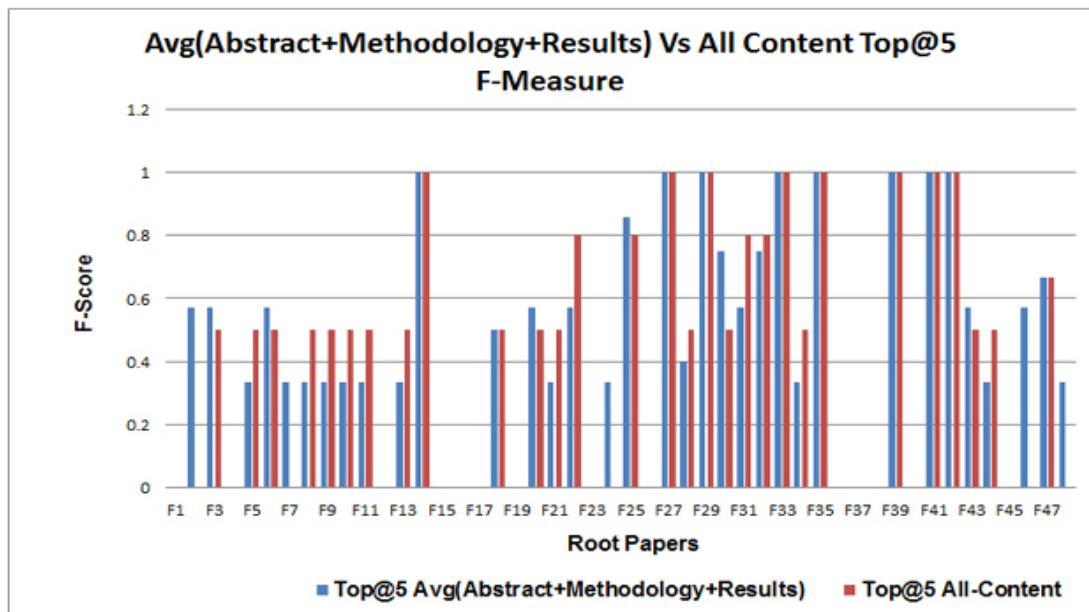


FIGURE 4.25: Avg(Abtract+Methodology+Results) Vs All Content Top 5

4.6 Combination of Triple Parameters using Weighted Average Results

4.6.1 Wt.AVR(Abtract-Methodology-Results-Sections Similarity-Based Ranking)

In this ranking technique, the contents of abstract and methodological and results sections of cited document was compared with the contents of corresponding logical sections (i-e abstract methodology and results) of all citing documents. Then the similarity score was computed between the logical sections of cited and citing documents. The combination of similarity scores of three logical sections (abstract methodology and results) was done by utilizing the weighted average formula depicted in equation 3.6. The computed score was then arranged in descending order. This score was utilized to rank the top 3 and top 5 documents. The F-measure for each cited document was computed individually. The cumulative F-measures for the Wt.AVR(abstract-methodology-results-section similarity-based ranking) were 0.68 for the top 3 documents and 0.67 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. Comparisons between the Wt.AVR(abstract-methodology-results-section similarity-based ranking) and full-content similarity are shown in **Figure 4.26** and **4.27** for the top 3 and top 5 ranked documents, respectively. The cited paper number is listed on the x-axis and the F-measure score on the y-axis. The red line shows the F-measure for each cited paper using the full-content approach, whereas the blue line shows the Wt.AVR(abstract-methodology-results-section similarity-based ranking) F-measure. For most of the cited papers, the red line and blue line follow the same path, meaning that both approaches produced the same results in these cases. It is also clear from **Figure 4.26** and **4.27** that the results of top 3 Wt.AVR(abstract-methodology-results-section similarity-based ranking) outperforms against the full content-based similarity, but the results of top 5 Wt.AVR(abstract-methodology-results-section similarity-based ranking) and full-content similarity based ranking are approximately equal.

