

CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD



Ranking Similar Papers based upon Section Wise Co-citation Occurrences

by

Riaz Ahmad

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

in the

Faculty of Computing

Department of Computer Science

2018

Ranking Similar Papers based upon Section Wise Co-citation Occurrences

By

RIAZ AHMAD

(PC-111004)

Dr. Atif Latif, Senior Researcher

Leibniz Information Centre for Economics, Hamburg, Germany

Dr. Nafees Ur Rehman, Senior Researcher

Konstanz University, Germany

Dr. Muhammad Tanvir Afzal

(Thesis Supervisor)

Prof. Dr. Nayyer Masood

(Head, Department of Computer Science)

Prof. Dr. Muhammad Abdul Qadir

(Dean, Faculty of Computing)

DEPARTMENT OF COMPUTER SCIENCE
CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
ISLAMABAD

2018

Copyright © 2018 by Riaz Ahmad

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

To my parents and family



**CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY
ISLAMABAD**

Expressway, Kahuta Road, Zone-V, Islamabad
Phone: +92-51-111-555-666 Fax: +92-51-4486705
Email: info@cust.edu.pk Website: <https://www.cust.edu.pk>

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the thesis, entitled “**Research Paper Recommendation by Exploiting Co-citation Occurrences in Sections of Scientific Papers**” was conducted under the supervision of **Dr. Muhammad Tanvir Afzal**. No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the **Department of Computer Science, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Computer Science**. The open defence of the thesis was conducted on **29 August, 2018**.

Student Name : Mr. Riaz Ahmed
(PC111004)

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

Examination Committee :

(a) External Examiner 1: Dr. Zahid Haleem
Associate Professor
GIKI, Topi

(b) External Examiner 2: Dr. Sharifullah Khan (TI)
Associate Professor
SEECs, NUST, Islamabad

(c) Internal Examiner : Dr. Aamer Nadeem
Associate Professor
CUST, Islamabad

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

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

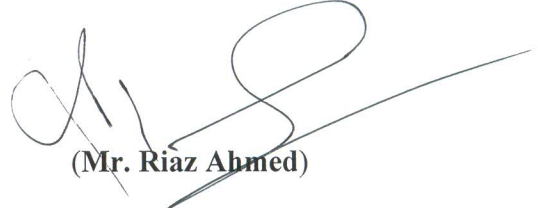
Name of Dean : Dr. Muhammad Abdul Qadir
Professor
CUST, Islamabad

AUTHOR'S DECLARATION

I, **Mr. Riaz Ahmed (Registration No. PC113003)**, hereby state that my PhD thesis titled, '**Research Paper Recommendation by Exploiting Co-citation Occurrences in Sections of Scientific Papers**' is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

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

Dated: 29 / August, 2018



(**Mr. Riaz Ahmed**)

Registration No : PC111004

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled “**Research Paper Recommendation by Exploiting Co-citation Occurrences in Sections of Scientific Papers**” is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete 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.

Dated: 29 / August, 2018



(Mr. Riaz Ahmed)

Registration No. PC111004

List of Publications

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

Journal Papers

1. **Ahmad, R.**, Afzal, M. T., & Qadir, M. A. (2017). Pattern Analysis of Citation-anchors in Citing documents for Accurate Identification of In-text Citations. *IEEE Access*, Vol (5), pp. 5819-5828.
2. **Ahmad, R.** & Afzal, M.T. (2018), CAD: an algorithm for citation-anchors detection in research papers. *Scientometrics*. Published online 29th September 2018. <https://doi.org/10.1007/s11192-018-2920-6>

Conference Papers

1. **Ahmad, R.**, Afzal, M. T., and Qadir, M. A. (2016, May). Information extraction from pdf sources based on rule-based system using integrated formats. In the semantic web: *ESWC 2016 Challenges*, Anissaras, Crete, Greece. pp. 293-308, Springer, Cham. [A Category Conference, Challenge Winner paper]
2. **Ahmad, R.**, Afzal, M. T., (2015, December). Research Paper Recommendation by exploiting co-citation occurrences in Generic Sections of Scientific Papers. PhD Symposium at 13th International Conference on Frontiers of Information Technology. Islamabad Pakistan. pp. 44-45.

Riaz Ahmad

(PC-111004)

Acknowledgements

First of all, I am thankful to Almighty Allah for granting me the health, wisdom and strength to start this PhD research work and enabling me to its completion.

Completion of this PhD thesis was possible with the support of several people. I would like to express my sincere gratitude to all of them. First of all, I am extremely owe to my research supervisor, Dr. Muhammad Tanvir Afzal, Associate Professor, for his valuable guidance, scholarly inputs and consistent encouragement I received throughout the research work. As my supervisor, Dr. Muhammad Tanvir Afzal worked closely with me during the proposal writing and during the period of my dissertation. Sir has always made himself available to clarify my doubts despite his busy schedules and I consider it as a great opportunity to do my PhD thesis under his guidance and to learn from his research expertise. Thank you Sir, for all your help and support.

I would also like to thank Professor Dr. Muhammad Abdul Qadir, the Dean of the Faculty of Computing, and Professor Dr. Nayyer Masood, Head of Computer Science Department for their support and encouragement.

Very special thanks to Dr. Muhammad Imran, my senior research fellows at CDSC, who had more confidence in me than I had in myself. They boosted my morale at every point I was feeling shaky, which was just about every other day. I am also thankful to other members of CDSC whose discussion and constructive criticism maintained an environment that was conducive for research.

There are numerous other people at C.U.S.T. who helped me in pursuit of my PhD in one way or the other including the faculty members at Department of Computer Science, managerial and support staff, and the librarian to mention a few. Thank you all.

I obliged a lot to my close friends, Mr. Ishaq Khan, Mr. Inamud din, Mr Tahir Khan and Mr Hassan Ali, for being constant source of inspiration and motivation throughout this time. There are so many other well wishers including friends,

colleagues and relations who remembered me in their prayers. Allah bless you all. I am also thankful to Prof. Mr. Muhammad Shahiq Shahid and Prof. Mr. Muhammad Amin who help me during my Ph.D study.

I owe a lot to my parents, who encouraged and helped me at every stage of my personal and academic life, and longed to see this achievement come true. I am very much thankful to my family, my wife, son and daughters, who supported me in every possible way to see the completion of this work.

Abstract

Citation indexes and digital libraries index millions of research papers and make them available to the scientific community; however, searching the intended information from these huge repositories remain a challenge. Everyday, the research papers in online digital libraries are increasing due to different number of conferences, workshop, and journals which are being arranged throughout the world. According to the statistic in 2017, one of the digital libraries in medical domain, such as PubMed consisted of 28 millions of research documents. The manual searching of relevant research papers from such a huge amount of documents is a very difficult task. Therefore, this area has attracted the attention of researcher's worldwide to propose and implement innovative techniques that could recommend relevant papers to researchers.

The identification of relevant research papers has become an important research area. For this, research community has proposed more than 90 different approaches in the past 15 years. These approaches have utilized different data sources, such as metadata, content, profile based data and citations of research papers. These techniques have certain strengths and limitations which have been critically reviewed and presented in this document.

One of the important approaches in this area is co-citation analysis which considers two documents as relevant if they are co-cited in other scientific documents. The original approach used references from the reference list of scientific documents to make such observations. However, in the recent years, the content of documents have also been exploited along with the reference list to enhance the accuracy. These approaches include Citation Proximity Analysis (CPA), Citation Order Analysis (COA), and exploit bytes of the content of scientific papers. These approaches conceptualize the occurrence of co-citations in different level of proximity and give more weights to the co-cited documents which are co-cited closely. However, the closely co-cited documents in the "Methodology/Results" section may be considered more relevant as compared to the closely co-cited papers in the "Introduction/Discussion" sections. This thesis explores structural organization

of scientific documents by giving weights according to the importance of different generic sections, and investigates that whether such approach may increase the accuracy of identifying relevant papers.

This work addresses the following important research challenges and can be considered as the contributions of the thesis: (1) generic section identification in citing document (2) in-text citation patterns and frequencies identification in citing document and (3) design of an algorithm that utilizes evidences from above mentioned sources (sections name, their weight, and the frequency of co-citations) to identify and recommend relevant papers.

For each contribution, the detailed architecture, dataset and evaluation have been discussed in this thesis. First the generic section identification component was designed, implemented and then evaluated with state-of-the-art approaches. The proposed approach was evaluated on two datasets consisted of 150 and 300 citing documents respectively. The aggregated F-score of proposed approach was 92% over the both datasets while the F-score of the state-of-the-art technique was 81%. Second, the component of in-text citation patterns and frequencies identification was implemented with detailed architecture, dataset, and evaluation. For the evaluation, two datasets were prepared from openly available digital libraries, Journal of Universal Computer Science (J.UCS)¹ and CiteSeerX². The proposed model was outperformed the state-of-the-art approach by increasing the F-score from 0.58 to 0.97. The third contribution of this thesis is section wise co-citation analysis which depends on earlier two components. The proposed approach was designed to rank the co-cited documents. For the evaluation purpose, two benchmarks such as JSD and cosine similarity based rankings were selected for the comparison of proposed and state-of-the-art approaches. The score has been compared between the proposed and state-of-the-art approaches using Spearman's and Kendall's tau measures. The results show that the proposed approach has outperformed comparatively the state-of-the-art techniques such as: standard co-citation and CPA based on bytes offset.

¹www.jucs.org/

²citeseerx.ist.psu.edu/

Contents

Author’s Declaration	v
Plagiarism Undertaking	vi
List of Publications	vii
Acknowledgements	viii
Abstract	x
List of Figures	xv
List of Tables	xviii
Abbreviations	xx
1 Introduction	1
1.1 Background	1
1.2 Basic Terminologies and Concepts	5
1.2.1 Citation	5
1.2.2 Citation Analysis	5
1.2.3 Co-citation Analysis	6
1.2.4 Co-citation Proximity Analysis	7
1.2.5 Co-citation Proximity Analysis Based on Byte-offset	8
1.2.6 In-text Citation Frequency Analysis (ICFA)	9
1.3 Research Motivation	10
1.4 Problem statement	12
1.5 Research Objectives	12
1.6 Scope of the research	13
1.7 Research methodology	13
1.8 Applications of the proposed research	16
1.9 Thesis Outline	16
2 Literature Review	17
2.1 Exploitation of IMRaD structure in Literature	17

2.2	In-text citation patterns and frequencies identification	19
2.3	Research Paper Recommendation Systems and Approaches	27
2.3.1	Research Paper Recommender Systems	28
2.3.2	Collaborative Filtering based Approaches	29
2.3.3	Metadata based Approaches	30
2.3.4	Citation Context based Approaches	33
2.3.5	Citation based Approaches	36
2.3.6	Hybrid Approaches	40
2.4	Summary	43
3	Proposed Approach Architecture	48
3.1	Data Preparation Phase	49
3.1.1	Key-Term based Crawler	50
3.1.2	Metadata Extractor	50
3.1.3	MetaDB Manager	52
3.1.4	Co-cited Pairs and Common Citing Documents Extraction	52
3.1.5	Citing papers downloader	54
3.1.6	PDF to Text and PDF to XML Convertors	54
3.2	Section Wise Co-citation Analysis Phase	54
3.3	Document Ranking and Result Evaluation Phase	56
3.3.1	Document Ranking	56
3.3.2	Result Evaluation	56
4	Identification and Mapping of Sections on ILMRaD Structure	57
4.1	ILMRaD structure Analysis	58
4.2	Proposed architecture for ILMRaD Structure Identification	60
4.2.1	Data Preparation	61
4.2.2	Structural component heading extraction phase	62
4.2.3	Structural component splitting and mapping phase	72
4.2.4	Rule Based Algorithm (RBA) for generic section identification	94
4.2.5	Generic section evaluation phase	97
4.3	Summary	101
5	In-Text Citation Patterns Identification	103
5.1	Overview of Basic Terminology	104
5.2	Pattern Analysis and Issues of Citation-Anchor	105
5.2.1	Numeric citation-tags problems	105
5.2.2	String-tags problems	108
5.3	Exploratory Analysis of GROBID AND CERMINE Tools	113
5.3.1	String Citation-anchor with Bracket problem	114
5.3.2	Citations with Same Author and Year problem	115
5.3.3	Multiple Numeric Citation-anchor with Semicolon Problem	116
5.3.4	CERMINE and GROBID tools Effected with Year Inclusion Problem	116
5.4	Proposed taxonomy of citation-anchor	118

5.5	Proposed Architecture for In-Text Citation Patterns and Frequencies Identification Approach	122
5.5.1	Data preparation phase	122
5.5.2	Automatic pattern detection of citation-anchors phase	125
5.5.3	Patterns for citation-anchors identification	128
5.6	Experimental setup	136
5.6.1	Datasets	136
5.6.2	Evaluation metrics	138
5.6.3	Results	139
5.7	Summary	143
6	Section Wise Co-citation Analysis	145
6.1	SWCA Algorithm	146
6.1.1	Dataset	147
6.1.2	Section Weights Identification	151
6.1.3	Relevancy Score (RS) Calculation	152
6.1.4	Document Ranking	155
6.1.5	Pseudo code for SWCA algorithm	156
6.2	Evaluation	158
6.2.1	Evaluation of generic section identification	158
6.2.2	Evaluation of In-text citation frequency Identification	159
6.3	Evaluation and comparison of SWCA approach with State-of-the-art approaches	160
6.3.1	Jensen-Shannon Divergence (JSD)	161
6.3.2	Content based Similarity	165
6.3.3	Co-citation Technique	169
6.3.4	Citation Proximity Analysis (Boyack et al)	170
6.3.5	Section Wise Co-citation Analysis(SWCA)	171
6.3.6	Results	172
6.4	Summary	185
7	Conclusion and Future Work	187
7.1	Conclusions	187
7.2	Contributions	188
7.3	Limitations of Proposed Approach	191
7.4	Future Work	192
	References	194

List of Figures

1.1	IMRaD structure of scientific document [23, 24]	3
1.2	Visual representation of Boyack et al approach with IMRaD structure	4
1.3	Citation Analysis of a cited document in citing documents [31]	6
1.4	Co-citation Analysis of cited-pair in citing documents [32]	7
1.5	Co-citation Analysis based on sentence level, paragraph level and article level in content of citing document [22]	8
1.6	Co-citation Analysis of cited pair in the citing document based on the chunk of Byte-offset [21]	9
1.7	In-text Citation Frequency Analysis in the content of citing document [33]	10
1.8	The methodological steps for the proposed research [44]	15
2.1	Example of reference string with citation-tag	20
2.2	Example of different formats of citation-tags in existing literature .	21
2.3	Various formats of citation-anchor in existing literature	22
2.4	Example of reference string without citation-tag	23
2.5	Citation-anchors in citing documents	24
2.6	Citation-anchor with part-of-speech (POS)	24
2.7	Mathematical ambiguity issues a) Reference string snapshot from paper b)Mathematical interval problem c)Reference string snapshot from paper and d)Mathematical parenthesis problem	27
3.1	Proposed architecture for section wise co-citation analysis	49
3.2	query paper link on CiteSeer site	50
3.3	CiteSeer link pattern with metadata information	51
3.4	Reference string extraction	51
3.5	Reference-string without citation-tag problem	52
3.6	Extracted metadata of query paper	52
3.7	Paper download link on CitSeer site	54
4.1	Proposed architecture for generic sections identification	61
4.2	Heading taxonomy for structural components	63
4.3	Analysis of section headings in both XML and Plain-text formats a) Snapshot of first level section headings in XML format b)Snapshot of first level section headings in plain-text format	65
4.4	Roman with capital case detection	68

4.5	Section heading recognition in XML document by section heading recognizer	70
4.6	Section heading conversion into structured elements	70
4.7	Structure of a research paper	71
4.8	Document structure splitting and integration	73
4.9	Snapshot of “citation-anchor patterns” from research papers	79
4.10	Snapshots of “Figure patterns” from a researcher paper	80
4.11	Snapshot of “Table pattern	81
4.12	Snapshot of “First person plural pronoun” patterns from a research paper	82
4.13	Snapshot of “Algorithm” pattern from a research paper	83
4.14	Structural components of a research paper mapped on generic Sections	86
4.15	Training dataset classification based on pages and structural components	89
4.16	Page and structural component based analysis for research papers with four pages	93
4.17	Proposed methods for section mapping	95
4.18	Aggregated precision, recall, and F-score of generic section identification for both approaches	101
5.1	Reference string, citation-tag and citation-anchor relationship	105
5.2	Mapping of numeric citation-tag on multiple citation-anchors	106
5.3	Mapping of numeric citation-tag on range citation-anchors	106
5.4	Incorrect citation-anchor due to mathematical ambiguity. a) Snapshot of reference or citation string with numeric-tag b)Content snapshot with valid and invalid citation-anchors for numeric citation-tag	107
5.5	Citation-tag mapping with compound citation-anchor	108
5.6	Format problems with one author, two authors and multiple authors anchor’ cases a) One-author case b) “&” symbol problem in two-authors case c)“et al” problem in multiple authors case	109
5.7	Carriage return and line feed problem	109
5.8	Year related problems a) Year format problem b)Year inclusion problem	110
5.9	Citation-anchor with space character problem	111
5.10	Citation-anchor with POS (part-of-speech) problem	111
5.11	Reference-string without citation-tag problem	112
5.12	Common character as Citation-anchor	112
5.13	Reference-string with superscript citation-anchor problem	113
5.14	CERMINE tool with String Citation-anchor with Bracket Problem a) Reference String with String Citation-tag with Bracket in Text and XML formats b)The Missed String Citation-anchors	114
5.15	Citations with Same Author and Same Year Problem a) Reference String in Text and XML formats b)CERMINE tool Assigned the Wrong Reference ID to Citation-anchors	115