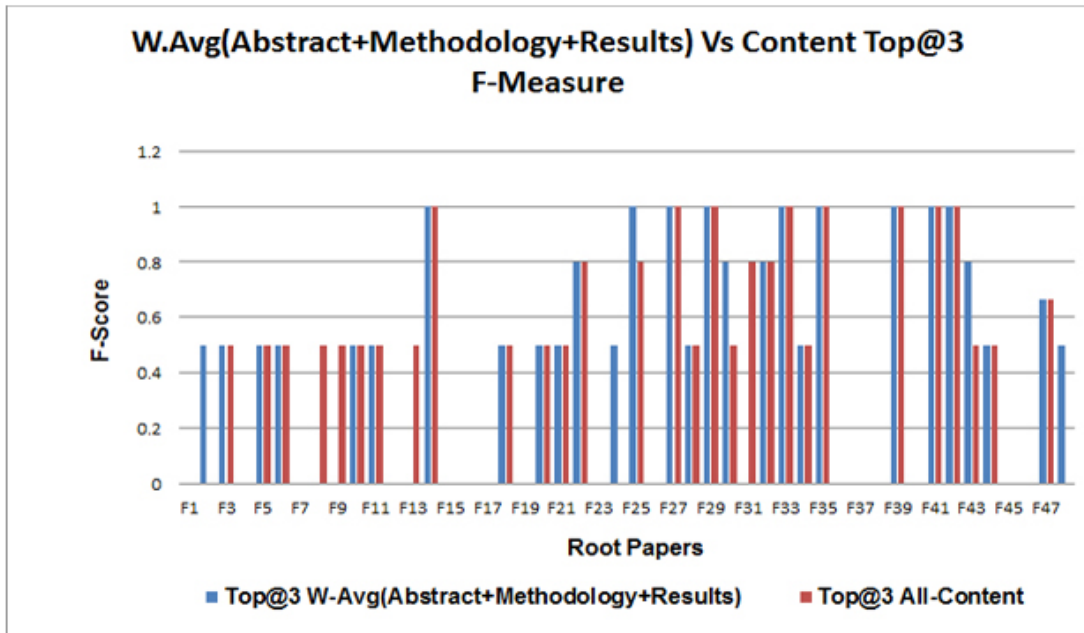


FIGURE 4.26: W-Avg(Abstr+Meth+Res) Vs All Content top 3

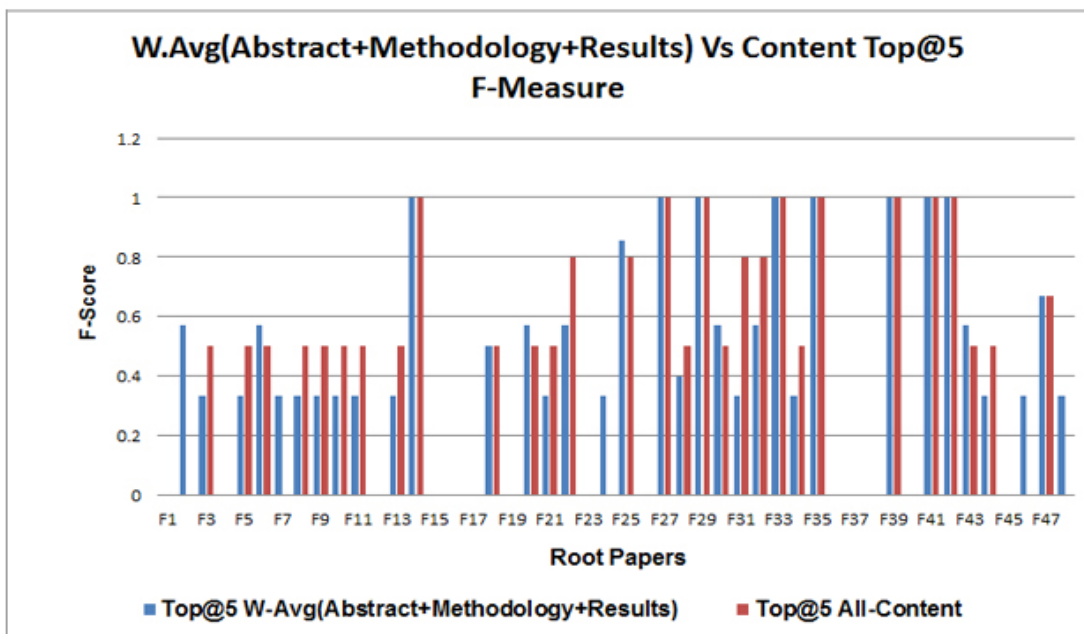


FIGURE 4.27: W.Avg(Abstr+Meth+Res) Vs All Content

4.7 Triple Ranking approaches Conclusion

Combining the individual sections of abstract, methodological, and results by using average and weighted average approaches yield the triple ranking parameters. The findings are derived from the top three and top five research publications.

The AVR(Abtract, Methodology, and Result) triple ranking parameters received F-measures of 0.75 and 0.71 for the top 3 and top 5 research publications, respectively. This triple ranking parameter is the most effective scorer parameter among single, double, and triple parameters. While the second triple parameter, W.Avg. (Abtract, Methodology, and Result), had F-measure scores of 0.68 and 0.67 for the top 3 and top 5 research papers, respectively.

4.8 Comparison to State-of-the-art Approaches

The previous section proposed six new rankings for identifying important citations for cited papers from the list of citing papers. Comparing abstracts was identified as the best ranking technique based on a critical analysis of the results for all six rankings. This section, in turn, compares the results of the best proposed ranking to the best parameter rankings achieved by current state-of-the-art approaches. The proposed approach was compared with the state-of-the-art approaches using the same dataset. The comparisons of precision and recall are depicted in **Figure 4.28** and **Figure 4.29**, respectively.

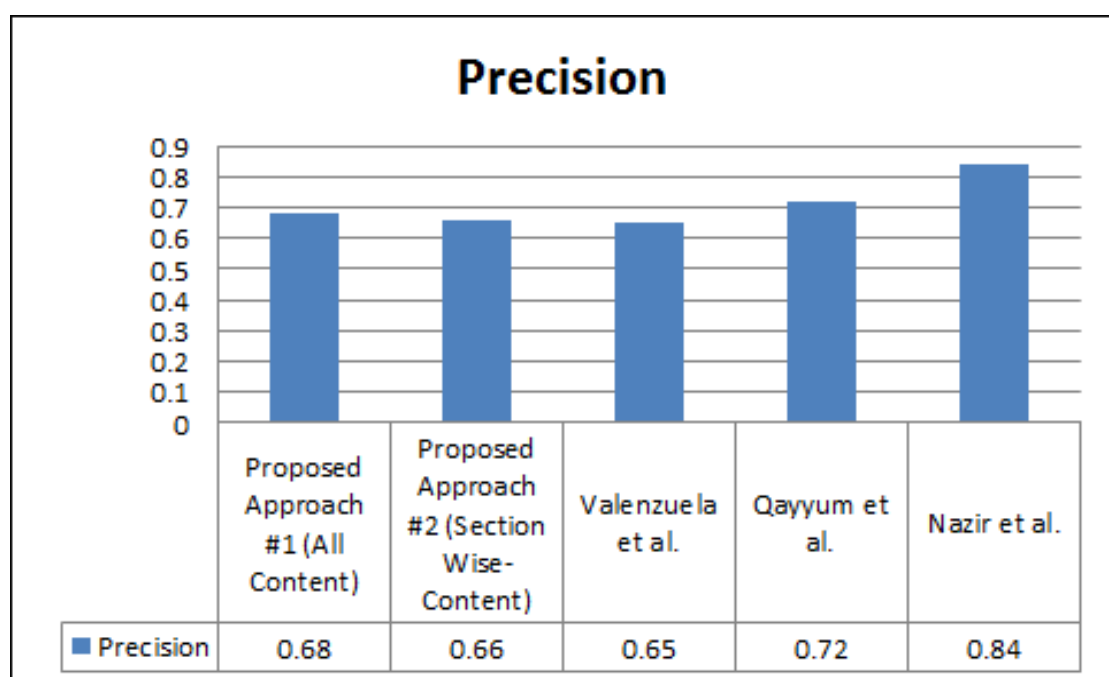


FIGURE 4.28: Comparison the precision of the proposed approaches with state-of-the-art rankings

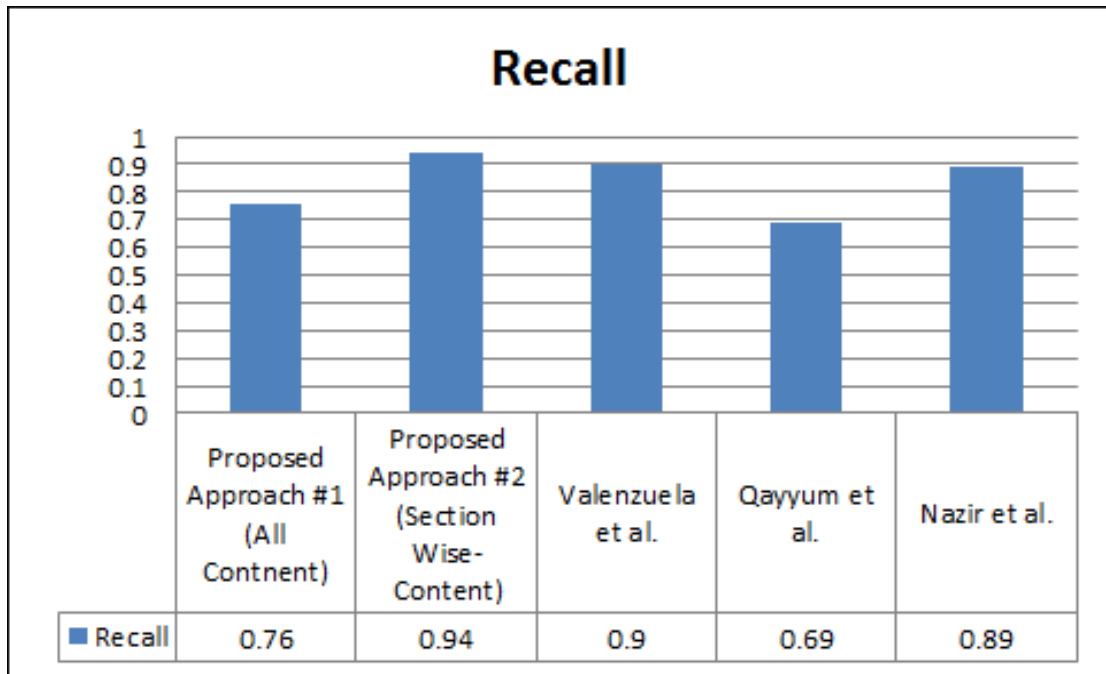


FIGURE 4.29: Comparing the recall of the proposed approaches with state-of-the-art rankings

The proposed approach was compared to the following state-of-the-art approaches. The first approach was presented by Valenzuela et al. [14], who conducted the pioneering work in this area and have made their dataset freely available online. This is the same dataset used by the proposed approach and the other approaches represented in **Figure 4.11** and **Figure 4.28**. Valenzuela et al. [14] tested 12 features as identifiers of important citations for the cited papers, with the in-text citation-based feature producing the best results.

The second state-of-the-art approach was proposed by Qayyum and Afzal [24]. They presented a hybrid approach that uses metadata and content-based features to identify important citations. The third state-of-the-art approach is the technique was proposed by Nazir et al. [26], who extended the approach of Valenzuela et al. [14]. They assigned weights to different sections of the paper to better capture the significance of in-text citation counts.

Figure 4.28 compares the precision results for the newly proposed approaches and three existing state-of-the-art approaches. The x-axis lists the approaches names and the y-axis the precision score. The results for proposed approach #1 (utilizing the full content of the cited-citing pair) is 0.68, and the result of proposed

approach #2 (examining content similarity in the abstract sections of the cited-citing pair) is 0.66. Nazir et al. [26] achieved the maximum precision of 0.84, followed by Qayyum and Afzal [24] with a precision score of 0.72. These are the best results from a variety of feature evaluations conducted within each study. The results indicates that the proposed approach outperformed Valenzuela et al.s [14] approach but was inferior to other state-of-the-art approaches. However, this is not in fact the case. To illustrate why, let us discuss the results of each approach one-by-one.

Valenzuela et al. [14] achieved a maximum precision of 0.37 when employing only a single parameter, namely direct citations per section. When examining only a single parameter, the newly proposed approach focusing solely on the abstract achieved a precision score of 0.66, thus outperforming Valenzuela et al. In comparison, Valenzuela et al. achieved a precision score of 0.65 when aggregating all 12 parameters, still slightly lower than the precision score of 0.66 obtained by Proposed Approach #2. Furthermore, to compare this value, one needs to consider the following facts. Valenzuela et al. have not discussed how accurately they extracted the 12 features. For example, metadata features like keywords are only available around 50% of the time [24]. Furthermore, the accurate extraction of in-text citation counts is not a trivial task and requires very sophisticated algorithms. This has been pointed out by [9], who achieved 58% accuracy in extracting in-text citations. Although an approach recently proposed by Ahmad and Afzal [110] raises this accuracy, it still needs to be verified on journals from diverse fields and different publishers styles.

Therefore, the precision score of 0.65 achieved by Valenzuela et al. is dependent on the accurate identification of in-text citation counts. If the approach by Valenzuela et al. [14] were to extract in-text citation counts automatically using the procedures presented by Shahid et al. [9], the precision score might remain in the range of around 0.3. Standard tools such as Content ExtRactor and MINER (CERMINE) [106] and GeneRation Of Bibliographic Data (GROBID) [111] could only achieve precision, recall, and F-measure scores in the range of 0.8 to 0.9 when evaluated by [110]. Thus, if Valenzuela et al. [14] were to apply the best

automated approach to detecting in-text citations, the precision score for finding important citations would drop from 0.65 to less than 0.5. In contrast, the proposed approach does not require any such complex parameter computations; it is based solely on the content of the abstract, which is freely available. Therefore, in terms of real applications, the proposed approach outperforms Valenzuela et al.'s approach in terms of precision score and thus can be considered a viable solution for citation indexes and digital libraries.