5.16	Missed Citation-anchors with GROBID tool due to Same Author and Year Problem	116
5.17	Multiple Numeric Citation-anchor with Semicolon Problem a) CER-MINE: Missed Multiple Numeric Citation-anchor b)GROBID:Missed Multiple Numeric Citation-anchor	117
5.18	Missed Citation-anchors with Year Inclusion Problem	117
5.19	Citation-anchor taxonomy	121
5.20	Proposed architecture for citation anchor detection	122
5.21	Metadata of cited and citing documents	124
5.22	Reference string extraction	126
5.23	Numeric citation-tag extraction	127
5.24	Regular expressions for citation-anchors identification	130
5.25	Precision, Recall, and F-score of both approaches over J.UCS dataset	140
5.26	Precision, Recall, and F-score of both approaches over CiteSeerX dataset	142
5.27	Comparison of Proposed approach with State-of-the-art Approach and Tools over CiteSeer Dataset	143
6.1	Proposed architecture for SWCA(Section wise co-citation analysis) .	146
6.2	The real snapshot of query papers from CiteSeerX site	148
6.3	The real snapshot of citations of query paper from CiteSeerX site .	149
6.4	Visual representation of Equation 3.1	150
6.5	The real snapshot of co-cited documents with a query paper from CiteSeerX site	150
6.6	Precision, Recall, and F-score of generic section Identification over CiteSeer dataset	159
6.7	Precision, Recall, and F-score of In-text citation frequency Identification over CiteSeer dataset	160
6.8	Proposed approach comparison with State-of-the-art approaches based on JSD ranking a) Average Correlation with JSD @ 3 b)Average Correlation with JSD @ 5 c)Average Correlation with JSD @ 7 d)Average Correlation with JSD @ 9	180
6.9	Comparison of Proposed technique with State-of-the-art techniques for different set of queries	181
6.9	Proposed approach comparison with State-of-the-art approaches based on Cosine ranking a) Average Correlation with Cosine @ 3 b)Average Correlation with Cosine @ 5 c)Average Correlation with Cosine @ 7 d)Average Correlation with Cosine @ 9	184
6.10	Comparison of Proposed technique with State-of-the-art techniques for different set of queries	185

List of Tables

2.1	Summary of reviewed literature	45
3.1	Key-Terms for query papers searching	50
4.1	Manual classification of section labels of structural components [28]	59
4.2	Manual classification of section labels over 211 research papers . . .	60
4.3	Heading analysis of structural components based on formats	64
4.4	Structural components offset dataset of a research paper	72
4.5	Key and Stemming words selection over training dataset of 211 research papers for heading label based analysis	75
4.6	Generic sections identification based on stemming words in 211 training dataset of research papers	76
4.7	Structural components mapping on generic sections	77
4.8	Training dataset for pages and structural components based analysis	88
4.9	Sequence patterns of Generic Sections in first subset of 4 pages Research Papers	90
4.10	Position frequency matrix (M_1)	91
4.11	Position probability matrix (M_2)	92
4.12	Sequence patterns with probabilities	93
4.13	Training and testing datasets for generic section identification task .	97
4.14	Confusion matrix of proposed approach for 50 papers in testing dataset1	98
4.15	Confusion matrix of State-of-the-art approach for 50 papers in test- ing set1	98
4.16	statistical data of proposed approach over testing dataset1	99
4.17	Statistical data of state-of-the-art technique over testing dataset1 .	100
4.18	Statistical data of proposed technique over testing dataset2	100
4.19	Statistical data of state-of-the-art technique over testing dataset2 .	101
5.1	Key-Terms for the selection of cited documents	123
5.2	Statistics of Datasets	137
5.3	CiteSeerX dataset specifications	137
5.4	Statistics of CiteSeerX Extended dataset	138
5.5	Frequency distribution of in-text citations in J.UCS Dataset	140
5.6	Frequency distribution of in-text citations in CiteSeerX dataset . .	141
6.1	Dataset of query paper,co-cited paper, and citing documents	152

6.2	One co-cited pair of research papers with three citing documents . . .	153
6.3	Co-citation frequencies and relevancy score (RS)	154
6.4	The Cumulative relevancy score of nine co-cited pairs	155
6.5	The cumulative relevancy score of nine co-cited pairs	156
6.6	Confusion matrix for generic sections identification over 150 papers	158
6.7	Cluster of documents	162
6.8	Word count and probability vectors for each document and cluster	163
6.9	Mean of ‘p1’, ‘p2’, and ‘p3’ with ‘q’ distribution	163
6.10	Kullback Leibler Divergence for ‘p’ and ‘q’	164
6.11	Ten rankings prepared for ten clusters of documents based on Di- vergence measure	165
6.12	Collection of text documents	166
6.13	Document TFV with tf-idf score	167
6.14	Terms with ‘tf-idf’ scores in d_1 , d_2 , and d_3	168
6.15	Ten rankings prepared for ten clusters based on cosine similarity score	169
6.16	Ten ranking prepared for ten cluster of documents based on Co- citation measure	170
6.17	Ten rankings prepared for ten clusters of documents based on Prox- imity measure	171
6.18	Ten rankings prepared for ten clusters based on Relevancy Score in SWCA approach	171
6.19	The ranking dataset of single cluster for proposed approach, state- of-the-art approaches, and JSD approach	173
6.20	Spearman rank correlation between JSD Vs Co-citation ranks . . .	173
6.21	Spearman rank correlation between JSD Vs Boyack et al ranks . . .	174
6.22	Spearman rank correlation between JSD Vs SWCA ranks	175

Abbreviations

ILMRAD	Introduction, Literature, Methodology, Result and Discussion
IA-STMP	International Association of Scientific, Technical and Medical Publishers
PSCA	Page and Structural Components Analysis
POS	Part of Speech
CPA	Citation Proximity Analysis
COA	Citation Order Analysis
CPI	Citation Proximity Index
ICFA	In-text Citation Frequency Analysis
MIR	Multiple In-text References
SWCA	Section Wise Co-citation Analysis
CPP	Co-cited Pairs
ESWC	European Semantic Web Conference
INTR	Introduction
LITR	Literature
MET	Methodology
RES	Result
DESC	Discussion
CON	Conclusion
GS_ID	Generic Section Identifier
RBA	Rule Based Algorithm
TP	Total Pages
SC	Structural Components
PFM	Position Frequency Matrix
PPM	Position Probability Matrix

SPM	Sequence Probability Matrix
CAD	Citation Anchors Detection
N-CAD	Numeric Citation Anchors Detection
S-CAD	String Citation Anchors Detection
TCD	Text of Citing Document
CT	Citation-Tag
SNAP	Single Numeric Anchor Pattern
MRCP	Multiple Range and Compound Patterns
RS	Relevancy Score
CRS	Cumulative Relevancy Score
KLD	Kullback Leibler Divergence
JSD	Jensen-Shannon Divergence
TFV	Term Frequency Vector
C-CIT	Correct Citations
IC-CIT	In-correct Citations
ZO	Zero Occurrences
CD	Citing Documents
CIT-CD	Citations Of Cited Documents
CIT-WNT	Citation With Numeric Tag
CIT-WST	Citation With String Tag
CIT-WCT	Citation Without Citation Tag

Chapter 1

Introduction

The flow of chapter is as follows: it covers the background and basic terminologies of co-citation analysis for identification of relevant documents. It is followed by the research motivation. The critical analysis of literature has led us to form the problem statement and research objectives which is explained after research motivation. Finally, chapter concludes with the methodology adopted for conducting this research and in the end of this chapter, the thesis outline is also presented.

1.1 Background

The publication and availability of scientific knowledge is increasing with great pace. Sometimes it is referred that the volume of knowledge is doubling every five years time [1, 2]. The major part of documents corpus consists of research articles due to continuous discoveries and inventions in science [3]. According to the recent IA-STMP Report [4], a variety of more than 10,000 publishers has collectively published more than 30,000 journals, representing millions of individual articles published to date. Citation indexes and scientific search systems index millions of research articles [5]. The identification of pertinent resources from these huge repositories becomes a challenging task [6, 7]. This has attracted scientific community to propose and implement state-of-the-art approaches in this area. Recently,

the research literature on research paper recommendation was reviewed critically by Beel et al [8, 9]. They highlighted 96 existing approaches in 217 research papers on the area of paper recommender system which were developed based on profile [10, 11], metadata [12–14], citation context [15, 16], citations [17, 18] and hybrid approaches [19, 20].

Beel et al [9] has recently identified that content based approaches remained dominant in the literature for research paper recommender systems. They have also identified that the citation-based approaches have a potential to identify the candidate relevant research documents because authors manually pick citations from literature when they are preparing their research work. The state-of-the-art technique presented by Boyack et al [21] combines the information of both content and co-citations to judge the relevancy and similarity between research documents. Their technique is the extension of Citation Proximity Analysis (CPA) [22].

In Boyack et al approach [21], the whole research paper document is considered as a set of bytes. To find relevancy between two co-cited papers, the byte offset between the citation-anchors of the two papers is calculated and a weight is assigned accordingly. If the byte offset between the citation-anchor positions of two co-cited papers A and B is 375, 1500, 6000 and over 6000, then the weights assigned will be 3, 2, 1 and 0 respectively. The byte offsets such as 375, 1500, and 6000 are ways to approximate the lengths of sentences, paragraphs, and sections without using the actual sentence structure, such as used in CPA [22]. They considered the average sentence length as 375 bytes and so the byte offsets 1500 and 6000 were considered equal to 4-16 sentences respectively.

Boyack et al [21] approach has a major shortcoming which can be highlighted with the help of two scenarios. In the first scenario, an author cites two papers A and B to provide introduction and background of his research and the byte offset between these two papers is 375 bytes. It means that the weight of two papers A and B is 3 which shows that the two co-cited papers are more relevant papers. It might be the case that these two papers are not more relevant to each other because the author has just cited these two papers of different domains for background study.

In another scenario, an author cites two papers A and C to conclude the result in his research paper and the byte offset between these two papers is 16 sentences. It means that the weight of 1 will be given to these papers. It is intuitive to consider that in this case both papers might be more relevant than the scenario mentioned earlier.

Therefore, it can be concluded from the above two scenarios that the assigned weights of two pairs (A, B) and (A, C) might not be feasible because the research papers follow the proper structure. Normally, in the structure of research paper, firstly authors discuss the background of research topic in their work. Secondly they explain the whole methodology of their research work. Thirdly, the findings of their experiments are discussed in result and then, finally the authors wind up the conclusion in the discussion. This structure exists for many years and is known as IMRaD (Introduction, Methodology, Result, and Discussion) [23, 24] as shown in Figure 1.1.

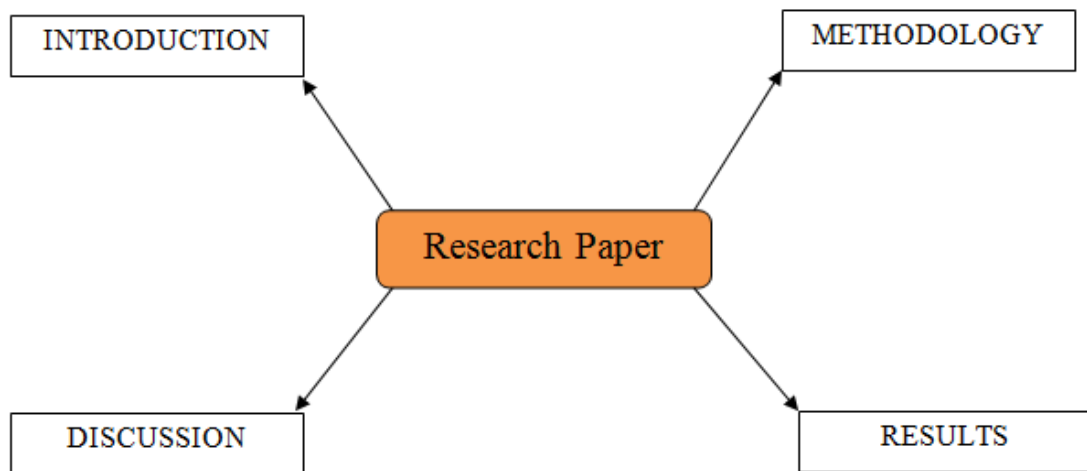


FIGURE 1.1: IMRaD structure of scientific document [23, 24]

In fact, during the last decades, IMRaD has imposed itself as a standard rhetorical framework for scientific articles in the experimental sciences [25]. Different authors [6, 26–28] have shown the significance and importance of using a research paper’s logical sections for finding relevant documents.

Assume a scenario to show a detailed example in Figure 1.2. The pairs (A, B) and (A, C) are co-cited in “Introduction” and “Methodology” sections respectively.

Generally, papers are cited in “Introduction” section just for background study of approaches. Therefore, it might be possible that the papers A and B are not closely related with each other. In this scenario, the Boyack et al [21] approach assigns the highest weight of 3 to both papers A and B due to minimum number of bytes offset, i.e., 375 between them. In the second scenario, the author cites the papers A and C in the Methodology section in the citing paper. It means that these two papers might be closely related with each other based on methods. In such case, the approach assigns the less weight of 1 to both papers A and C due to the maximum number of byte offset (6000 bytes). Therefore, it is concluded from the given scenario in Figure 1.2 that the IMRaD structure of research papers should be exploited for co-citation analysis to recommend relevant research papers instead of just relying on statistical distribution of bytes and sentences.

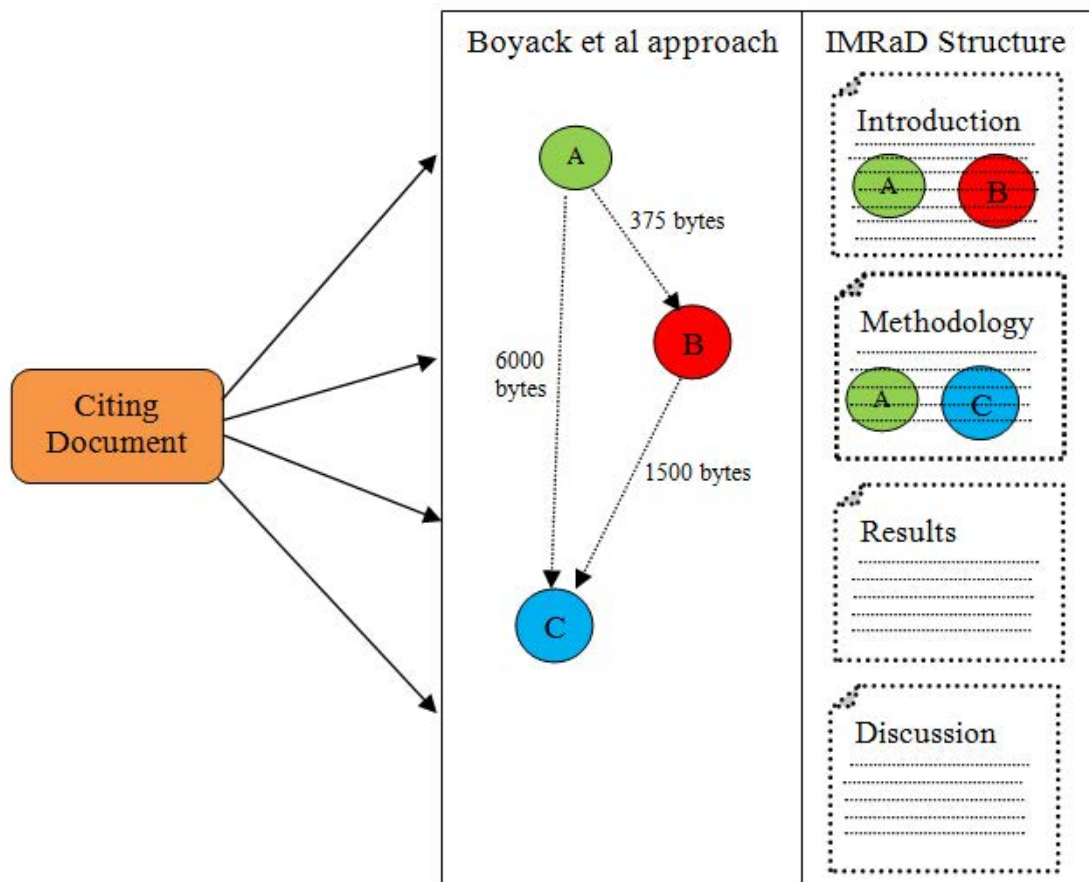


FIGURE 1.2: Visual representation of Boyack et al approach with IMRaD structure

1.2 Basic Terminologies and Concepts

In this section, we have discussed some of key terminologies and concepts to understand the proposed approach in this research work.

1.2.1 Citation

A citation is an explicit connection in citing documents in a published or unpublished research work. More specifically, a citation is an abbreviated alphanumeric expression embedded in the body text of citing documents that denotes a reference string in the bibliographic section of the research work for the purpose of recognizing the relevance of the research works of other researchers to the topic of discussion at the spot where the citation appears [29]. Generally the citation is prepared by the combination of both the in-text citation-anchor(i.e Liu2014) and the reference strings. Citations allow authors to refer to past research in a formal and highly structured way [30].

In the below part of this section, different types of citation-based analysis are shown with proper examples.

1.2.2 Citation Analysis

Initially in the citation analysis, the reference strings of citations are only analyzed in the bibliography section of the citing documents [31]. The importance of citations was not considered in the body of the citing document. This type of citation analysis is also called direct citation. In the direct citation,i.e., the cited document is directly cited into the citing document. For example, in Figure 1.3, the cited document A which is published in 2000. This document is cited in the bibliography sections of the three cited documents,i.e., A, B, and C with published years 2003, 2006 and 2008 respectively. The citation count measure is also calculated based on the citation analysis. For example, the citation count of document A in Figure 1.3 is 3 because the document A is cited by three citing documents.

Citation count is also called a dynamic measure because the citation count of a particular paper may be increase with the passage of time.

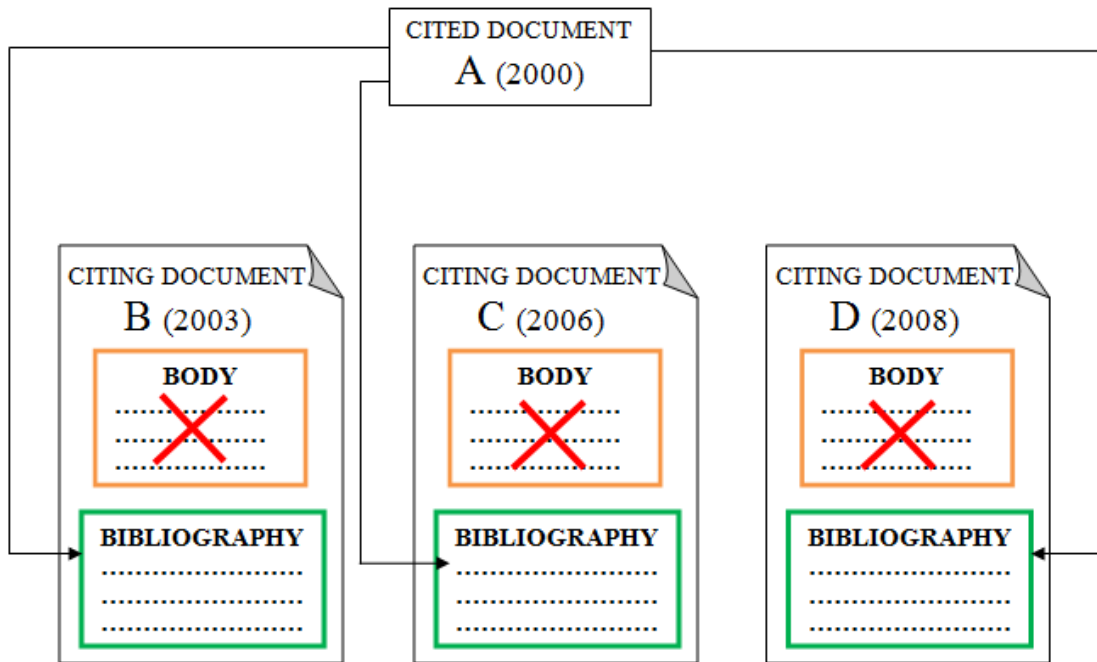


FIGURE 1.3: Citation Analysis of a cited document in citing documents [31]

1.2.3 Co-citation Analysis

Co-citation analysis [32] considers two cited documents similar if both have been cited in the bibliography section by one or more citing documents. For example in Figure 1.4, the both cited documents D and E are cited together in the bibliography section of the citing documents A, B and C respectively. In this way, the co-citation strength of two co-cited documents D and E will be 3. In conventional co-citation analysis, the content of the citing document is not considered for the recommendation of the research paper.