The second state-of-the-art approach was presented by Qayyum and Afzal [24]. They classified citing papers as important/non-important citations for the cited paper using metadata and the papers content. The best individual feature they examined achieved a precision score of 0.35. Thus, with respect to single features, the proposed approach utilizing the abstract alone outperforms Qayyum and Afzal [24]. However, when Qayyum and Afzal [24] aggregated four metadata elements, the precision score using the random forest classifier reached to 0.72. Important to consider here is that this score can only be obtained when all four metadata elements are available. For example, only 58.3% of Qayyum and Afzal's [24] dataset included keywords. This approach is not applicable in the scenarios wherein metadata is not present in equal ratio. Furthermore, cue phrases need to be identified for each individual dataset. This makes the method impractical to use in real systems. In contrast, the newly proposed approach does not rely upon defining a cue phrase dictionary or the availability of keywords.

The third approach selected for comparison is the technique proposed by Nazir et al. [26]. They used section-based in-text citation frequencies to classify citations as important or non-important. A further novel element of this approach is their identification of suitable section weights using linear regression. The approach achieved a precision score of 0.84. However, the present comparison demonstrates the pitfalls of this state-of-the-art approach. Specifically, it is necessary to calculate in-text citation frequencies, which is quite challenging to perform automatically, as noted above. Another challenge concerns mapping section headings onto logical sections (such as Introduction, Literature, Methodology, Results and Discussion). Shahid and Afzal [24] achieved the highest accuracy for this task which is 78%.

Considering all these factors, the proposed approach is comparable to the best-known existing approach as it does not require any complex calculations to be performed unlike other state-of-the-art approaches.

The recall of both proposed approaches and existing state-of-the-art approaches is compared in **Figure 4.30** Proposed Approach # 2 (section-wise similarity between abstracts) achieved the best recall of 0.94, higher than existing state-of-the-art approaches. Content-based approaches such as those used by search engines and citation indexes are considered the best approaches to obtain maximum recall.

This means that 90% of the time, important citations are identified as such by the proposed approach, with some noise. The proposed approach not only achieves better recall, its implementation is also more viable for the following two reasons: (1) it does not require complex calculations of in-text citation frequencies, mapping section headings to logical sections, the availability of all metadata fields, or identifying cue phrases for each dataset, and (2) abstracts are generally available online for free.

4.9 Comparisons of n-gram Features

This section highlights the results n-gram features. Here n-gram means the bi-gram features and tri-gram features. In bi-gram features, all possible combination of two sections have been used. For the tri-gram features, all possible combinations of three sections have been used.

Our proposed approach depends on the different logical sections of research papers. These logical sections include Abstract, Introduction, Literature Review, Methodology, Result, and all possible combinations. We have compared our proposed approach results with three recent state-of-the-art approaches of Valenzuela [14] and Qayyum & Afzal [24] due to the following two reasons. (1)-Both of the state-of-the-art approaches of Valenzuela [14] and Qayyum & Afzal [24] published their research work in well reputable journals and conferences. Valenzuela [14] published their work "in Workshops at the twenty-ninth AAAI conference on artificial

intelligence with titled "Identifying meaningful citations." in the year 2015. While Qayyum & Afzal [24] published research work in *Scientometrics* journal with titled "Identification of important citations by exploiting research articles metadata and cue-terms from the content." in the year 2019. (2)- We have used the same data set (i.e. Valenzuela comprehensive annotated dataset) as used by Valenzuela [14] and Qayyum & Afzal [24] for comparison. This thesis has considered three top performing logical sections as explained in the section 4.1.1.7.

The similarity score of all logical sections has been calculated by using the cosine similarity approach and different logical sections have been combined by using average and weighted average techniques. The results have been extracted by using every possible combination of logical sections from single combinations parameters to triple combinations. The extracted results have been sorted in descending order and then Precision, Recall, and F-measure scores have been calculated of the top 3 and top 5 ranked papers. The Valenzuela's [14] proposed approach achieved an average precision score of 0.65 by combining all of their 12 features. While Qayyum & Afzal [24] approach achieved 0.73 average precision by combining all the features. In our proposed approach, the single parameter scored an F-Measure score of 0.63 of the top 3 ranked papers and 0.64 of the top 5 ranked papers. By using two parameters combined with a weighted average technique system scored an F-measure score of 0.69 of the top 3 ranked papers and 0.66 of the top 5 ranked papers. When two parameters combined with average technique then the system scored an F-measure score of 0.68 of the top 3 and top 5 ranked papers. When we combined three parameters through weighted average technique then the system achieved 0.68 F-Measure score of top 3 ranked papers and F-Measure score of 0.67 of the top 5 ranked papers. Similarly, when three parameters were combined through an average technique then our system achieved 0.75 F-Measure scores of the top 3 ranked papers and 0.71 F-Measure scores of top 5 ranked papers. The best feature (i.e. triple parameter) of the proposed approach yielded a precision of 0.75 when top 3 ranked papers were considered. While when considering the top 5 ranked papers then the best feature (i.e. triple parameter) yielded a precision of 0.73.

Chapter 5

Discussion

5.1 Conclusion

Identifying the set of important citations for the cited paper from the list of citing papers is a challenge that has led the scientific community to propose a wide range of techniques. The critical analysis of more than 80 papers have revealed the following facts. Researchers have tried to classify papers into different number of reasons ranging from 2 reasons to 150 reasons using manual, automated and hybrid approaches. However, there is a paradigm shift for citation classification from larger number of reasons (classes) to binary classes such as: Important citation and non-important citation. There have been many recent studies which have focused to classify citations into binary classes. Therefore, this thesis has focused to contribute in this area and have raised the bar for the citation classification for binary classes. This research has critically evaluated the literature and identified three such state-of-the-art approaches to classifying citations into two classes, namely important and non-important. These existing approaches have utilized different set of features and have proposed diversified classification methods . The precision of these state-of-the-art approaches range from 0.72 to 0.84. However, they are dependent on the accurate identification of some complex features, such as in-text citation frequencies, mapping section headings onto section labels, availability of metadata elements, and constructing dataset-dependent dictionaries of cue- phrases. The accuracy values achieved by the state-of-the-art

approaches mentioned above are only possible when all of these parameters are extracted accurately.

However, a critical analysis shows that the accuracy of identifying in-text citations varies from 58% to 90%, as highlighted by different research and state-of-the-art tools. The accuracy of mapping section headings onto logical sections is just 78%. Keyword metadata is available only 53% of the time. Cue-phrases developed for one particular dataset needs to be re-developed for another dataset. Currently, state-of-the-art approaches extract such features in a semi-automatic way, and incorrect values are corrected manually. However, when all of these features are extracted fully automatically, the precision score may drop to one-third of the reported values.

This research presents a method that does not require the computation of such complex features. In the similar domain of identifying relevant research papers, papers content has been successfully used for nearly two decades to identify relevant papers. Based on these findings, this paper adopted the content-based similarity approach to identify the similarity between pairs of cited and citing papers. Furthermore, a novel approach involving section-based similarity was proposed, implemented, and evaluated. An in-depth analysis of both proposed approaches indicated that the abstract alone is sufficient to decide whether the citing paper is important or non-important for the cited paper. The proposed approach achieved precision scores of 0.68 for full content and 0.66 for the abstract section, respectively, outperforming existing state-of-the-art approaches when considering the facts presented above.

Furthermore, the recall of existing state-of-the-art approaches range from 0.7 to 0.9, while the proposed approach has achieved a recall score of 0.94. Thus, the proposed approach significantly outperformed existing approaches in terms of recall, particularly when considering that inaccurate calculations of in-text citations, section mapping, metadata availability, and cue-phrase construction will significantly reduce recall scores for the state-of-the-art approaches when conducted automatically. In contrast, there is no need for such complex calculations in the proposed approach. Almost all of the existing approaches have given importance to the met-

adata, in-text citation counts, and their positions in the research article. It is important to note that the metadata of the research article does not contain any domain-specific terms that could have a detrimental effect on accuracy. Currently, none of the existing approaches have compared the important terms represented in different corresponding logical sections of research articles in order to determine the important and non-important relationship between two articles (i.e. citing and cited articles). We have proposed a comprehensive methodology for dealing with the issue that has been raised. Our methodology compares the content of corresponding logical sections of cited and citing research articles.

We have evaluated five different logical sections of research paper. These logical sections are: Abstract, Introduction, Literature Review, Methodology, and Results sections. The similarity between the corresponding logical sections and their possible combinations are carried out by utilizing cosine similarity technique, figure 4.7 shows all possible combinations. These combinations include single parameters as well as other combination. The combinations of two or more logical sections are achieved using average and weighted average approaches. A comprehensive dataset of Valenzuela et al was used to perform experiments as it was used by all three approaches selected for comparisons. The given dataset contains 465 research articles that are considered to be paper-citations pairs. Out of which 48 are root papers and the remaining are citations. In pre-processing phase, stop words were removed and content terms were converted into their root words.

The F-measure scores of the top three and top five ranked papers based on the logical sections are computed independently and compared to the scores of the top three and top five ranked of All-Content (i.e. whole content) of research papers. The F-measure score that are computed for abstract, methodology, and results sections outperforms the score computed for All-Content (i.e. whole content) of the research papers. Whereas introduction and literature review sections scores are considerable low than the scores of All-Contents. Consequently, for further experiments, only top three performing sections (abstract, methodology and result) have been used for developing the combinations. The given experiments were performed on the comprehensive dataset of Valenzuela et al. The building mecha

nism of above said combinations were shown in the figure 4.7. We took all possible combinations of the three selected logical sections (i-e abstract, methodology, and results) and computed the F-Measure score of each combination of the top three and top five ranked articles.

In the case of double parameters, two logical sections were combined using average and weighted average techniques. When combined with an average approach, the combination of Avg. (abstract, methodology) outperformed other double parameters, with F-measures of 0.74 and 0.70 of the top three and top five ranked research articles, respectively. When combined with a weighted average technique, the W.Avg. (abstract, methodology) sections combinations produced better results than other double parameters. This combination received an F-measure score of 0.72 for the top three and 0.68 for the top five ranked research articles, respectively. Three logical sections (abstract, methodology, and result (A, M, R)) were combined with the help of average and weighted average techniques in the triple parameters. When combined with an average technique, this triple combination is the top scorer parameter, with F-measures of 0.75 and 0.71 of the top 3 and top 5 ranked research articles, respectively. When this triple combination was performed using the weighted average technique, it achieved an F-measure score of 0.68 and 0.67, respectively, of the top three and top five ranked research articles.

The outcomes of the above experiments were compared to the outcomes of the state-of-the-art methods of Qayyum & Afzal et al, Valenzuela et al, and Nazir et al [24, 14, 26] and All-Content based approaches. By integrating all 12 different characteristics, Valenzuela et al obtained an accuracy (i.e. precision) of 0.65, while Qayyum & Afzal et al achieved an average accuracy (i.e. precision) of 0.72. The F-Measure score for the All-Content-Based Approach was 0.63 for the top three ranked papers and 0.65 for the top five ranked articles. While the best combination of our suggested approach's abstract, methodology, and result (A, M, R) parts had an F-Measure score of 0.75. The finding of this study is that research articles are produced using domain-specific terminology and expertise. It is more likely that both papers will employ similar language and concepts since they belong to the same group or are closely working on the same subject or expanding previous work

in another work. If the citing article utilizes domain specific words and expertise in the abstract, methodology, and results section, there is a greater probability that both publications belong to the same domain and have significant citations. The logical parts of a research paper are of varying significance, and there is a greater likelihood that the citing and cited article utilize comparable vocabulary words in the same sections of their papers.

5.2 Limitations

In this thesis, we utilized a publicly accessible annotated dataset that was mainly used in Valenzuela's et al method. This dataset has 465 total annotated paper-citation pairings, which is insufficient to draw broad conclusions. Because this area has a scarcity of big annotated datasets, further study will need the creation of huge standard annotated datasets. The datasets that should cover various writers from various geographical locations and topics.

5.3 Future Work

In the field of text mining, it is still hard to convert PDF to XML and text. About 5% of PDFs were not in a format that cutting-edge technologies could recognize, and these files were also not good for converting text, that is, from PDF to XML and text. even though a researcher got a PhD for coming up with new ways to convert PDF files to XML and text. It is evident that approximately 5% of documents were not converted correctly. So, new experiments and new ways of doing things are needed. This could lead to all PDFs being changed to XML and text formats.