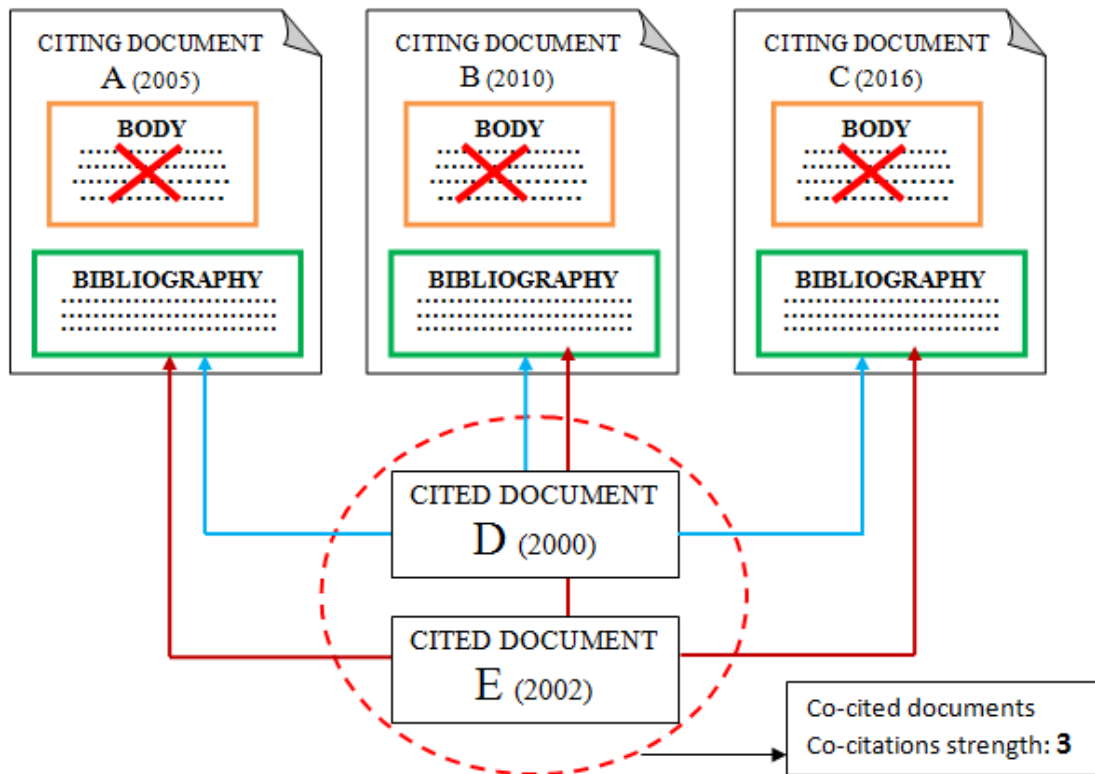


FIGURE 1.4: Co-citation Analysis of cited-pair in citing documents [32]

1.2.4 Co-citation Proximity Analysis

Co-citation proximity analysis [22] is a further extension of co-citation analysis. In this analysis, the proximity or distance of citations is analyzed to each other within full-text of a citing document. If the two citations occur closer to each other in the full-text document, then these citations will be considered that they are related. The measure CPI (Citation Proximity Index) is used to check the similarity between two co-cited documents. If for example two citations are given in the same sentence the probability that they are very similar is higher ($CPI = 1$) as if they were only in the same paragraph ($CPI = 1/4$). For example in Figure 1.5, paper B and C are more related because they are cited by the paper A at sentence level.

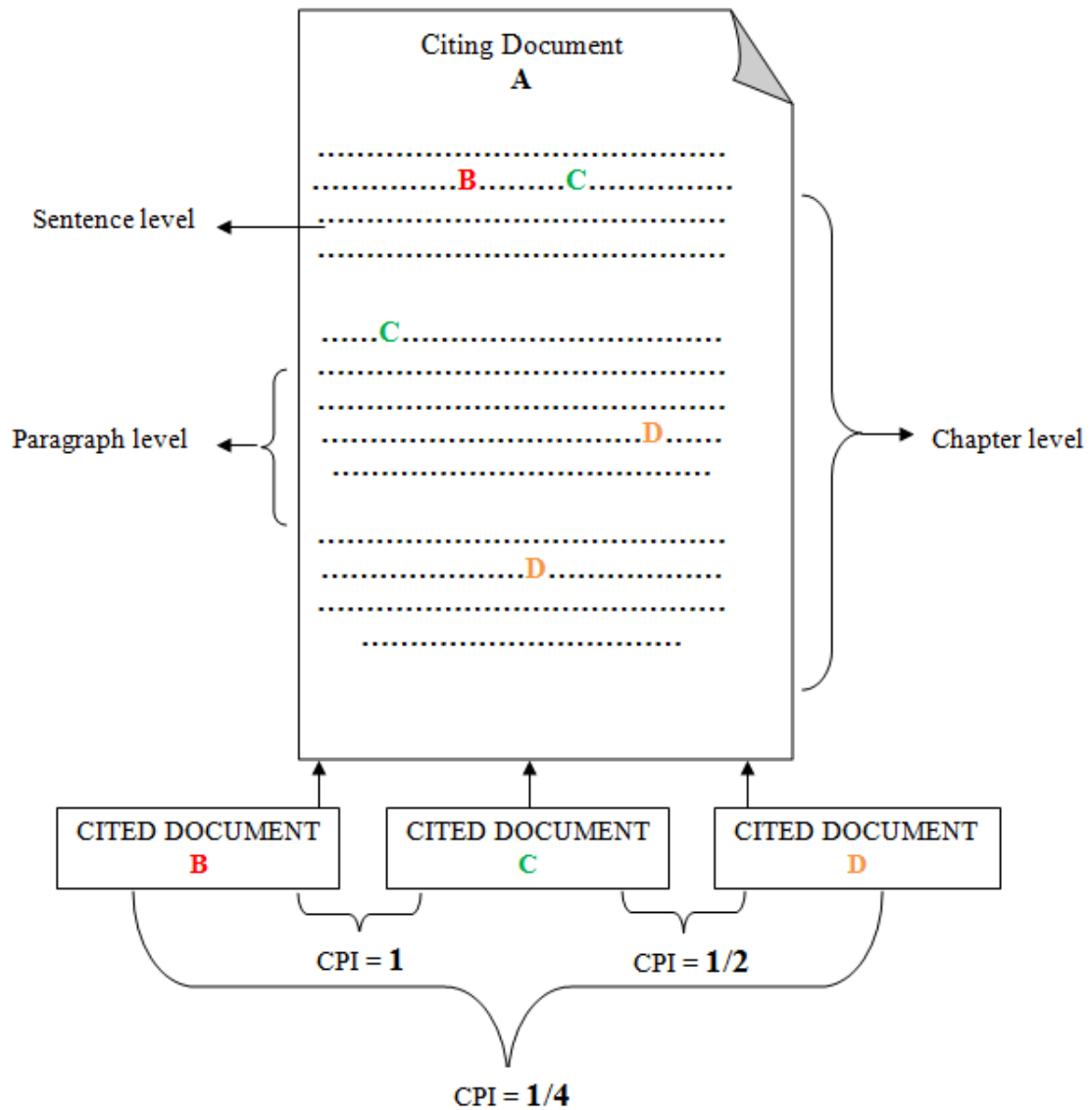


FIGURE 1.5: Co-citation Analysis based on sentence level, paragraph level and article level in content of citing document [22]

1.2.5 Co-citation Proximity Analysis Based on Byte-offset

Boyack et al [21] performed the co-citation proximity analysis based on byte-offset in a full-text document. They analyzed the citations into different size of byte chunks such as 375, 1500, and 6000 with the assigned weights 3, 2 and 1 respectively. For example in Figure 1.6, the five cited documents B, C, D, E and F are cited in text of a full-text citing document A. Here, we have shown four pairs of cited documents such as (B,C), (B, D), (B, E) and (B,F). The citations B, C

in pair (B,C) that are within the same bracket, a weight of 4, while citation pairs i.e (B, D), (B, E), and (B, F) within 375, 1500, and 6000 bytes are given weights of 3, 2, and 1 respectively. Citation pairs that are more than 6000 bytes apart are given a weight of zero.

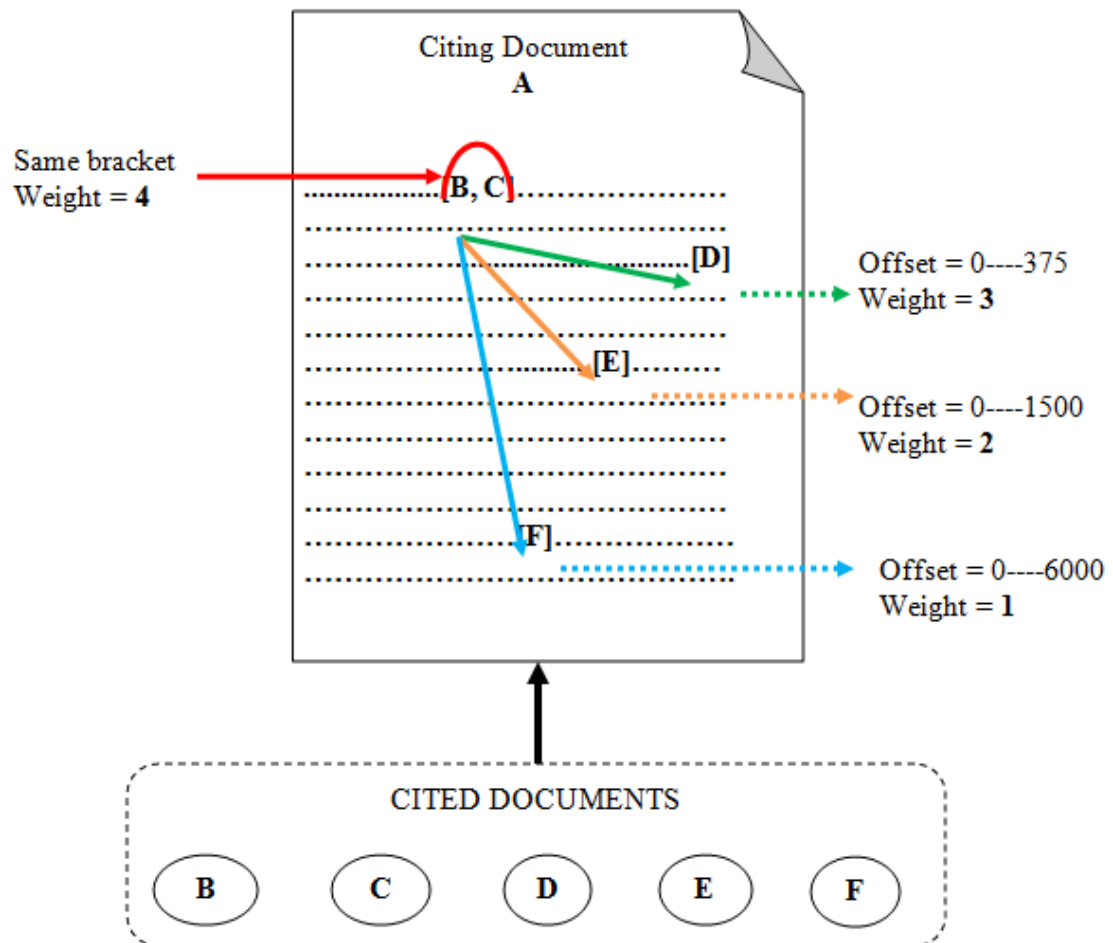


FIGURE 1.6: Co-citation Analysis of cited pair in the citing document based on the chunk of Byte-offset [21]

1.2.6 In-text Citation Frequency Analysis (ICFA)

Initially, the new measure intext citation frequency was introduced by Gipp et al [33]. Recently, the Shahid et al [34] have also used this measure to find the relationship of citations across the sections of citing documents. ICFA analyses the frequency with which a research paper or article is cited within the citing document. In Figure 1.7, the three cited documents B, C, and D are cited in

citing document A. The in-text citation frequency of cited document B is 4 which shows the strong relationship with document A.

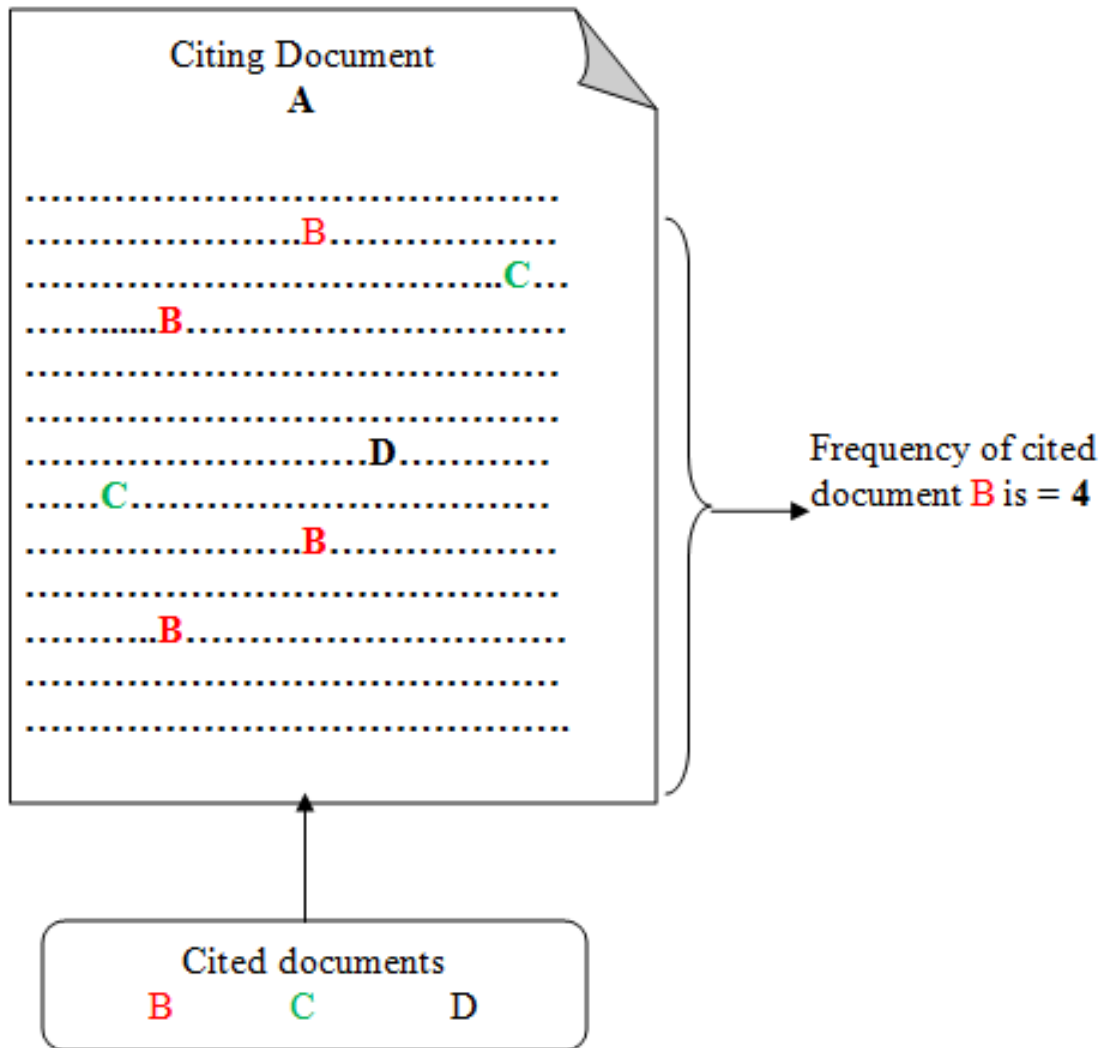


FIGURE 1.7: In-text Citation Frequency Analysis in the content of citing document [33]

1.3 Research Motivation

This section presents an overview of citation analysis, co-citation analysis, co-citation analysis based on proximity (CPA), and in-text citation frequency analysis, for better understanding of the domain. Citations have been used as an important evidence to recommend relevant research papers using a number of approaches, such as bibliographic coupling [35], citation count [31], co-citation

analysis [32], and citation context [36]. Different co-citation models have been proposed in the literature. The foundation work of co-citation analysis was proposed by Small [32]. The philosophy of their proposal was to consider paper A as relevant to paper B, provided that paper A and paper B have been co-cited in many other scientific documents.

The idea of co-citation analysis was extended by different authors using text of citing papers. Gipp and Beel [22] evaluated the co-citation position in the text of citing documents based on proximity using co-citation weights, such as 1, 1/2, 1/4, and 1/8. The citation proximity analysis increased the accuracy of co-citation by 55% [22]. The order (occurrence sequence) of co-cited papers was also exploited by Gipp and Beel [37].

Boyack et al [21] distributed the full-text document into different size of byte chunks such as 375, 1500, 6000 with the assigned weights 4, 3, 2, 1, and 0 respectively. If the numbers of bytes between the occurrence positions of co-cited papers is greater than 6000, weight of zero is assigned. Above approaches used a variety of ways to exploit the content of scientific papers and extended co-citation analysis to recommend relevant research papers.