Bibliography

- [1] P. Knowledge, “An essay concerning the social dimension of science,” *London: Cambridge UP*, 1968.
- [2] F. Narin, “The use of publication and citation analysis in the evaluation of scientific activity,” *Evaluative Bibliometrics, Computer Horizons, Inc., Cherry Hill, NJ for NSF Contract NSF-C627*, 1976.
- [3] H. Inhaber and K. Przednowek, “Quality of research and the nobel prizes,” *Social Studies of Science*, vol. 6, no. 1, pp. 33–50, 1976.
- [4] R. C. Anderson, F. Narin, and P. McAllister, “Publication ratings versus peer ratings of universities,” *Journal of the American Society for Information Science*, vol. 29, no. 2, pp. 91–103, 1978.
- [5] A. T. Smith and M. Eysenck, “The correlation between rae ratings and citation counts in psychology,” 2002.
- [6] S. Maqsood, M. A. Islam, M. T. Afzal, and N. Masood, “A comprehensive author ranking evaluation of network and bibliographic indices,” *Malaysian Journal of Library & Information Science*, vol. 25, no. 1, pp. 31–45, 2020.
- [7] J. E. Hirsch, “An index to quantify an individual’s scientific research output,” *Proceedings of the National academy of Sciences*, vol. 102, no. 46, pp. 16 569–16 572, 2005.
- [8] A. Mazloumian, D. Helbing, S. Lozano, R. P. Light, and K. Börner, “Global multi-level analysis of the scientific food web’,” *Scientific reports*, vol. 3, no. 1, pp. 1–5, 2013.

-
- [9] A. Shahid, M. T. Afzal, and M. A. Qadir, "Lessons learned: The complexity of accurate identification of in-text citations." *Int. Arab J. Inf. Technol.*, vol. 12, no. 5, pp. 481–488, 2015.
- [10] E. Garfield *et al.*, "Can citation indexing be automated," in *Statistical association methods for mechanized documentation, symposium proceedings*, vol. 269. Publication National Bureau of Standards, Miscellaneous, 1965, pp. 189–192.
- [11] I. Spiegel-Rosing, "Science studies: Bibliometric and content analysis," *Social Studies of Science*, vol. 7, no. 1, pp. 97–113, 1977.
- [12] L. Bornmann and H.-D. Daniel, "What do citation counts measure? a review of studies on citing behavior," *Journal of documentation*, 2008.
- [13] M. J. Moravcsik and P. Murugesan, "Some results on the function and quality of citations," *Social studies of science*, vol. 5, no. 1, pp. 86–92, 1975.
- [14] M. Valenzuela, V. Ha, and O. Etzioni, "Identifying meaningful citations," in *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, 2015.
- [15] X. Zhu, P. Turney, D. Lemire, and A. Vellino, "Measuring academic influence: Not all citations are equal," *Journal of the Association for Information Science and Technology*, vol. 66, no. 2, pp. 408–427, 2015.
- [16] R. Benedictus, F. Miedema, and M. W. Ferguson, "Fewer numbers, better science," *Nature*, vol. 538, no. 7626, pp. 453–455, 2016.
- [17] M. H. MacRoberts and B. R. MacRoberts, "The mismeasure of science: Citation analysis," *Journal of the Association for Information Science and Technology*, vol. 69, no. 3, pp. 474–482, 2018.
- [18] J. Wilsdon, "The metric tide: Independent review of the role of metrics in research assessment and management," 2016.

- [19] S. Teufel, A. Siddharthan, and D. Tidhar, "Automatic classification of citation function," in *Proceedings of the 2006 conference on empirical methods in natural language processing*, 2006, pp. 103–110.
- [20] D. O. Case and G. M. Higgins, "How can we investigate citation behavior? a study of reasons for citing literature in communication," *Journal of the American Society for Information Science*, vol. 51, no. 7, pp. 635–645, 2000.
- [21] T. A. Brooks, "Private acts and public objects: An investigation of citer motivations," *Journal of the American Society for Information Science*, vol. 36, no. 4, pp. 223–229, 1985.
- [22] B. Finney, "The reference characteristics of scientific texts," Ph.D. dissertation, City University (London, England), 1979.
- [23] M. Garzone and R. E. Mercer, "Towards an automated citation classifier," in *Conference of the canadian society for computational studies of intelligence*. Springer, 2000, pp. 337–346.
- [24] F. Qayyum and M. T. Afzal, "Identification of important citations by exploiting research articles metadata and cue-terms from content," *Scientometrics*, vol. 118, no. 1, pp. 21–43, 2019.
- [25] R. Ahmad, M. T. Afzal, and M. A. Qadir, "Information extraction from pdf sources based on rule-based system using integrated formats," in *Semantic web evaluation challenge*. Springer, 2016, pp. 293–308.
- [26] S. Nazir, M. Asif, S. Ahmad, F. Bukhari, M. T. Afzal, and H. Aljuaid, "Important citation identification by exploiting content and section-wise in-text citation count," *PloS one*, vol. 15, no. 3, p. e0228885, 2020.
- [27] A. Shahid and M. T. Afzal, "Section-wise indexing and retrieval of research articles," *Cluster Computing*, vol. 21, no. 1, pp. 481–492, 2018.
- [28] 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.

- [29] E. Garfield, ““ science citation index” a new dimension in indexing: This unique approach underlies versatile bibliographic systems for communicating and evaluating information.” *Science*, vol. 144, no. 3619, pp. 649–654, 1964.
- [30] H. Small, “Co-citation in the scientific literature: A new measure of the relationship between two documents,” *Journal of the American Society for information Science*, vol. 24, no. 4, pp. 265–269, 1973.
- [31] 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.
- [32] K. W. Boyack, H. Small, and R. Klavans, “Improving the accuracy of co-citation clustering using full text,” *Journal of the American Society for Information Science and Technology*, vol. 64, no. 9, pp. 1759–1767, 2013.
- [33] B. Gipp, J. Beel, and C. Hentschel, “Scienstein: A research paper recommender system,” in *Proceedings of the international conference on Emerging trends in computing (ICETiC09)*, 2009, pp. 309–315.
- [34] A. Shahid, M. Afzal, and M. Qadir, “Discovering semantic relatedness between scientific articles through citation frequency,” *Australian Journal of Basic and Applied Sciences*, vol. 5, no. 6, pp. 1599–1604, 2011.
- [35] D. M. Pennock, E. J. Horvitz, S. Lawrence, and C. L. Giles, “Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach,” *arXiv preprint arXiv:1301.3885*, 2013.
- [36] W. Paik, S. Yilmazel, E. Brown, M. Poulin, S. Dubon, and C. Amice, “Applying natural language processing (nlp) based metadata extraction to automatically acquire user preferences,” in *Proceedings of the 1st international conference on Knowledge capture*, 2001, pp. 116–122.
- [37] Y. Al Murtadha, M. N. Sulaiman, N. Mustapha, and N. I. Udzir, “Improved web page recommender system based on web usage mining,” 2011.

- [38] K. Sugiyama and M.-Y. Kan, "Scholarly paper recommendation via user's recent research interests," in *Proceedings of the 10th annual joint conference on Digital libraries*, 2010, pp. 29–38.
- [39] K. Uchiyama, H. Nanba, A. Aizawa, and T. Sagara, "Osusume: cross-lingual recommender system for research papers," in *Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation*, 2011, pp. 39–42.
- [40] K. Jack, "Mendeley: recommendation systems for academic literature," *Presentation at Technical University of Graz (TUG)*, 2012.
- [41] P. Jomsri, S. Sanguansintukul, and W. Choochaiwattana, "A framework for tag-based research paper recommender system: an ir approach," in *2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops*. IEEE, 2010, pp. 103–108.
- [42] S.-Y. Hwang, W.-C. Hsiung, and W.-S. Yang, "A prototype www literature recommendation system for digital libraries," *Online Information Review*, 2003.
- [43] J. Bollen and H. Van de Sompel, "An architecture for the aggregation and analysis of scholarly usage data," in *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, 2006, pp. 298–307.
- [44] M. Mönnich and M. Spiering, "Adding value to the library catalog by implementing a recommendation system," *D-Lib Magazine*, vol. 14, no. 5/6, pp. 1082–9873, 2008.
- [45] R. Kumar, *Research methodology: A step-by-step guide for beginners*. Sage, 2018.
- [46] E. Rich, "User modeling via stereotypes," *Cognitive science*, vol. 3, no. 4, pp. 329–354, 1979.
- [47] 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.

- [48] M. M. Kessler, “Bibliographic coupling between scientific papers,” *American documentation*, vol. 14, no. 1, pp. 10–25, 1963.
- [49] J. Beel, “Towards effective research-paper recommender systems and user modeling based on mind maps,” *arXiv preprint arXiv:1703.09109*, 2017.
- [50] J. Beel, S. Langer, G. Kapitsaki, C. Breitingner, and B. Gipp, “Exploring the potential of user modeling based on mind maps,” in *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 2015, pp. 3–17.
- [51] J. Beel, B. Gipp, and E. Wilde, “Academic search engine optimization (aseo) optimizing scholarly literature for google scholar & co.” *Journal of scholarly publishing*, vol. 41, no. 2, pp. 176–190, 2010.
- [52] S. Pohl, “Using access data for paper recommendations on arxiv. org,” *arXiv preprint arXiv:0704.2963*, 2007.
- [53] 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*, 2007, pp. 417–418.
- [54] P. Resnick, “An open architecture for collaborative filtering of netnews,” in *Proc CSCW’94*, 1994.
- [55] R. Torres, S. M. McNee, M. Abel, J. A. Konstan, and J. Riedl, “Enhancing digital libraries with techlens+,” in *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, 2004, pp. 228–236.
- [56] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, “Collaborative filtering recommender systems,” in *The adaptive web*. Springer, 2007, pp. 291–324.
- [57] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, “An algorithmic framework for performing collaborative filtering,” in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 230–237.

- [58] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.
- [59] 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.
- [60] S. M. McNee, N. Kapoor, and J. A. Konstan, "Don't look stupid: avoiding pitfalls when recommending research papers," in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, 2006, pp. 171–180.
- [61] 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*, 2002, pp. 116–125.
- [62] C. Yang, B. Wei, J. Wu, Y. Zhang, and L. Zhang, "Cares: a ranking-oriented cadal recommender system," in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, 2009, pp. 203–212.
- [63] A. Naak, H. Hage, and E. Aïmeur, "A multi-criteria collaborative filtering approach for research paper recommendation in papyres," in *International conference on e-technologies*. Springer, 2009, pp. 25–39.
- [64] I. G. Councill, C. L. Giles, and M.-Y. Kan, "Parscit: an open-source crf reference string parsing package." in *LREC*, vol. 8, 2008, pp. 661–667.
- [65] A. Vellino, "Usage-based vs. citation-based methods for recommending scholarly research articles," *arXiv preprint arXiv:1303.7149*, 2013.
- [66] S. Sosnovsky and D. Dicheva, "Ontological technologies for user modelling," *International Journal of Metadata, Semantics and Ontologies*, vol. 5, no. 1, pp. 32–71, 2010.

- [67] S. S. Sundar, A. Oeldorf-Hirsch, and Q. Xu, “The bandwagon effect of collaborative filtering technology,” in *CHI’08 extended abstracts on Human factors in computing systems*, 2008, pp. 3453–3458.
- [68] P. Lops, M. d. Gemmis, and G. Semeraro, “Content-based recommender systems: State of the art and trends,” *Recommender systems handbook*, pp. 73–105, 2011.
- [69] B. Mehta, T. Hofmann, and P. Fankhauser, “Lies and propaganda: detecting spam users in collaborative filtering,” in *Proceedings of the 12th international conference on Intelligent user interfaces*, 2007, pp. 14–21.
- [70] B. Mehta, T. Hofmann, and W. Nejdl, “Robust collaborative filtering,” in *Proceedings of the 2007 ACM conference on Recommender systems*, 2007, pp. 49–56.
- [71] B. Mehta and W. Nejdl, “Attack resistant collaborative filtering,” in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 75–82.
- [72] Y. Seroussi, “Utilising user texts to improve recommendations,” in *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 2010, pp. 403–406.
- [73] Y. Seroussi, I. Zukerman, and F. Bohnert, “Collaborative inference of sentiments from texts,” in *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 2010, pp. 195–206.
- [74] F. Esposito, S. Ferilli, T. Basile, and N. D. Mauro, “Machine learning for digital document processing: from layout analysis to metadata extraction,” in *Machine learning in document analysis and recognition*. Springer, 2008, pp. 105–138.
- [75] C. K. Shin and D. S. Doermann, “Classification of document page images based on visual similarity of layout structures,” in *Document Recognition and Retrieval VII*, vol. 3967. SPIE, 1999, pp. 182–190.