Both Gipp and Beel [22] and Boyack et al [21] studies do not consider co-citation analysis with semantic evidences. They have only statistically analyzed the co-citation distribution and proximity based on the number of occurrences and number of bytes. Furthermore, the proximity based co-citation analysis has some inherited limitations. For example, consider two papers A and B co-cited ten times in the text where the author was only introducing the readers to the overall domain (e.g., in the introduction section) in a citing paper, In another case, two papers A and C co-cited five times in the text where authors were concluding their findings (e.g., in the result section) in a citing document. In such a case, Paper A and B might not be relevant as compared to the papers A and C as was mentioned in Figure 1.2.

For such section based analysis, IMRaD structure is well known structure in the scientific community [25, 28] and has been utilized for different purposes by scientific community [18, 24, 38], Therefore, the IMRaD structure of research papers should be analyzed for co-citation analysis to recommend relevant research papers. Different authors [6, 26, 27] have shown the significance and importance of using sections for finding relevant documents in a paper recommender systems. This has motivated the author of this thesis to systematically explore this area.

1.4 Problem statement

Based on the research motivation in the previous section, this thesis has focused on the following three research problems.

1. The accuracy of structural components mapping on ILMRaD structure is 78% in the recent approach [28]. This need to be improved.
2. The accuracy of in-text citation patterns and their frequencies is just 58% in the state-of-the-art approach [18]. This need to be approved.
3. The exiting state-of-the-art co-citation approach [21] has used the statistical measure ,i.e., bytes offset as illustrated in Figure 1.6 in the content of the citing documents for the ranking of relevant research documents. They do not consider the structural measure of the citing document. First we will solve the above two problems and then we will develop such approach for the co-citation analysis which will use the structural measure instead of statistical measure in the content of citing documents.

1.5 Research Objectives

Our first research objective is to improve the accuracy of ILMRaD structure identification by analyzing different patterns in the contents of the citing document instead of the section label as used in the previous approach [28].

In the second research objective, first we analyze the previous approaches for in-text citation frequency identification, existing standard formats of in-text patterns, and also conduct new experiment to make new rules and heuristics. These rules and heuristics will identify all those patterns of in-text citation-anchor in the citing documents which are not properly detected by the exact matching as discussed in state-of-the-art approach [18]. Based on these rules and heuristics, we will develop the complete approach for the in-text pattern and their frequencies identification.

The third and final research objective is to develop the approach of co-citation analysis which will use the structural measure and the co-citation frequencies in the citing document to rank the relevant research papers.

1.6 Scope of the research

Citation analysis is an important domain in the field of research and development. Citation analysis, other than recommending related scientific research documents has been used for different purposes, such as finding relationship between authors [39–41], and measuring influence of a journal [30, 42, 43]. The scope of the current research is to evaluate whether the co-citation of two or more documents in different generic sections can be used to improve the ranking of relevant documents. The aim of this research work is to develop a state-of-the-art Co-citation analysis technique for research documents. The proposed system does not focus on text similarity or metadata of documents to find out the relatedness among the scientific documents. It focuses on exploring co-citing patterns and co-citation frequencies of in-text citation tags in various generic sections of a citing document.

1.7 Research methodology

For conducting this research, the three-phase, eight-step model has been followed as proposed by Kumar [44] with slight modifications as per the requirements of

this research. The activities carried out during the course of this research are described below, and a mapping between these activities and Kumar's model is also highlighted in Figure 1.8.

Phase I: Deciding what to Research

Step 1: Research Problem: This step consists of three tasks (1) Literature review (2) Research gap identification, and (3) Research problem formulation.

Phase II: Planning to Research Study

Step 2: Proposed Approach Architecture: In this step, first we have proposed the novel approach "Section Wise Co-citation Analysis (SWCA)" based on *step 1* and then designed the proposed methodology for the conducting the suggested approach.

Step 3: Data Collection Method: In *step 3*, the automatic tool is designed to achieve research documents collection.

Step 4: Sample Selection: In this step, we selected randomly the sample of research documents from research documents collection that was achieved in *step 3*.

Step 5: Synopsis: In this step, we have prepared the synopsis document after the initial experiment in this research work.

Phase III: Implementation of Research Study

Step 6: Dataset pre-processing: This step is used to prepare the comprehensive datasets of semi-structured research papers documents with required pre-processing.

Step 7: Evaluation and Results: In this step, the result of proposed approach will be evaluated and discussed with state-of-the-art approaches.

Step 8: Thesis: This is the last step of our research methodology in which we have prepared the thesis document.

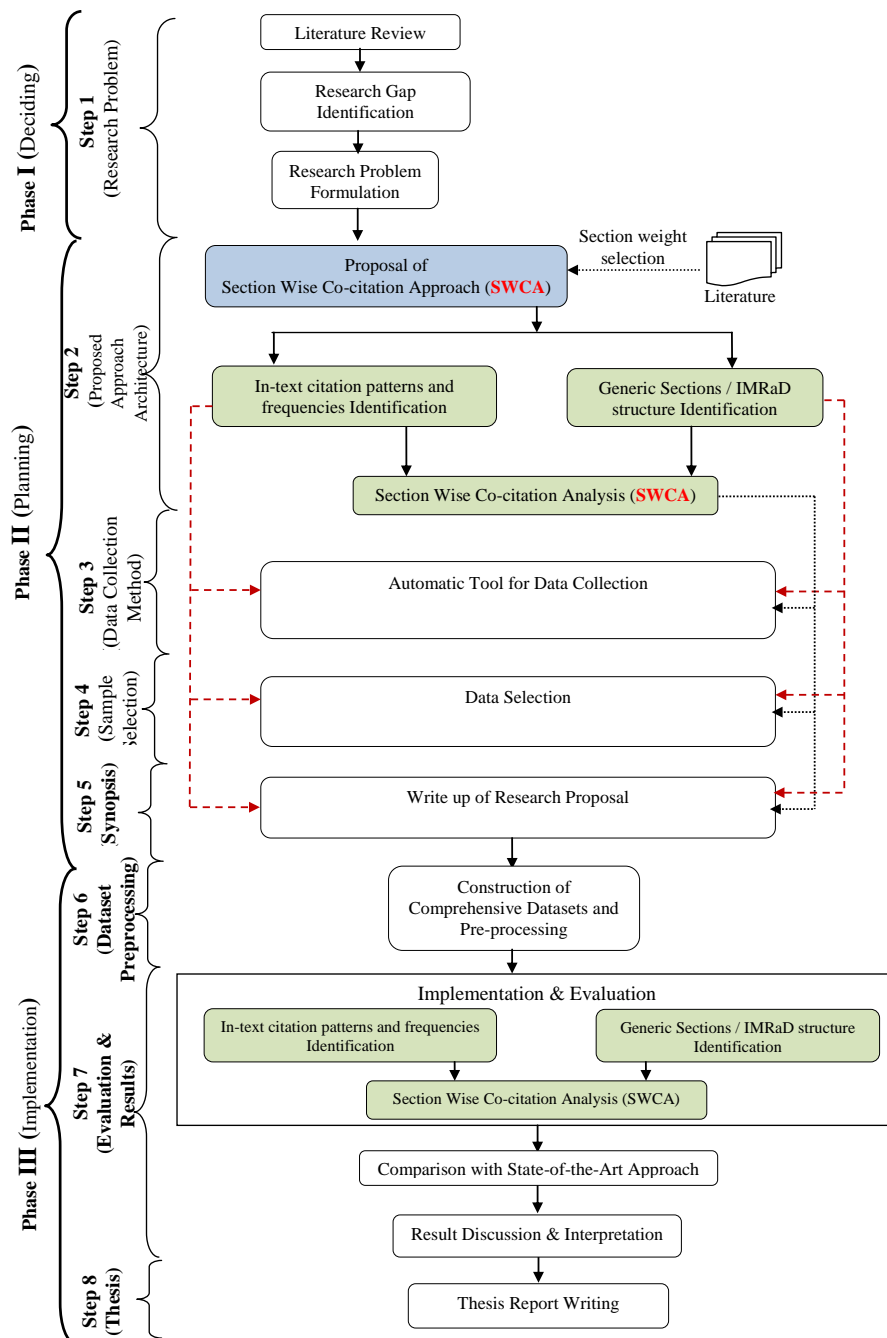


FIGURE 1.8: The methodological steps for the proposed research [44]

1.8 Applications of the proposed research

The proposed research work can be utilized in various application domains and contexts. Some of them are given below:

1. Digital libraries (ACM, IEEE, Springer etc),
2. Citation Indexes (Google scholar, ISI Web of Knowledge, CiteSeerX),
3. Conference and Journals etc.

1.9 Thesis Outline

This dissertation consists of seven chapters. In Chapters 1 and 2, Introduction and literature about proposed research work have been discussed respectively. In Chapter 3, the architecture of the proposed approach SWCA has been elaborated along with main contributions or research tasks. These research contributions are (1) ILMRaD structure identification (2) in-text citation patterns and their frequencies identification and (3) section wise co-citation analysis. These three contributions shows three research problems. Each of these research problems are comprehensively discussed in Chapters 4, 5, and 6 respectively. In the last chapter, conclusions, limitations and future work of our proposed approach are discussed.

Chapter 2

Literature Review

In this chapter, the literature survey and critical analysis is carried out to understand the scope and importance of those three tasks (1) IMRaD structure and section mapping, (2) in-text citations identification, and (3) research paper recommender systems. The following sections present detailed literature review and the current state-of-the-art in all three dimensions in which this thesis has made contributions.

2.1 Exploitation of IMRaD structure in Literature

The organization of scientific papers typically follows a standardized pattern, the well-known IMRaD structure (introduction, methods, results, and discussion) [24]. The idea that the section structure of papers plays an important role in determining the function and importance of citations was first developed by McCain and Turner [45]. To some extent, citation location can reveal the citation motivation. If we are aware of the section where a citation is located, the role of the citation can be figured out to some extent [38]. The Introduction section explains the scope and objective of the study in the light of current knowledge on the subject; the Materials and Methods describes how the study was conducted; the Results section

reports what was found in the study; and the Discussion section explains meaning and significance of the results and provides suggestions for future directions of research [23].

Recently, different authors have exploited IMRaD structure for different purposes. In fact, during the last decades, IMRaD has imposed itself as a standard rhetorical framework for scientific articles in the experimental sciences [24]. In 1998, Maricic et al [46] studied a collection of 357 papers focusing on three components: locations of references, levels of citation, and age. They suggested that if the section structure is derived from publishing practices, it also reflects the structure of scientific papers. As a result, references have different values according to their location, that is, the section in which they appear. To express these differences they assigned weights to the different sections using a ranking scale (Introduction: 10, methods: 30, results: 30, discussion: 25). Bertin & Iana [47] presented a large-scale approach for the extraction of verbs in reference contexts. They analyzed citation contexts in relation with the IMRaD structure of scientific documents and used rank correlation analysis to characterize the distances between the section types. The results show strong differences in the verb frequencies around citations between the sections in the IMRaD structure.

Bertin and Iana considered sentences that contain multiple in-text references (MIR) and their position in the rhetorical structure of articles. Different authors [6, 26, 27] and Shahid and Afzal [28] have shown the significance and importance of using sections for finding relevant documents. Hu et al [48] visualized and analyzed the distributions of citations in articles that are organized in a commonly seen four-section structure, namely, introduction, method, results, and conclusions (IMRC). They measured the proportion of each section by height of blocks. Usually the first and the last sections occupy the lowest shares of the full text. In the 4-section articles, for example, the proportions for each section from first to last are 20.8, 31.5, 35.9 and 11.9%. Ding et al [49] performed an analysis of citations in 866 articles from the *Journal of the American Society of Information Science and Technology*. They studied the number of times each citation was cited across sections and obtained citation frequencies per section.

In the most recent study of Bertin and Iana [25], the references distribution are analyzed in the structure of scientific papers as well as the age of these cited references express the negational citations. They identified the section structure in each article by analyzing the section titles, in order to identify the four main section types in the IMRaD structure (Introduction, Methods, Results, and Discussion). More than 97% of all research articles in the corpus contain these four section types.

In Shahid and Afzal [28] approach, 329 papers were randomly selected from the total of 1,200 documents. The total 1833 sections were extracted from 329 research papers. The section “Introduction” was noted as the most compliant section ,i.e., in 78% of the documents, the section “Introduction” was referred with the same names. However, the section “Methodology” was not referred even a single time with the term “Methodology”. The section “Related Work” was referred with the same or similar terms as “Related Work” in only 30% of the documents. The section “Results” was mentioned with the term “Results” only by 1% of the documents. The system was evaluated based on well-known measure of precision and recall. Precision and recall values were computed for each standard section ,i.e., “Introduction”, “Related Work”, “Methodology”, “Results”, “Discussion” and “Conclusion”. The overall F1 measure score received is 0.78%.

2.2 In-text citation patterns and frequencies identification

A citation is an explicit connection in citing documents to a published or unpublished research work. More specifically, a citation is an abbreviated alphanumeric expression embedded in the body text of citing documents that denotes a reference string in the bibliographic section of the research work for the purpose of recognizing the relevance of the research works of other researchers to the topic of discussion at the spot where the citation appears [29]. Generally the citation is prepared by the combination of both the in-text citation-anchor “Liu2014” and

the reference strings. Citations allow authors to refer to past research in a formal and highly structured way [30]. It has been used for knowledge diffusion studies [50], network studies, and in finding relationships between documents [32]. Impact factor measurements, as derived from citation counts have been applied in making important decisions hiring, tenure decisions, promotions and the award of grants [51].

The reference string of each citation in the citing paper contains citation tags “[1], 1, (Author, 2000)”, and metadata like authors name, title, and year. The approaches [18, 22, 52] were developed using citation tag and the citation anchor. When the citation tag is cited in the text of the citing paper, it is called citation anchor. The red circle shows the citation tag of the reference string while the green circle shows the reference or citation anchor inside the text of document as shown in Figure 2.1.

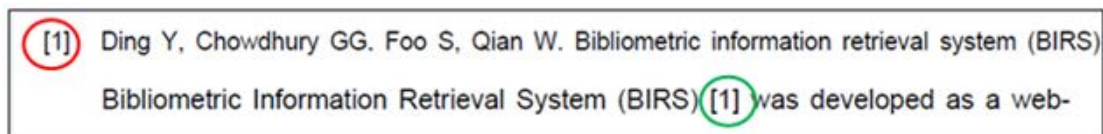


FIGURE 2.1: Example of reference string with citation-tag

Citation tag identification of cited papers in the citing document is an important issue [53]. The reason of wrong identification is the various formats of citation-tags and citation-anchors. The examples of diversified reference tags taken from different real papers are shown in Figure 2.2.

8) J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In ACM conference on Computer supported cooperative work, 2001.

[18] L. Ungar and D. Foster. Clustering methods for collaborative filtering. In *Workshop on Recommendation Systems*, 1998.

63. Ungar, L., Foster, D.P.: Clustering methods for collaborative filtering. In: *Proceedings of the Workshop on Recommendation Systems*. (1998)

[SW93] M.F. Schwartz and D.C.M. Wood. "Discovering Shared Interests Using Graph Analysis". *Communications of the ACM*, Vol. 36(8):pp. 78-89, August 1993.

[Sutton & Barto, 1998] Sutton, R. S., Barto, A. G., *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, Massachusetts, 1998.

[Resnick and Varian, 1997] Paul Resnick and Hal R. Varian. Recommender systems. *Communications of the ACM*, 40(3):175-188, 1997.

[Hofmann 2004] Hofmann, T.: "Latent Semantic Models for Collaborative Filtering"; *ACM Trans. on Information Systems*, 22, 1 (2004), 89-115.

[Hofmann, 2004] Hofmann, T.: *Latent Semantic Models For Collaborative Filtering*. *ACM Transactions on Information Systems (TOIS)* (2004) 22(1): p. 89-115

[RIS*94] RESNICK P., IACOVOU N., SUCHAK M., BERGSTROM P., RIEDL J.: GroupLens: An open architecture for collaborative filtering of netnews. *Proceedings of the ACM Conference on Recommendation Systems at the 15th National Conference on Artificial Intelligence*, 1994.

[Res*94] Resnick, P.; Iacovou, N.; Suchak, M.; Bergstrom, P.; Riedl, J.: GroupLens: an open architecture for collaborative filtering of netnews. *Proceedings of the ACM Conference on Recommendation Systems at the 15th National Conference on Artificial Intelligence*, 1994.

[UnFo98] Ungar, L.H.; Foster, D.P.: Clustering methods for collaborative filtering. *Workshop on Recommendation Systems at the 15th National Conference on Artificial Intelligence*, 1998.

[Herlocker et al., 1999] Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An Algorithmic Framework For Performing Collaborative Filtering. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 1999.

[Ungar et al. 98] Ungar, L.H. AND Foster, D.P.: "Clustering Methods for Collaborative Filtering"; *Proc. Workshop on Recommendation Systems*, AAAI Press, Menlo Park California, 1998.

[Sarwar et al., 2000a] Sarwar, B., Karypis, G., Konstan, J., Riedl, J. Application of dimensionality reduction in recommendation systems. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2000.

[Sarwar, B. M.; Karypis, G.; Konstan, J. A.; and Reidl, J. 2001. Item-based collaborative filtering recommendation systems. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2001.

FIGURE 2.2: Example of different formats of citation-tags in existing literature

The citation tags is the combination of either bracket ,i.e., [], parenthesis (), alphabets, numeric, dot, comma and some special symbols like *, +. Some of the citation tags contain last names of the first author and year informations "[Hoffman 2004]", "[Herlocker et al., 1999]". Some citation tags are prepared by combination of the first two characters of author names and the last two digit of the year ,i.e., "[UnFo98]". The last reference string in Figure 2.3 contains no citation tag at all.

In different domains computer science, medical etc the researchers are using different types of citation anchors that are given in Figure 2.3. The numerical citation anchors are like "[1]", "[1][2]", "[1, 2, 3]", "[12-15]", "[1]-[5]" and "[1-3, 8, 9]". Some researchers are using citation anchors as superscript like text¹ or text⁵⁻⁶. The alphanumerical citation anchors are like "author, (year)", "author [2002, 2003]", "author et al., 2003, author et al., 2003a, author & author, 2003, and author and

author, [2005]”. These different formats of citation anchors reduce the accuracy of in-text citation frequency calculation of cited papers as highlighted by [18].

from the early work on filtering netnews [15] and the movie recommender system for collaborative filtering by default focuses on rating-based user profiles [1, 9, 10, 16, 21]

Numerous studies on searching research documents have been proposed [1]-[5]. A Filtering (MFCF) is widely used by other researchers [6-8].

Sarwar, Karypis, and Konstan (2000, 2002) applied dimensionality reduction for the user based CF approach. He also used SVD for generating predictions. In contrast to our work, Sarwar et al. (2000, 2002) do not consider based algorithm have been suggested, e.g., (Breese, Heckerman, & Kadie, 1998; Herlocker, Konstan, Borchers, & Riedl, 1999).

Hofmann (2004) proposed a model-based algorithm which matrix factorization techniques (Canny, 2002; Hofmann, 2004; Sarwar et al., 2000; Srebro et al., 2005)

correlation [Resnick et al., 1994] The Pearson correlation coefficient is layers in the user competences profile [Brut et al., 2008d].

haviour is dominated by hedonic motivations [Oli03, Zil88b, Zil88a, ZB85]. According to this Movielens web site [RIS+94]. An evaluation was performed using two MovieLens datasets regression models [SKKR01] based on user and item features

[RIS+94]. Given the data privacy restrictions imposed by the multiple segments and describe the strength of each relationship.⁷

FIGURE 2.3: Various formats of citation-anchor in existing literature

The accurate identification of citation tags and matching of them with the various formats of citation anchors in text is difficult task. The contemporary systems have used diversified approaches such as string matching [53, 54] and set of heuristics [18] to achieve the accuracy of both types of citation ,i.e., citation-tag and citation-anchor.

Giles et al developed heuristic over 5093 documents consisted of 89,614 references. The documents of the corpus existed in Postscript format and identified by “.ps” or “.ps.Z” or “.ps.gZ” with web crawler. They extracted the set of references from the reference sections of the citing papers and then parsed each citation into metadata, such as citation tags, authors, title, and page number. The reference section

is identified by the keyword “REFERENCES” or “References”. They first identified the most regular features based on their position and composition in the reference string. The position means that the citation tags occur at the start of the citation, the author information precedes the title information. The composition means that the year of publication contains four digit beginning with the digits “19”. They also used the database of author and journal names to identify more subfields of citations. They used the citation tags to match with the citation anchor to extract the citation context. The text around the citation tag in the document is called context of citation. However, this method is unable to identify the citation tag in the reference string as given in Figure 2.4. Reference string without citation tag is another problem that affects the accuracy of in-text citation frequency calculation. Gipp et al [53] did not use the citation tag for finding in-text frequency of citation anchors. Furthermore, Giles et al have claimed an accuracy of 80% for the identification of metadata from the papers..

Claypool, M., A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin, 1999. Combining content-based and collaborative filters in an online newspaper. In *ACM SIGIR'99. Workshop on Recommender Systems: Algorithms and Evaluation*.

FIGURE 2.4: Example of reference string without citation-tag

The Bergmark [54] proposed four steps approach ,i.e., (1). The identification of citation anchors in-text, (2)The extraction of reference section (3) Parsing the Reference Strings, and (4) Matching reference anchors to the reference tags of reference strings. They converted the documents into XHTML format for the analysis. In the first step, they identified the anchors along with context informations in the body of each documents. Anchors are tags of cited paper that are used in the text of the citing paper. They identified the citation anchors in Figure 2.5 based on the occurrence of “(”, “[” and “{” for the papers published in D-Lib.

[1] or [1,3] or [8-10]	See Hakkala (1996)
[Bruce and Wayne]	Bruce and others (1997)
[Bruce et al.]	Bruce and Wayne (1998)
(Bruce & Wayne, 1998)	
(Bruce,1998)	
(Bruce, 1998, Wayne, 1999)	
(Bruce et al., 1998)	
(CNRI, 1997)	
{Digital Library Initiative}	
[Bruce, 1996; Wayne, 1999]	

FIGURE 2.5: Citation-anchors in citing documents

But the problem with bracket based search is that it will create the mathematical ambiguity like equation no “(1)”, interval “[-2, 2]”. They handled the numerical ranges by replacing [1-3] with [1][2][3]. They also broke the comma and semi-colon lists into individual citation anchor “[Bruce, 1996; Wayne, 1999]” into “[Bruce, 1996]” and “[Wayne, 1999]” and also highlighted the problem that some authors use the anchors as part of speech. The POS is usually used between authors and year of publication as shown in red circle in Figure 2.6.

Shardanand and Maes measured "reversals" – large errors between the predicted and actual rating [1995];

FIGURE 2.6: Citation-anchor with part-of-speech (POS)

In the second step of Bergmark approach [54] extracted the reference section based on keyword “References”, “Bibliography”, “Notes and References”, “Note and References”. The reference section identification approach will suffer by these problems: when there is no reference section, references are in a different file in the case of HTML documents, and when reference section loses its markup during the conversion of HTML document into XHTML document like JTidy tool remove the “<H3>” markup due to the syntax problems. In the third step, they extracted the citation tags along with the metadata, such as authors name, title, year, and page number. In the fourth step, they proposed exact and approximate matching algorithms for the matching of reference anchor and reference tag. usually the reference anchor and reference tag are different to each other, e.g., (reference

anchor [10], reference tag 10.) and (reference anchor [Borden and locks, 1998], reference tag “Bordon, Fred and Galdie locks”). They showed 86.7% reference or citation tag accuracy over 66 D-Lib papers. The reference tag accuracy for one reference string is the percentage of its elements that are correctly parsed. The elements consist of each author, title, year, contexts and URL if present. Bergmark [54] did not use the citation tags for the in-text citations frequency.

Nadirman et al [55] worked over 242 research papers to trace the reference strings from the reference section of research articles. They converted the 242 papers into text files. They extracted attributes title, author, and year and shown 91.54% accuracy of these attributes from the reference strings of 242 citing research papers. However, they did not identify the citation tags in their work.

In Tkaczyk et al [56] research study, different tools have been compared for the extraction of metadata from the reference strings in reference section of articles. The metadata consisted of author, title, journal, pages, volume, year etc. According to their evaluation, the best performing tools are CERMINE [57] and GROBID [58]. The authors of these tools were not highlighted the accuracy of the in-text citation frequency. The citation-anchors detection of these tools have been suffered by the different problems, such as string citation-anchor with bracket problem, citation with same author and year problem, multiple numeric citation-anchor with semicolon problem, and year inclusion problem.

Shahid et al's [18] evaluated the string comparison based methods to highlight the problems of identification of in-text citation from the corpus of research documents. They created the dataset that consisted of 1200 PDF files and 16,000 references. The proposed methodology gives 58% accuracy of in-text citations frequencies identification. The 42% error was due to the problems, such as mathematical ambiguity, wrong allotments, commonality in content, and string variations with citation tags. They categorized the citation tags into different groups, such as Numeric, Alphabetic, and Single character. The numeric citation tags are like “1. , [1], 1), (1)”. The example of alphabetic citation tags are such as “Srinivasan, Scherbakov 1995”, “[Davenport and Prusak, 1998]”, “[Staiger 1993]”, and “[Olson

et al. 2002]”, “[MPEG-7]”. The single character citation tags are “[N]”, “[P]”. The mathematical ambiguity occurs when a reference string has a numeric tag, such as “2.” In figure 2.7(a). The identification of this citation tag in the text of document will give some wrong citation anchors, such as the mathematical intervals like “[−2, 2]”, and equations (2) mentioned in the text of the paper and have been highlighted in Figure 2.7(b). They have shown the mathematical interval problem and the mathematical parenthesis problem in below Figure 2.7. They have shown the mathematical ambiguity problem with Figure 2.7(c) & 2.7(d). The citation tag “8.” can occur in various formats in text, such as [8], [1, 2, 8], [1][2][4], and [1-9].

The string variations problem occurs due to the inclusion of hyphen (-) in the reference anchor, such as “Law-verre and Schanuel 1997” that will not match with reference tag in the reference section. They highlighted the problem of same first author with different co-authors in the same year in different research papers, such as “Viroli and Omicini, 2001” and “Viroli et al., 2001”. According to Shahid et al’s [18], this problem could not be solved with first author and year information alone. They have further shown citation tags, such as “[P]”, “[A]” are very common citation tags that are matched mostly with the content of the paper.

Some of the problems do not detect with the exact matching of citation tag with citation-anchor. These problems are “multiple-anchor problem”, “range-anchor problem”, “compound-anchor problem”, “format problems”, “hyphen with carriage return and line feed problem”, “year related problem”, “citation-anchor with POS problem”, and “reference string with superscript citation-anchor”. These problems should be consider in the detection of in-text citation patterns and their frequencies in the full-text document.

2. M.V.A. Andrade, J.L.D. Comba and J. Stolfi *Affine Arithmetic*, Interval'94, St. Petersburg (Russia), March 5-10 - (1994).

(a)

2. Square (or even power) of affine form:
 Consider the square function x^2 with $x \in X = [-2, 2]$

in 1994 by Andrade, Comba and Stolfi [2] and was first applied in computer graphic problems [4] and surface intersections [6]. Recently, these methods were

It is not an in-text citation of ref no. 2

It is in-text citation of ref no.2

(b)

8. S.Rudeanu: *Lattice Functions and Equations*. Springer-Verlag, London 2001.

(c)

In the sequel we assume that the function f is given in Table 1, while g, h_1 and h_2 are of the form (8), (9) and (10), respectively.

It is reference to an equation no. 8 instead in-text citation of reference no. 8

(d)

FIGURE 2.7: Mathematical ambiguity issues a) Reference string snapshot from paper b) Mathematical interval problem c) Reference string snapshot from paper and d) Mathematical parenthesis problem

2.3 Research Paper Recommendation Systems and Approaches

In the first subsection of this part, we shall describe two state-of-the-art research paper recommender systems, namely Google Scholar [59] and CiteSeerX [60]. These systems are openly available for researchers who want to search multidisciplinary literature. In the subsections, we shall highlight various approaches for the research paper recommendation that are proposed in the literature. On the basis

of analysis of existing techniques, the approaches have been categorized into collaborative filtering based approaches, citations context based approaches, citations based approaches, meta-data based approaches and hybrid approaches.

2.3.1 Research Paper Recommender Systems

In this part, two state-of-the-art research paper recommender systems will be discussed, namely Google Scholar and CiteSeerX. These system are being widely used for literature selection by researchers from different domains.

Google scholar [59] is the internet-based search system that is freely available to find scholarly documents like academic papers from conferences and journals, books, abstracts, technical reports and other academic literature from various fields of research. It can also help researchers find different metadata that are freely available in full text research documents. Google scholar offers a variety of options, such as creating a link between cited documents and citing document, and also allow users to maintain a customized library of research documents. Google scholar exploits the keyword searching to return most relevant results. This search tool provides the results in ranked format. The exact algorithm behind Google scholar for searching of relevant documents is unknown [61].

CiteSeerX [53, 60] is the openly available digital library and search engine which consists of academic literature in PDF and Postscript format. This electronic library has focus on the publications in computer science domain. This tool is used to provide the most recent relevant research documents based on cited by and co-citation datasets. CiteSeerX also has capabilities to provide the relevant result based on keyword searching, citation and citation context from the huge amount of academic documents. This tool can easily index the full-text documents.

2.3.2 Collaborative Filtering based Approaches

Collaborative Filtering (CF) remained an important approach in the literature to build recommender systems. It uses the known preferences of a group of users to make recommendations or predictions of the unknown preferences for other users. The fundamental assumption of collaborative filtering is that if users X and Y rate n items similarly, or have similar behaviors, such as buying, watching, listening and therefore they will rate or act on other items similarly. The Collaborative Filtering has been applied in the past in diversified domains, such as mineral exploration, environmental sensing, financial data, electronic commerce, and web applications data.

Goldberg et al [62] initially used collaborative filtering. Collaborative filtering approaches have been used for various purposes in various domains, such as USENET articles [63], jokes [64], college courses [65], and commerce site including Amazon.com, Ebay.

Zhang et al [66] designed and implemented a paper recommender system based on semantic concept similarity. It is computed from collaborative tags. Semantic concepts are used to represent user profiles and item profiles. Collaborative tagging describes the process by which users add metadata in the form of keywords to content. The neighbor users are selected using collaborative filtering and content-based filtering approach is utilized to generate a recommendation list from the papers, tagged by their neighbors. They evaluated their approach on a large dataset comprising of 220,723 papers from CiteULike. In the dataset, there were 6800 users and 70,796 tags. The semantic concept similarity algorithm was trained on 90% dataset and approach was evaluated on 10%. This approach does not work when the numbers of neighbors are small. They observed during evaluation that if the size of neighbor users set increases, the hit percentage also increases. They identified that user groups were not accurate therefore, it was concluded as a future work that clustering of users may improve the quality of neighbor users.

The traditional collaborative filtering approaches such as user based and item based CF get the user's preferences at a low-level (item level). The systems use the co-rated items of users to find the user's similarity. In reality, the user's may like to gather similar items into categories for corresponding user groups. There are scenarios in which users x and y rated the five different items in the same group respectively. The main challenges of Collaborative Filtering are data sparsity, scalability, synonymy, gray sheep, shilling attacks, privacy protection [10].

2.3.3 Metadata based Approaches

Another important approach to recommend relevant papers is by exploiting the metadata of research papers like title, author, and keywords. A recent study of metadata based recommendation system has been performed by [14]. They designed a novel approach to identify the relevant papers of a user interest based on given keywords. The proposed technique consists of three steps (1) fuzzy clustering of papers to get the group of related papers based on topic similarity, (2) selection of a summary paper among a group of same papers and (3) finally performed ranking on summary papers to get good quality papers on the top of the list to complete the user needs. The summary paper allows us to summarize the set of papers into a single representative one. It also simplifies users interaction with huge number of papers from literature. They constructed a corpus from Web of Science, DBLP, CiteSeerX and local database sources. The dataset consisted of common attributes from papers, such as title, authors, published date, journal, and citation or reference list. The title and abstract features used to find all those papers which have similar topics and interest based on partial keyword matching. In this research paper, they have used the co-citation criteria to identify the group of papers which share common interests. They used two measures recall and precision for the evaluation process.

Chen et al [12] proposed methodology based on citation network which is called Citation Authority Diffusion (CAD). The approach was developed to retrieve and

recognize the important papers from survey documents collection. The SIM (Survey Importance Measurement) System has been developed based on CAD approach. It is available as online web service. The metadata such as title, abstract, keywords, and bibliography of a target research are used as input for SIM system. The proposed methodology is composed of three modules: 1) Information Collection; 2) Information Organization; and 3) Information Presentation. The first module is designed to extract the concepts from the target research and then these concepts are used to retrieve the survey papers collection. The second module is constructed to discover the potential paper and relationships among survey documents. Whereas the relationship between the generated surveys model and the target research is presented by Information presentation module. Thus, this module computes a survey novelty score to the target research which helps the people (users) in understanding what have to do or what they have done so far? For evaluation, they selected a corpus of papers that were published before 2008 in CiteSeerX. The dataset was constructed by 456,787 unique papers. They prepared 1,612 papers set with quality references for testing purpose. The dataset was limited to specific domain such as computer science in CiteSeerX that is not enough to check the accuracy of proposed system. Hence, the system can be evaluated against different datasets. They further planned to extract more concepts from the target research to retrieve more relevant papers from survey documents.

Livne [13] have explored the future citation counts of papers based on given information's that are available at the time of publication. They prepared dataset from Microsoft Academic Search consisted of 38 millions papers; 19 millions authors belong to over 15 academic domains. The metadata or features such as author, venue, references, and citations were extracted automatically. It was a huge size of dataset for experiment, Hence they selected the papers set that were published from 2000 to 2005 across seven domains, such as Biology, Chemistry, Medicine, Computer science, Mathematics, Engineering, and Physics. The proposed model predicted citation counts well in some domains, e.g., 39% in Medicine, 35% in Biology, 33% in Chemistry, and 30% in Computer Science based on all given features. It means that more work may be expected in these domains in future. The

proposed technique can be extended across sub-domains as well as to predict the impact of high level entities, e.g., researchers and universities.

Hong [67] proposed IARS (Interesting Area Recognition System) to find a user interest in research field, and then employed it to create user profiles. At the user end, the recommender system also filters and suggests the research papers to users based on user's given implicit feedback. IARS uses the Category, Journal information, Scope and paper information, such as title, author, and year of publication, keyword, and abstract to recognize user parts of interests. The category, journal information and paper information are acquired by a crawlers and extractors from Google and Google Scholar respectively. These metadata are stored in an information database by a database manager. It also provides a list of recommended papers based on metadata for a users. In the implicit feedback, the users are not aware of the fact that they are providing feedback or their behavior is being used by a recommender system. The feedback or user information, number of clicks, stay time and the records of purchase is observed by the recommender system. The clicked information is filtered by Feedback filter module to find the user interest and then it is utilized by Profile Manager to create the user profiles. The user profiles consist of user preferences that express the interest of user research field. User profile renewal is performed whenever a user clicks research papers. The proposed approach was evaluated only on journal papers in the field of computer science. Their system provided over 88% average precision.

Hoxa et al [68] proposed a paper recommender system based on the literature that generated by the Albanian researcher in their country or across its neighboring countries. The scientific documents were written in Albanian language and there was no such system to find a relevant paper in such articles. The dataset was very small consisted of 226 articles for experiment. They designed a modular system architecture consists of few modules, Articles database, Database Populator, Metadata Extractor, Articles Searcher and Articles Recommender. They extracted metadata such as title, authors, abstract, keywords, body and the articles parts by metadata extractor. The proposed system also extracted the terms frequency

across the body of articles, title, and abstract as well as across the different sections, such as introduction, related work. Database populator is used to store all these metadata in articles' database. The module Articles Searcher is used for keyword based queries and also indexes the metadata in article database and returns search results based on the presence of term in the document. Articles recommender recommends similar articles to the one that the user is currently viewing. The results are ranked by the frequency of searched term in the documents. They proved that the top results contained the relevant items.