- [76] D. Buttler, “A short survey of document structure similarity algorithms,” Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), Tech. Rep., 2004.
- [77] C. D. Manning, *Introduction to information retrieval*. Syngress Publishing,, 2008.
- [78] 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*, 2011, pp. 297–306.
- [79] C. L. Giles, K. D. Bollacker, and S. Lawrence, “Citeseer: An automatic citation indexing system,” in *Proceedings of the third ACM conference on Digital libraries*, 1998, pp. 89–98.
- [80] J. Lin and W. J. Wilbur, “Pubmed related articles: a probabilistic topic-based model for content similarity,” *BMC bioinformatics*, vol. 8, no. 1, pp. 1–14, 2007.
- [81] N. Lao and W. W. Cohen, “Relational retrieval using a combination of path-constrained random walks,” *Machine learning*, vol. 81, no. 1, pp. 53–67, 2010.
- [82] L. Rokach, P. Mitra, S. Kataria, W. Huang, and L. Giles, “A supervised learning method for context-aware citation recommendation in a large corpus,” *INVITED SPEAKER: Analyzing the Performance of Top-K Retrieval Algorithms*, p. 1978, 1978.
- [83] 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*, 2010, pp. 159–166.
- [84] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, “Context-aware citation recommendation,” in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 421–430.

- [85] Y. Jiang, A. Jia, Y. Feng, and D. Zhao, "Recommending academic papers via users' reading purposes," in *Proceedings of the sixth ACM conference on Recommender systems*, 2012, pp. 241–244.
- [86] Z. Huang, W. Chung, T.-H. Ong, and H. Chen, "A graph-based recommender system for digital library," in *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, 2002, pp. 65–73.
- [87] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed-membership models of scientific publications," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl.1, pp. 5220–5227, 2004.
- [88] S. Kataria, P. Mitra, and S. Bhatia, "Utilizing context in generative bayesian models for linked corpus," in *Twenty-fourth AAAI conference on artificial intelligence*, 2010.
- [89] S. E. Middleton, D. C. De Roure, and N. R. Shadbolt, "Capturing knowledge of user preferences: ontologies in recommender systems," in *Proceedings of the 1st international conference on Knowledge capture*, 2001, pp. 100–107.
- [90] S. E. Middleton, N. R. Shadbolt, and D. C. De Roure, "Ontological user profiling in recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 54–88, 2004.
- [91] F. Ferrara, N. Pudota, and C. Tasso, "A keyphrase-based paper recommender system," in *Italian research conference on digital libraries*. Springer, 2011, pp. 14–25.
- [92] S. Pruitikanee, L. Di Jorio, A. Laurent, and M. Sala, "Paper recommendation system: A global and soft approach," in *Future computing*, 2012.
- [93] M. T. Afzal, W.-T. Balke, H. Maurer, and N. Kulathuramaiyer, "Improving citation mining," in *2009 First International Conference on Networked Digital Technologies*. IEEE, 2009, pp. 116–121.
- [94] M. T. Afzal, H. A. Maurer, W.-T. Balke, and N. Kulathuramaiyer, "Rule based autonomous citation mining with tier1." *J. Digit. Inf. Manag.*, vol. 8, no. 3, pp. 196–204, 2010.

- [95] J. Bichteler and E. A. Eaton III, "The combined use of bibliographic coupling and cocitation for document retrieval," *Journal of the American Society for Information Science*, vol. 31, no. 4, pp. 278–282, 1980.
- [96] 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*, 2007, pp. 705–706.
- [97] M. Krapivin and M. Marchese, "Focused page rank in scientific papers ranking," in *International Conference on Asian Digital Libraries*. Springer, 2008, pp. 144–153.
- [98] S. Lawrence, C. L. Giles, and K. Bollacker, "Digital libraries and autonomous citation indexing," *Computer*, vol. 32, no. 6, pp. 67–71, 1999.
- [99] S. Bonzi, "Characteristics of a literature as predictors of relatedness between cited and citing works," *Journal of the American Society for Information Science*, vol. 33, no. 4, pp. 208–216, 1982.
- [100] B.-A. Lipetz, "Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators," *American documentation*, vol. 16, no. 2, pp. 81–90, 1965.
- [101] C. Oppenheim and S. P. Renn, "Highly cited old papers and the reasons why they continue to be cited," *Journal of the American Society for Information Science*, vol. 29, no. 5, pp. 225–231, 1978.
- [102] S. B. Pham and A. Hoffmann, "A new approach for scientific citation classification using cue phrases," in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2003, pp. 759–771.
- [103] 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, pp. 25–31, 2013.
- [104] M. Reyhani Hamedani, S.-W. Kim, S.-C. Lee, and D.-J. Kim, "On exploiting content and citations together to compute similarity of scientific papers," in

- Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 1553–1556.
- [105] 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.
- [106] D. Tkaczyk and Ł. Bolikowski, “Extracting contextual information from scientific literature using cermine system,” in *Semantic Web Evaluation Challenges*. Springer, 2015, pp. 93–104.
- [107] M. F. Porter, “An algorithm for suffix stripping,” *Program*, 1980.
- [108] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 39.
- [109] Z. Gu, Y. Cai, S. Wang, M. Li, J. Qiu, S. Su, X. Du, and Z. Tian, “Adversarial attacks on content-based filtering journal recommender systems,” *Computers, Materials & Continua*, vol. 64, no. 3, pp. 1755–1770, 2020.
- [110] R. Ahmad and M. T. Afzal, “Cad: An algorithm for citation-anchors detection in research papers,” *Scientometrics*, vol. 117, no. 3, pp. 1405–1423, 2018.
- [111] P. Lopez, “Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications,” in *International conference on theory and practice of digital libraries*. Springer, 2009, pp. 473–474.