2.3.4 Citation Context based Approaches

The citation context has also been used to recommend most relevant research papers, for example, Kaplan introduced a new method based on co-reference chain for extracting citation context from research papers [15]. Co-reference occurs when two or more expressions or sentences in a text refer to the same person or thing; they have the same referent, e.g., Bill said he would come; the proper noun "Bill" and the pronoun "he" refers to the same person, namely "Bill". The co-reference chains match noun phrases that appear with other noun phrases to which they refer. The citing paper contains citations that are represented by citation markers. Citation marker, like "[1], [abc et al]" are called the citation-anchors. The text around the citation-anchor is called citation site (c-site) for short or citation context. Each sentence in citation site is known as c-site sentence to represent the block of text that refers to the cited work. The approach worked on the identification of citation contexts with background information from research papers. The term background information is to refer to any running text that elaborates the c-site but strictly it is not a part of c-site. Background information may need to be included for the citation to be comprehensible. This information is important to understanding the c-site sentences. Background information is a form of meta-information about the c-site. The proposed architecture contains two major modules (1) corpus construction and analysis (2) creation and evaluation of the conference resolver. The corpus was created which were consisting of 38 papers, 50

citation contexts and 90 citation context sentences. The algorithm behind the co-reference resolver was working in the following manner. The algorithm first finds the anchor sentence. Then it tries to search noun phrase in the anchor sentence. The algorithm begins sentence by sentence search from right to left. If a noun phrase occurs in the sentence, then the searched sentence will be concatenated with an anchor sentence. The same process will be repeated up to a specified distance threshold or until a noun phrase sentence occurs. The same process will be iterated for the new noun phrases in anchor sentences. They have evaluated their technique with cue-phrases technique and concluded that co-reference chain outperforms cue-phrases, i.e., the previous technique have identified 64.9% correct sentences out of 94 sentences while the co-reference chain technique have obtained 74.4% correct sentences . However, the proposed method has some limitations that it was not tested over a large dataset of citation context and the noun-phrase feature was not enough for the improvement of their co-reference chain method.

He et al [69] have developed a context aware citation recommender system that can recommend a highly quality set of citations for a paper. They have implemented a prototype system in CiteSeerX to recommend bibliography to a document and providing the ranked set of citations to a specific citation placeholder in a query paper. Citation placeholder is the location to cite a particular reference or citation marker [15] in the text of the paper. The steps of the developed system are: (1) query document preprocessing, and (2) selection and ranking of recommended citations. In the first step, they extracted the global context and local context from a query document. In the second step, they associated the local context with each placeholder in query document and then generated the bibliography list for the query document by the selection and ranking of citations. Title and abstract of the paper is global context. The local context is the text surrounding a citation or placeholder. The different sizes of local contexts impact the information retrieval performance. Therefore, they have selected the fixed window contexts, i.e., size of 100 words) for their experiments. After removing all stop words, they have selected 50 words before and 50 words after the citation anchors. They have prepared the dataset consisting of titles, abstracts and 1,810,917 local citation contexts from

456,787 unique documents in the corpus. They have used 1,612 papers as a testing data set. They evaluated the proposed approach against many baselines in the CiteSeerX digital Library. The system performance was also evaluated by user studies and click through monitoring. Their technique was based on a partial list of citations. The system might not work well when unknown terms or features are scanned from a documents. This might be overcome using autonomous learning of new key-terms or features from the dataset.

Tuarob et al [16] presented an initial effort in understanding the descriptions of algorithms from the content of the research documents. Specifically, they identified how an existing algorithm can be used in scholarly works and proposed a classification scheme for algorithm function. The scheme consisted of 9 classes of algorithm citation functions. They divided these classes into three categories such as favorable, neutral and critical based on the Authors' attitudes. They used the dataset of 2000 papers from CiteSeerX along with 300 algorithm citation contexts. Algorithm citation contexts consisted of algorithm citation sentence, i.e., a sentence in which one or more algorithms are cited, and sentences that immediately precede and follow it. They find that authors are mostly 60.99% of the time neutral, 28.34% critical and 10.67% favorable towards other algorithms.

The hypothesis of Sugiyama and Kan [6] was that the author published work shows the interest of a researcher. They designed approach that was capable to increase the author's profile based on references lists in their publication history and citing papers of each profile paper. PageRank is the general ranking scheme and it does not consider the user interest in ranking. Previous recommender systems considered the user interest in limited sense by using metadata or collaborating filtering. The technique used the contextual information from neighbors, i.e., citing and referenced papers by of the target paper. It is domain independent. The proposed method consisted of four steps: (1) user profile construction and its conversion into feature vector (2) feature vectors construction for candidate papers (3) similarity identification between feature vector of user profile and candidate papers and (4) finally recommend papers with high similarity. For the experiment, they selected publication lists of those researchers who have publications' in DBLP

source. The corpus of candidate papers consisted of 597 full text papers. The dataset also contained information about the citation and reference papers for each author's profile paper. They evaluated recommendation accuracy of their approach using NDCG and MRR, and achieved better results than Pagerank as baseline. The efficiency of this approach relied on the complete user profile. There is a need of other approaches for user profile construction. In addition, there is a need to develop methods for recommending papers that are easier to understand to quickly gain knowledge about their intended research.

2.3.5 Citation based Approaches

Citation or direct citation is one of the popular measures to find the relationship among documents. If a citing paper refer to the published or unpublished work in the reference section by including some citation tag is called direction citation or citation. Citation tag can exist in different formats such as “[1], [xyz et al., 2009], 1), [HKKR002]”. The researchers believe that most of the references in bibliography are very important to describe the idea in the citing document [70]. There are many approaches to recommend relevant scientific literature proposed in the literature using the citations of research documents, such as Bibliographic Analysis [35], co-citations analysis [32], Citation Proximity Analysis [22], and Citation Order Analysis [37].

One of the famous citation based approach is known as bibliographic coupling [35]. In bibliographic coupling, two papers P1 and P2 are considered similar, if they share some common references in their bibliographic sections. These common references define the bibliographic strength between two or more research documents. In other words, if two documents share a large number of common references in their bibliographic sections then it means that the bibliographic strength between these two documents is greater and hence they are highly relevant to each other. For experiment they used a dataset consisted of 8521 articles which generated 137,000 references. Experimental results proved that bibliographic coupling performed well in recommending relevant articles. However bibliographic coupling

depends on the references contained in the coupled documents. Therefore it is fixed and can only identify permanent relationship between research articles. Similarly this approach may fail to provide all relevant documents if all the research papers are not listed in the citation.

Small H [32] proposed a new measure called co-citation. He used it to find the document relationship. It is the frequency of two documents cited together in other papers. The co-citation frequency of two cited documents can be determined by comparing the lists of citing documents and counting identical entries. The bibliographic coupling and direct citation were two measures which were used to find documents relationship before co-citation. The co-citation links cited documents while the bibliographic coupling links the source documents. Strong co-citation links represent the subject similarities and association or co-occurrence of ideas. The proposed technique not performs well against the dataset which have no citations in their papers. However, the frequency of citations and citations in different logical sections are not used to identify relationship strength between co-cited papers.

Gipp and Beel have proposed new approach called Citation Proximity Analysis [22] developed based on existing co-citation technique [32]. They have checked the proximity or position of co-citations to each other within full text of a paper. According to authors' analysis, if co-citations occurred closely to each other, the papers are more related. They denoted the proximity of co-citations in different parts of document by different CPI (Citation Proximity Index) values or weights. The CPI values were 1, 1/2, and 1/4 etc for the co-citations in sentence, paragraph and chapter respectively. They selected CPI based on occurrences of co-citation. They used three steps to calculate the CPI values. In the first step, the document is parsed and a series of heuristics are used to process the citations including their position within the document. In the second step, the citations are assigned to corresponding items in bibliography. In the third step the proximity among co-citation is examined. The dataset was prepared from research paper recommender system called Scienstein.org. It contained 1.2 million papers. The technique was

used to analyze the similarity and classification of selected corpus. The evaluation was conducted with the existing techniques, such as Bibliographic coupling, Co-citation analysis, and Keyword based approaches. The CPA produced better precision over these techniques. They do not consider the citation context. They also give same weight to the co-cited papers if they are co-cited in results sections and in related work sections.

Gipp and Beel introduced another approach COA (Citation Order analysis) [37] which is a variant of co-citation analysis. In COA, the order of citations are considered that is used for the identification of a text similar to one that has been translated from language A to language B, as the citations would still occur in the same order. CPA and COA do not replace the text analysis and existing citation analysis techniques. The CPA and COA offer substantial advantages in identifying related documents in comparison to existing approaches. CPA assigned different weights to article, paragraph and sentence. The weights are used to represent the importance of the different parts. These technologies can be used with collaborating filtering to identifying more relevant documents for new researchers.

In Boyack et al approach [21], the whole research paper document is considered as a set of bytes. To find relevancy between two co-cited papers, the byte offset between the citation-anchors of the two papers is calculated and a weight is assigned accordingly. If the byte offset between the citation-anchor positions of two co-cited papers A and B is 375, 1500, 6000 and over 6000, then the weights assigned will be 3, 2, 1 and 0 respectively. The byte offsets such as 375, 1500, and 6000 are ways to approximate the lengths of sentences, paragraphs, and sections without using the actual sentence structure, such as used in CPA [22]. They considered the average sentence length as 375 bytes and so the byte offsets 1500 and 6000 were considered equal to 4-16 sentences respectively.

In the recent work of Colavizza et al [71], the similarity of research paper pairs at different levels of co-citation such as journal, article, section, paragraph, sentence, and bracket are analyzed in fulltext documents. They consider section as anything which has a heading. However, generally more than one headings belong

to one logical section ,i.e., introduction, related work, methodology, result, and discussion. They do not consider the structure like IMRaD [38] for the co-citation analysis to find the papers similarity.

Hou et al [17] and Shahid et al's [34] proposed a new measure called citation frequency or citation counts within the text of citing paper. It can be used to improve the accuracy of citations. The hypothesis of Hou et al was that most frequently occurred citations in text are considered most important references for citing article. They have used strategy to classify the closely related references (CRR) and less related references (LRR) in the reference list of citing document based on common references between cited documents and citing document. They analyzed 651 papers published in 2008 and after experiments, averagely they found that each CRR appeared 3.35 times and each LRR appeared 1.88 times in corpus. It was concluded that the CRR occurs frequently in the text of citing paper.

Whereas Shahid et al's [34] proposed the idea to retrieve most relevant citing papers of the cited document. They introduced a new measure known as in-text citation frequency to find the relationship strength between documents. The number of times a particular citation occurred in the text of citing paper is called in-text citation frequency. The existing approaches such as Text based similarity approaches, Context based approaches, and co-citation analyses do not consider such semantic information. The proposed technique contains different modules like 1) Document Fetcher and 2) document Parser modules. The first module is used to get the document from dataset and convert into xml format. The converted file is used as input for Document parser module. It has been further divided into sub-modules such as: (i) Citation Tag Identifier (ii) Section Identifier, and (iii) Citation Tag Frequency Calculator. Citation Tag Identifier is used to identify the citation tag in a text. Section Identifier exploits the layout information of research papers and a domain specific dictionary to identify sections in the document. The citation tag frequency calculator is utilized to find the frequency of particular citation in a text. They have used the dataset from the J.UCS containing 1460 documents. It was found that if a citing paper cites a cited paper in the full text more than five times, then there exists a strong relationship among documents.

However, the approach was evaluated on a one typical journal; there is a need to evaluate the technique for more venues.

Hence, in [18] they have proposed technique that is used to find the accurate patterns of citation tag in text of the citing document. The whole approach consisted of three steps: (1) PDF to XML conversion (2) Calculating the citation frequencies (3) Clustering of citations based on frequencies. The dataset contained of over 1200 papers and 16000 citations. They extracted more than 3000 accurate citations. Most of the citations missed due the concern problems such as wrong allotment, mathematical ambiguities, commonality in content, String variations exist in scientific document. The accuracy of automatically identifying in-text citations remains 58%. To prove their concept that more in-text citations would denote strong relation, they have manually corrected all wrongly identified in-text citations. In the citation based approaches, an interesting observation of Hou et al was that some of the references are used for only background purpose or incidentally mentioned. Therefore, such observation of researchers creates doubts about the citations performance [17].

2.3.6 Hybrid Approaches

To recommend most relevant research papers, different researchers used hybrid approaches by combining different approaches. Strohman et al proposed a hybrid approach to get the related work for unpublished document based on text and citations graph of previous work [72]. The unpublished document was used as query in the system. Most current literature search systems focus on short query while their system is based on large query. This query may comprise one or more than one pages. Authors have argued that the text similarity computation is not enough to find the relevant document. The authors have exploited graph-based features as well in the retrieval process to achieve high quality retrieval results. The retrieving of relevant documents is particularly a challenging task because the concept of relevance is much severe. Most papers could cite hundreds of topically similar papers, but contain just a few highly relevant citations. The proposed

method was consisted of three steps. In the first step, they search a collection of over a million papers and returned the top 100 most similar papers to the query document as the set R . In the second step, they increased the set R of query documents from 1000 to 3000 papers with all the citing documents based on existing documents in set R . Text-based features are good for finding some similar related work. Citation features are useful to identify the conceptually related work than text features but may do a poor job at coverage (since recent documents may have no citations). Both features are not working well in isolation. Hence, in the third step, they utilized both types of features, such as publication year, text similarity, co-citation coupling, katz measure, same author (papers written by the same team of authors), and citation count to rank the documents in set R . They created dataset from Rexa collection that comprising of 964,977 papers, 105,601 full text papers, 1.46 million citations and 672,372 cited papers. One thousand papers were selected as sample queries. The evaluation study was conducted with text similarity technique as baseline. Experimental results show the effectiveness of their system in mean average precision. The large query size decreased the performance of the proposed system during matching process with the huge dataset and therefore, it was concluded that query size can be reduced to increase the performance.

Liang et al modified the links of citation network of scientific documents with citation relations represented by some weights [19]. They classified the citation relations into three categories such as (1) based-on, (2) comparable and (3) general. Based on relation is a relation when a citing paper is based on a cited paper to some extent, e.g., technique based relation. In comparable relation, the cited and citing paper is compared in terms of differences or resemblances, e.g., solve same problem with different methods. In last type of citation relation, they checked the background information similarity of citing paper with cited paper. The dataset prepared from ACL Anthology network consisted of 12409 papers and 61527 citation links. They conducted both offline evaluation and expert evaluation with five baselines techniques, such as Co-citation, Co-coupling, CCIDF, HITS Vector-based, and Katz graph distance. Experimental results show that their proposed

approach is more effective than the state-of-the-art methods for finding relevant papers. They plan to integrate their model with topic analysis to find more relevant papers.

Afzal et al proposed rule based Autonomous Citation Mining technique called Template based Information Extraction using Rule based Learning (TIERL) to improve the state of the art in Autonomous Citation mining based on some common heuristics [73]. It was used to overcome the limitations of existing current leading citation indexes, such as ISI Web of knowledge, CiteSeerX, and Google Scholar. These limitations are style of citation, spelling errors, improper citation linking and its extraction from PDF document. The approach consisted of two steps, 1) Template based Information Extraction (TIE); 2) Rule based Learning. In the first step, they extracted the reference entries based on the defined template. In the second step, they used heuristic rules to extract the data, such as authors, title, venue, and also control the uncertainty and approximate matching of citations. They considered more than 1200 papers in J.UCS journal for experiments. For evaluation, they selected the ISI, Google Scholar and CiteSeerX as baseline for the proposed approach. The overall accuracy of the system was 99.23% that shows better performance than the existing approaches to identify citations and these citations were then used to recommend relevant papers.

Author Co-citation Analysis is effective method based on co-citation counting. It was used to identify, trace, and visualize the intellectual structure of an academic discipline by counting the frequency with which any work of an author is co-cited with another author in the references of citing documents. The traditional approach assigned equally weight to all co-citation pairs without considering the variation of citing content. In paper [74], they further extend the current author co-citation analysis method by incorporating citing sentence similarity into citation counts. Citing sentences are used to obtain the topical relatedness between the cited authors. This similarity is measured by topical relatedness between two citing sentences. In the traditional approach of calculating the co-citation similarity, any author pair is counted as 1. But in the proposed approach, the author pair was

weighted by the similarity of sentences that these two authors were cited in the full-text article. They selected the dataset consisting of 1420 full text articles having 600,68 references. The dataset consisted of authors, titles of the cited documents and citing references. The results show the content-based ACA method reveal more specific subject fields than the traditional ACA.

Digital Libraries is very important tool for searching the scientific literature. Ranking algorithms are used to rank the search content of digital library. It depends upon many factors like citations to paper, content, authors and publications of the paper etc. Singla et al [75] have developed C3 ranking algorithm based on two parameters i.e. citation to paper and relevancy of content with the query. The proposed approach comprising of five steps: (1) Extraction of Keywords from given paper, (2) Extraction of summaries from query paper based on top ranked keywords (3) Retrieval of citing papers of given paper (4) Finding the similarity score between the summaries of query paper and each citing papers. Total similarity score can be obtained by adding the individual similarity score of each citing papers with query paper. C3 rank is the mean of total similarity score that can be obtained by total similarity score divided by total number of cited paper. They used only ten papers for experiment. The results of C3 ranking algorithm are compared with Citation count ranking algorithm and Content based ranking algorithm and it was concluded that C3 ranking algorithm performed better than existing approaches.

2.4 Summary

In this chapter, we have critically reviewed various existing research paper recommender systems that have been proposed in the literature. On the basis of existing techniques, the reviewed literature has been categorized into different categories such as Collaborative filtering based approaches, Citations context based approaches, Citations based approaches, Metadata based approaches and Hybrid approaches.

According to Beel et al [9] study, 55% approaches for research paper recommender systems have been developed based on content filtering and only 10% of research-paper recommender systems use a co-citation method. We have studied the literature about the existing research of co-citation method. According to our study as shown in Table 2.1, the co-citation approach has recently exploited to check the relatedness of research papers in-text of citing document based on proximity measure [22] and character offset [21]. This study shown that no one has analyzed the co-citation method across the logical sections of research papers such as introduction, methodology, result, and discussion. This structure exists for many years and is known as IMRaD (Introduction, Methodology, Result, and Discussion) [23, 24]. In this chapter, the literature study of research paper recommender systems, IMRaD structure, and citation-based approaches have been highlighted. After this study, we found the research gap that the section wise co-citation analysis should be analyzed and may recommend the relevant research paper.

TABLE 2.1: Summary of reviewed literature

S.No	Cited documents	Data source	Methodology	Strength	Limitations
GENERIC SECTIONS EXPLOITATION IN RESEARCH					
1	[26]	Research papers, citations, sections, sentences	Citation and sentiment based analysis across rhetorical sections	Retrieval of relevant documents	Need the proper citations and sections of a research papers
2	[6]	Research paper Citations, Citation-context	Citation Analysis, Citation context analysis, Section analysis	To enhance the author profile based on citation list, the author profile shows the user's interest clearly	Paper recommendation not possible without citations of a paper
3	[27]	Research papers, Citations, logical sections	Citation analysis, Section identification analysis	For Search, Navigation, Summarization	Need proper tool to convert the PDF document, Without sections and citations, the documents can not be possible to processed
4	[47]	Research papers, Citations, Citation contexts, IMRaD structure	Citation analysis, Section identification analysis, verb or lexical analysis		required proper citations in text of citing document, proper tool required for citation anchor detection
5	[38]	Research papers, Citations, Sections	Citation analysis, Section identification analysis	Research paper recommendation based on repetitive occurrence of citations in sections of papers	Need proper tool to convert the PDF document, Without sections and citations, the documents can not be possible to processed
6	[25]	Research paper, Citation, Citation context, Semantic data, IMRaD structure data	Citation, and Citation context analysis, In-text reference analysis, Semantic annotation, IMRaD structure analysis	Find the negational citations to improve the information retrieval of scientific papers	This analysis not possible for the papers which have no citations

Table 2.1 Continued...

S.No	Cited documents	Data source	Methodology	Strength	Limitations
7	[28]	Research paper, IMRaD structure data	IMRaD structure analysis	The technique depends on heading keyterm	The section identification not possible without using proper keyterm in section heading labels
IN-TEXT CITATION ANALYSIS IN EXISTING LITERATURE					
8	[22]	Citing documents, References, citation-anchor	Citation analysis, Co-citation analysis	Recommend relevant papers to authors	The paper without references is not processed
9	[37]	Citing documents, References, citation-anchor	Citation analysis, Order based Co-citation analysis	Recommend relevant papers to authors	The paper without references is not processed
10	[52]	Citing documents, References, citation-anchor	Co-citation analysis, citation context analysis	Recommend the most relevant sections in the documents based citation distribution	Required the research paper with full-text, PDF to XML or plain-text is also required
11	[18]	Citing documents, References, citation-tag, citation-anchor	Citation analysis, citation context analysis	Recommend the most relevant documents based on in-text citation frequencies	Required the research paper with full-text, Required proper tool for PDF to XML or plain-text
12	[21]	Citing documents, References, citation-tag, citation-anchor	Co-citation analysis, citation context analysis	Recommend the most relevant documents based on the co-cited frequencies	Required the research paper with full-text, Required proper tool for PDF to XML or plain-text
13	[76]	Citing documents, References, citation-tag, citation-anchor	In-text Citation anchor analysis	Improved the in-text citation frequency of citation	Required the research paper with full-text, Required proper tool for PDF to XML or plain-text
RESEARCH PAPER RECOMMENDER SYSTEMS					

Table 2.1 Continued...

S.No	Cited documents	Data source	Methodology	Strength	Limitations
14	[66]	user profiles, tags	Collaborative filtering, Content based filtering		This approach does not work when number of users are small
15	[10]	user profiles, items rating	Collaborative filtering		
16	[14]	Metadata (title, author, year etc)	Citation analysis, Metadata analysis		The metadata extraction does not possible without reference section in the papers
17	[13]	Metadata, Citations	Citation count analysis		
18	[6]	User profile, reference list, citations, citation context, citing documents	Citation context analysis		
19	[16]	Citation context of cited document and citing document	Citation context analysis for algorithm based relevancy		

Chapter 3

Proposed Approach Architecture

This chapter is dedicated to explain the architecture of the proposed approach. The parts for the proposed approach have been shown in the form of block diagram in Figure 3.1. The architecture has been divided into three phases (1) data preparation phase (2) section wise co-citation analysis phase and (3) document ranking and result evaluation phase. In data preparation phase, comprehensive dataset has been created for three main tasks of this thesis shown in phase 2. The main steps in second phase are (a) generic sections/ILMRaD structure identification (b) in-text co-citation patterns and frequencies identification and (c) section wise co-citation analysis SWCA. Third phase of the methodology ranks documents based on the proposed approach SWCA followed by the evaluation of proposed approach.

The architecture of the proposed approach has been constructed by automated, semi-automated, and manual components. The automated components are represented by dotted circle while the semi-automated component is represented by solid circle. The keyterms, section weights, and result evaluation parts are manually operated.

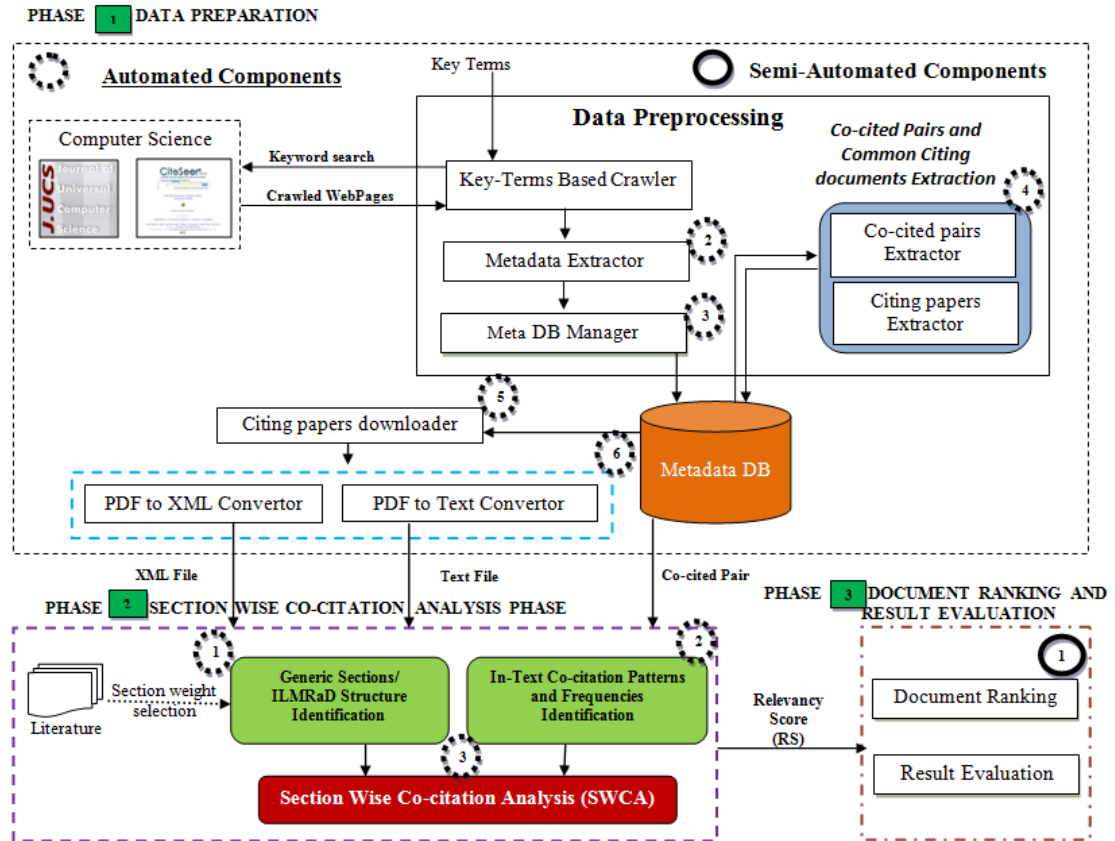


FIGURE 3.1: Proposed architecture for section wise co-citation analysis

3.1 Data Preparation Phase

For evaluation of the proposed approach SWCA, two comprehensive datasets have been prepared from two scientific digital libraries that include J.UCS¹ and CiteSeerX². The dataset of Journal of Universal Computer Science (J.UCS) has been selected from Afzal et al [77] research work due to the fact that it contains comprehensive selection of papers from all topics in Computer Science. The dataset of CiteSeerX has selected from CiteSeerX open digital library which consists of metadata about query papers, co-cited papers, and citing papers. The CiteSeerX dataset has been prepared by the combination of different components: (1) Key-Terms based crawler (2) Metadata Extractor (3) Co-cited pair Extractor (4) Citing papers downloader, and (5) PDF to xml or PDF to plain-text convertors. These components are briefly described below.

¹www.jucs.org/

²citeseerx.ist.psu.edu/

3.1.1 Key-Term based Crawler

Initially, some key-terms are selected from computer science domain as shown in Table 3.1.

TABLE 3.1: Key-Terms for query papers searching

Key-Terms
Collaborative Filtering
Information Visualization
Data Mining
Information Retrieval
Web Mining

These key-terms are exploited by the key-term based crawler to search the relevant webpages on CiteSeer site. For example, the key-term “Collaborative Filtering” is first split into two keywords “Collaborative” and “Filtering”, then the crawler uses these keywords in the link as given in Figure 3.2. Finally, the key-term based crawler returns the webpage which may contain of the links of query papers, co-cited papers, and citing papers.

["http://citeseerx.ist.psu.edu/search?q=collaborative+filtering&submit.x=0&submit.y=0&sort=rlv&t=doc"](http://citeseerx.ist.psu.edu/search?q=collaborative+filtering&submit.x=0&submit.y=0&sort=rlv&t=doc)

FIGURE 3.2: query paper link on CiteSeer site

3.1.2 Metadata Extractor

The webpage returned by the crawler of earlier step contains the links of queried papers (cited papers). The query paper link consists of metadata, such as ‘paper title’, ‘Authors list’, ‘year’, ‘citationid’, ‘number of cited by or citing documents’, and ‘doi’ as shown in Figure 3.3.

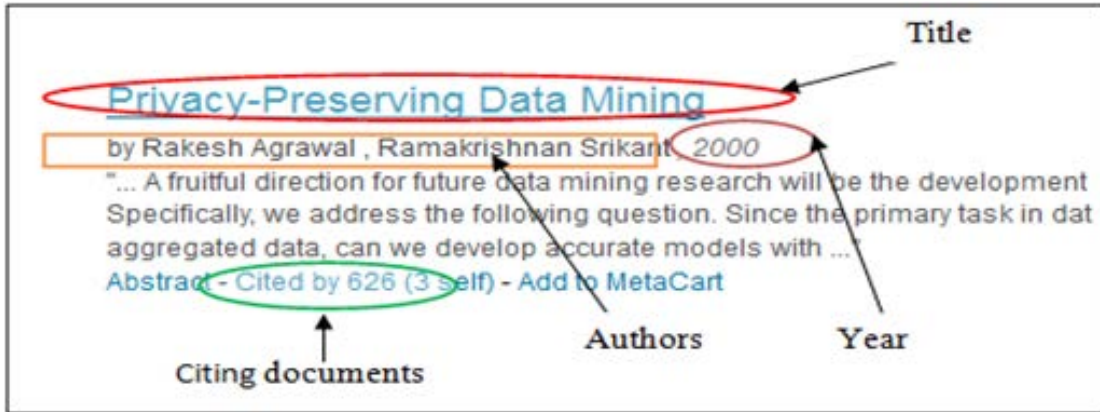


FIGURE 3.3: CiteSeer link pattern with metadata information

Furthermore, we extracted two more pieces of information from the ‘Author list’ metadata such as ‘First Author name’, and ‘the number of authors’. All these metadata are used in the preparation of the final dataset. The ‘paper title’, ‘First author name’ and ‘year’ are used to detect the occurrence of cited document in the reference section in the citing document as shown in Figure 3.4.

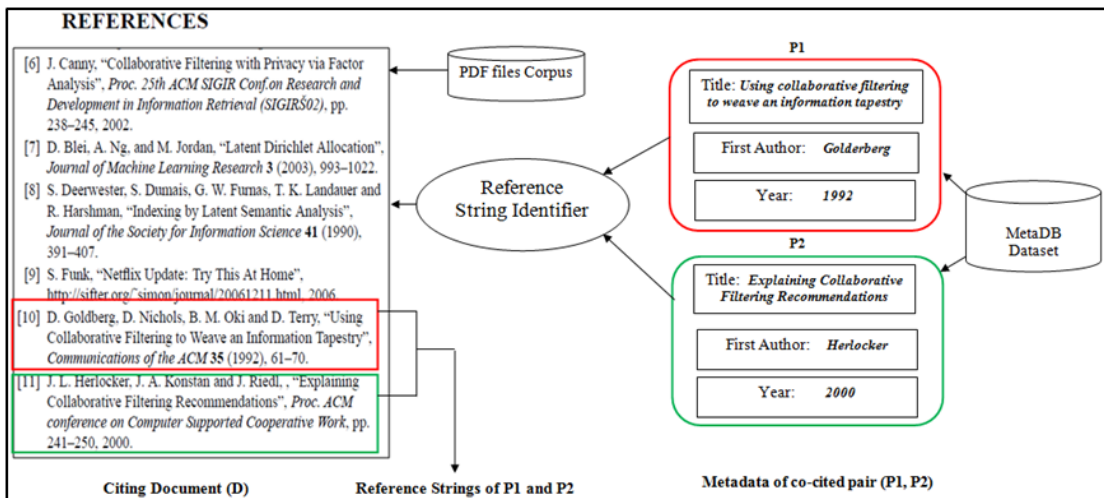


FIGURE 3.4: Reference string extraction

The ‘first author name’, ‘number of authors’ and ‘year’ information are also used to construct the citation anchor in case of those references which have no citation-tags as given in Figure 3.5. All the above metadata are extracted by using the metadata extractor in data preparation.

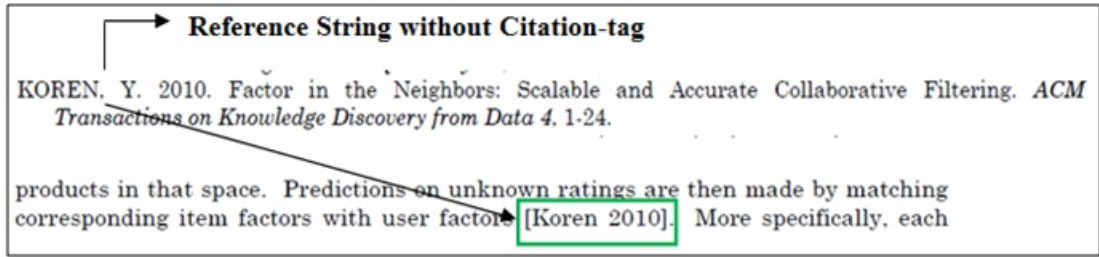


FIGURE 3.5: Reference-string without citation-tag problem

3.1.3 MetaDB Manager

The retrieved metadata is stored in metadata DB by metaDB manager for dataset preparation. The metadata of some queried papers is given in Figure 3.6. In the same way, all above mentioned metadata is also extracted for the co-cited documents. The nine co-cited documents or co-citations are mentioned for each query paper on the CiteSeer site. Therefore, we will select nine co-cited documents for each query paper.

Query paper	citations	citationid	Title	Number of Authors	First Author	Year	Doi
QP1	394	112703	Explaining Collaborative Filtering Recomm	3	Herlocker	2000	doi=10.1.1.32.252&rank=4
QP2	330	24487	Latent Semantic Models for Collaborative f	1	Hofmann	2004	doi=10.1.1.331.3636&rank=8
QP3	211	112227	Clustering Methods for Collaborative Filter	7	Ungar	1998	doi=10.1.1.33.4026&rank=9
QP4	233	10136101	Collaborative filtering with temporal dynar	1	Koren	2009	doi=10.1.1.379.1951&rank=10
QP5	368	121796	Eigentaste: A Constant Time Collaborative	4	Goldberg	2000	doi=10.1.1.37.8212&rank=14

FIGURE 3.6: Extracted metadata of query paper

The set of co-cited pairs and citing papers are prepared by using the co-cited pair extractor. This component uses the ‘citationid’ metadata to get the common citing documents between a query paper and a cited document. The set of citing documents is represented in Equation 3.1.

3.1.4 Co-cited Pairs and Common Citing Documents Extraction

It is a very important component of the data preparation phase. This part is used to prepare the set of co-cited pairs (cited documents) and the set of citing

documents. The query papers (X) are retrieved based on key-terms. The set of co-cited pairs (CCPs) of research papers are prepared based on metadata of query papers (X) and co-cited papers (Y). The set of common citing documents (D) for each co-cited pair can be obtained by the intersection of citing papers of “x” and citing papers of “y” as shown in Equation 3.1. Each pair is represented by (x, y). The ‘x’ and ‘y’ are co-cited papers. The set (D) of common citing documents can be represented by Equation 3.1.

$$D = \{p \mid \forall (x, y) \in CCPs, p \in (citedby(x) \cap citedby(y))\} \quad (3.1)$$

The above equation is explained by an example.

Let us take the set of co-cited pairs (CCPs). The CCPs set consist of four pairs of co-cited documents i.e $CCPs = (x_1, y_1), (x_1, y_2), (x_1, y_3), (x_1, y_4)$. In all pairs, the document x_1 is co-cited with other documents ($y_1, y_2, y_3, \text{ and } y_4$).

$$CCP = \{(x_1, y_1), (x_1, y_2), (x_1, y_3), (x_1, y_4)\}$$

Now, we take the pair of co-cited documents (x_1, y_3) from the set CCPs to find the set of common citing documents (D) in it.

$$p = (x_1, y_3)$$

For the set (D) preparation, it is necessary to get the cited-by sets of both co-cited-documents x_1, y_3 in pair P. Let us take the citation identifiers of x_1 and y_3 for further understanding of the given equation.

$$Citedby(x) = \{101, 102, 103, 104, 105, 106, 107, 120, 121\}$$

$$Citedby(y) = \{101, 103, 107, 108, 109, 112, 114, 120\}$$

The set of common citing document (D) for co-cited pair $p(x_1, y_3)$ can be obtained by the intersection of cited-by of x_1 and cited-by of y_3 . The pair (x_1, y_3) is co-cited in citing documents with citation identifiers (101, 103, 107, and 120). The same process will be repeated for all set of co-cited pairs (CCPs).

$$D = Citedby(x) \cap Citedby(y) = \{101, 103, 107, 120\}$$

The sets of co-cited pairs (*CCPs*) and citing documents (*D*) will be used in section wise co-citation analysis as will be discussed in chapter 6.

3.1.5 Citing papers downloader

The ‘doi’ metadata is used by citing papers downloader to download the PDF files for the common citing documents as shown in Figure 3.7.

```
"http://citeseerx.ist.psu.edu/viewdoc/download?.$doi."&rep=rep1&type=pdf";
```

FIGURE 3.7: Paper download link on CitSeer site

3.1.6 PDF to Text and PDF to XML Convertors

In our research task, we are considering two formats of PDF file (1) Plain-text format and (2) XML format. In this component, the PDF file is converted into two formats by using PDF to Text and PDF to XML convertors respectively. Both formats files will be used as input for second phase of our proposed architecture as shown in Figure 3.1.

3.2 Section Wise Co-citation Analysis Phase

The section wise co-citation analysis is the main phase of our research work. This phase consists of three main steps (a) Generic sections/ILMRaD structure identification (b) In-text co-citation patterns and frequencies identification, and (c) Section wise co-citation analysis (SWCA) as highlighted by red circles in Figure 3.1. As the first step, the sections of citing documents are extracted and mapped on the generic section or ILMRaD structure. In the second step, the rule based approach is applied to find the patterns and frequencies of co-cited documents in the text of

citing documents. In the third step, The co-citation analysis has been performed across the generic sections or ILMRaD structure of citing documents.

(a) Generic Sections or ILMRaD Structure Identification

Generic sections identification is the first component of section wise co-citation analysis phase. In this step, we have extracted the sections in citing documents and then mapped these sections on (ILMRaD) structure by three proposed methods (1) Section headings labels based analysis (2) In-Text patterns based analysis and (3) Pages and structural components based analysis. In section heading labels based analysis; the sections are mapped on the (ILMRaD) structure based on the section heading. In in-text patterns based analysis, the sections are mapped based on some in-text patterns and defined rules. In pages and structural components based analysis, the sections are mapped based on pages, sections, and pre-defined set of section patterns. The details discussion is given in chapter 4.

(b) In-Text Co-citation Patterns and Frequencies Identification

In-text co-citation patterns and frequency identification is the second step to find the patterns and frequencies of citations in the text of citing documents. The accuracy of co-citations frequencies depends on the accurate identification of citation-tags and citation-anchors. This section consists of four key components including (1) Reference string identifier (2) Citation-tag identifier (2) Mapping section and (4) Citation-anchors taxonomy. The details of this part are given in chapter 5.

(c) Section Wise Co-citation Analysis

The third and final component is the section wise co-citation analysis. This component depends on the output of the first two main components as mentioned above. In this section, we have calculated the document relevancy score between co-cited documents using ILMRaD structure along with section weights and co-citation frequencies. The details of this component has been given in chapter 6.

3.3 Document Ranking and Result Evaluation Phase

This phase has been divided into two parts (1) Document Ranking and (2) Result Evaluation.

3.3.1 Document Ranking

The documents are ranked based on the document relevancy scores produced by the previous phase. The papers with highest relevancy score will come on the top of the ranked list as discussed in chapter 6.

3.3.2 Result Evaluation

This section explains the evaluation process of the proposed approach. The proposed approach consists of three important contributions: (1) Generic sections identification (2) In-text co-citation patterns and frequencies identification, and (3) Section wise co-citation analysis(SWCA). The accuracy of each of mentioned components needs to be comprehensively evaluated. The details of each of above mentioned respective contribution have been given in chapter 4, chapter 5, and chapter 6 respectively. The evaluation of each contribution is also included in the same chapter.

Chapter 4

Identification and Mapping of Sections on ILMRaD Structure

Note: The part of this chapter has been published in conference¹

In chapter 3, the three main tasks in second phase (Section Wise Co-citation Analysis Phase) of proposed approach were highlighted in Figure 3.1. This chapter explores the first core component “Generic Sections ILMRaD Structure Identification” of the section-wise co-citation analysis phase. ILMRaD is the short form for Introduction-Literature-Methodology-Result and Discussion.

In this chapter, first we will analyze the ILMRaD structure in research papers. Second the proposed architecture has been defined for the identification of generic sections or ILMRaD structure in research documents. Subsequently, the proposed approach and state-of-the-art technique [28] are implemented over the two datasets. Finally, the experimental results of proposed approach are compared and evaluated with the state-of-the-art technique [28].

¹Ahmad, R., Afzal, M. T., and Qadir, M. A. (2016). Information extraction from pdf sources based on rule-based system using integrated formats. In the semantic web: ESWC 2016 Challenges, Anissaras, Crete, Greece. 641:293-308, Communications in computer and information science. Springer. [A Category Conference], Challenge Winner paper.

4.1 ILMRaD structure Analysis

Usually, most of the academic research articles for various journals and conferences are prepared by different combinations of structural components, such as Title, Authors, keyword, Abstract, Introduction, Related Work (literature), Methods, Experiment, Results, Discussion, Future work, Conclusion, Acknowledgement, and References [23, 27]. However, the majority of research articles follow standardized or generic structure IMRaD [78] explicitly or implicitly Introduction, Methods, Results and Discussion. The IMRaD structure of scientific papers is properly followed in BioMedical domains.

On the other hand, the research papers of computer science domain also consist of “Related work(Literature)” section. Therefore, in this research work, the “Literature” section is also considered with IMRaD structure. This modified structure will be referred to as ILMRaD in the rest of the document. The definition of each generic section represented in ILMRaD is as follows: The “Introduction” section is followed by the abstract section in majority of research papers and normally the term “Introduction” is used by most of papers to represent this section. The term “Related work” is used to represent the literature part in the citing documents. The “Methodology” section in ILMRaD structure is rarely represented with the terms “Method”, “Methodology” and “Methods and Materials” as shown in Table 4.1 [28]. According to the experiment of 329 research papers in Shahid and Afzal approach, the section “Introduction” was noted as the most frequent section in 78% of the documents, the section “Introduction” was referred with the same names. However, the section “Methodology” was not referred even a single time with the term “Methodology”. The section “Related Work” was referred with the same or similar terms “Related Work” in only 30% of the documents. The section “Results” was mentioned with the term “Results” only by 1% of the documents.

TABLE 4.1: Manual classification of section labels of structural components [28]

Class Name	Total papers	Entries	Section label same as standard sections labels	Section label different from standard sections labels
Introduction	329	378	78%	22%
Related Work	158	184	30%	70%
Methodology	322	829	0%	100%
Results	59	62	1%	99%
Discussion	95	110	20%	80%
Conclusion	263	270	60%	40%

However, mostly the “Methodology” section is represented with different number of structural components [27, 79]. Usually these structural components are used with different headings in research papers such as “Problem Definition and Architecture”, “The Candidate Set”, and “Modeling content-based Citation Relevance” as shown in Table 4.4. The “Result” section in ILMRaD structure is prepared by the various combinations of structural components such as “Experiment”, “Evaluation” and “Results”. The last generic section “Discussion” in ILMRaD structure is also prepared by different combinations of “Discussion”, “Future work”, “Conclusion” in various research documents.

Apart from the above mentioned experiment, another experiment has been performed in this research thesis and that experiment again evaluated whether the scientific authors use the similar name of the sections as per their generic section names. Therefore firstly, the generic sections have been analyzed based on their occurrences in research articles. In this analysis, it is observed that the “Introduction” section mostly existed with 94% in 211 research papers. The other sections have been found with different section labels. It is also observed that the methodology, result, and discussion sections widely occurred with different section labels in research papers. This discussion motivated us to map section headings onto

the logical sections. This will help us to achieve the overall task of section wise co-citation analysis. The details of proposed approach are as follows.

TABLE 4.2: Manual classification of section labels over 211 research papers

Generic Sections	Section label same as standard sections labels	Section label different from standard sections labels
Introduction	94%	6%
Related Work	39%	61%
Methodology	1%	99%
Result	5%	95%
Discussion	6%	96%
Conclusion	26%	74%

4.2 Proposed architecture for ILMRaD Structure Identification

Manually, mapping structural components onto generic sections is very difficult for the large corpus of research papers. Therefore, the architecture has been proposed and designed for the automatic identification of generic sections as shown in Figure 4.1. The proposed architecture consists of three phases (1) Structural component heading extraction phase (2) Structural component splitting and mapping phase, and (3) Generic section evaluation phase. In first phase of our proposed architecture, the heading labels, the contents boundary and page number of each structural components in research papers have been extracted and then they are stored for the next phase processing. The second phase splits and maps the structural components of research papers on the generic sections. In the third step, we have evaluated the corpus of generic sections research papers using the developed gold standard. The detail of each phase is given below in respective sections.

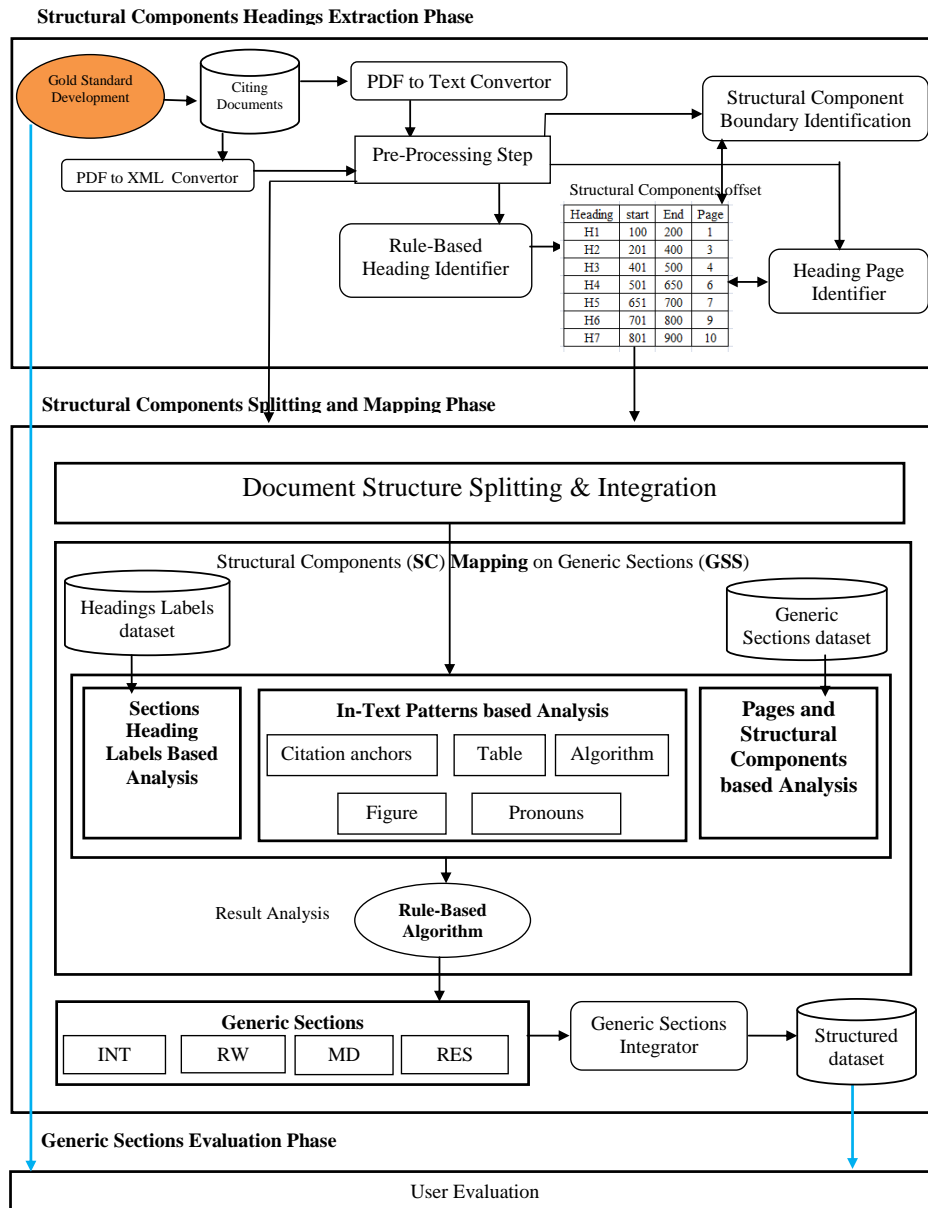


FIGURE 4.1: Proposed architecture for generic sections identification

4.2.1 Data Preparation

For the generic section identification task, two datasets have been prepared (1) Training dataset and (2) Testing dataset. The training dataset of 211 research papers is available by Nguyen and Kan [80]. The proposed technique for the generic

sections identification has been tested on the training dataset. The testing dataset is prepared by the combination of two annotated datasets of generic sections. Both of these datasets were made by extensive user studies by three researchers actively developing approaches which need section annotations. These annotated datasets have been selected for the evaluation of our proposed approach that was developed for generic section identification. The training dataset consisted of 211 research papers with 1,220 sections. The first test data consisted of 150 unique papers out of 499 papers and second test dataset consisted of 500 research papers.

4.2.2 Structural component heading extraction phase

It is the first step of generic section or ILMRaD structure identification. The citing documents in this phase are used as input. This section identifies and extracts the structural component information, such as “Heading labels”, “Content boundary”, and “Heading label page number”. Therefore, three main modules have been included in this phase along with some additional parts. The modules are (1) Rule-based heading identifier (2) Structural component boundary identification and (3) Heading page identifier. The additional parts are PDF to text convertor, PDF to XML convertor, and Pre-processing step. First, the PDF file is converted into plain-text or XML formats by PDF to text convertor or PDF to XML convertor and then the converted file is processed in the pre-processing step for further operations of three modules.

Module 1: *Rule-based heading identifier*

Each research paper is organized in different structural components with headings and body of contents by authors of research papers. We are interested to automatically identify the headings and corresponding content of each structural component. In different research papers, different types and styles of headings are being used to identify structural components. In this study of research, the headings taxonomy for structural components was constructed by comprehensive evaluation of research papers published in diversified venues. This taxonomy as presented in Figure 4.2 shows types and styles variations of headings. Figure 4.2

presents two categories of headings (1) With numerals and (2) Without numerals. The “with numerals” category of structural component heading is classified into two sub-categories (a) Numeric numerals and (b) Roman numerals. Numeric numerals consist of four types of headings such as ‘Uppercase’, ‘Title Case’, ‘Sentence Case’, and ‘Mixed Case’. All the cases of heading in Numeric numerals category are started with numbers, such as “1. INTRODUCTION & MOTIVATION”, “1 Introduction & Motivation”, “1. Introduction & motivation”. The roman numerals headings are started with roman numbers, such as “I INTRODUCTION”, and “II INTRODUCTION & MOTIVATION”. The “without Numerals” heading category also consists of four types without Numeric and roman numerals, such as “INTRODUCTION & MOTIVATION”, “Introduction & Motivation”, and “Introduction & motivation”.

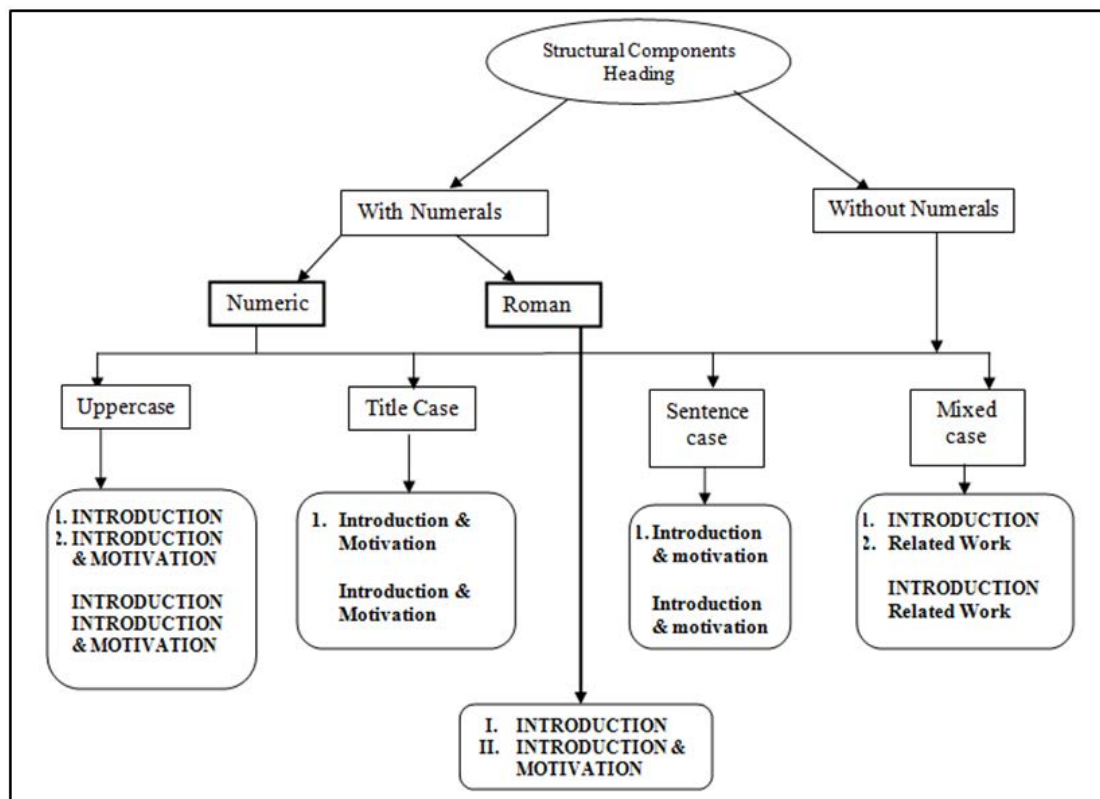


FIGURE 4.2: Heading taxonomy for structural components

The module “Rule-Based Heading Identifier” uses the headings taxonomy to identify the headings labels of structural components in an automatic way and then

it stores the headings labels in the structural component offset dataset for future use.

Our initial experiment was conducted over training dataset of 211 research papers to evaluate the occurrences of headings taxonomy. The statistics of experiment are automatically prepared as given in Table 4.3.

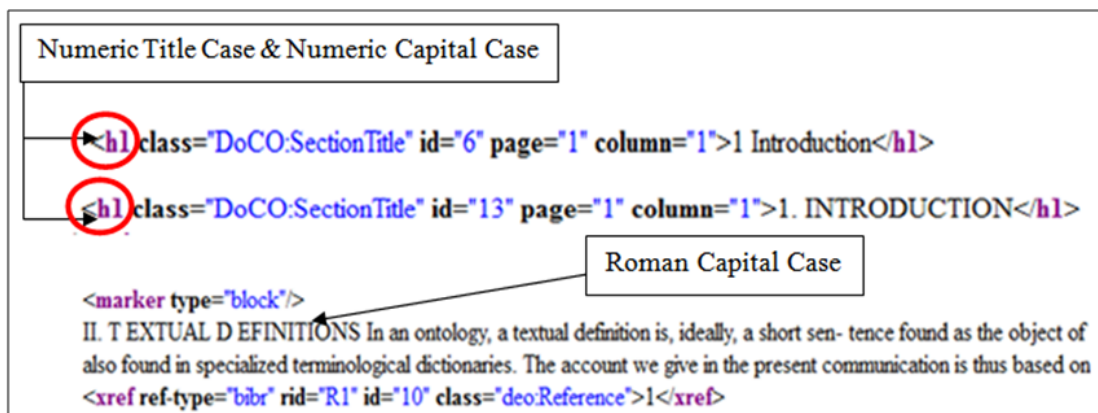
TABLE 4.3: Heading analysis of structural components based on formats

Heading analysis within 211 research papers	
Heading formats	Total number of papers
Numeric with capital case	155
Numeric with title case	28
Numeric with sentence case	6
Numeric with mixed case	2
Roman with capital case	10
Capital case without Numeric	8
Sentence case without Numeric	2
Title case without Numeric	4

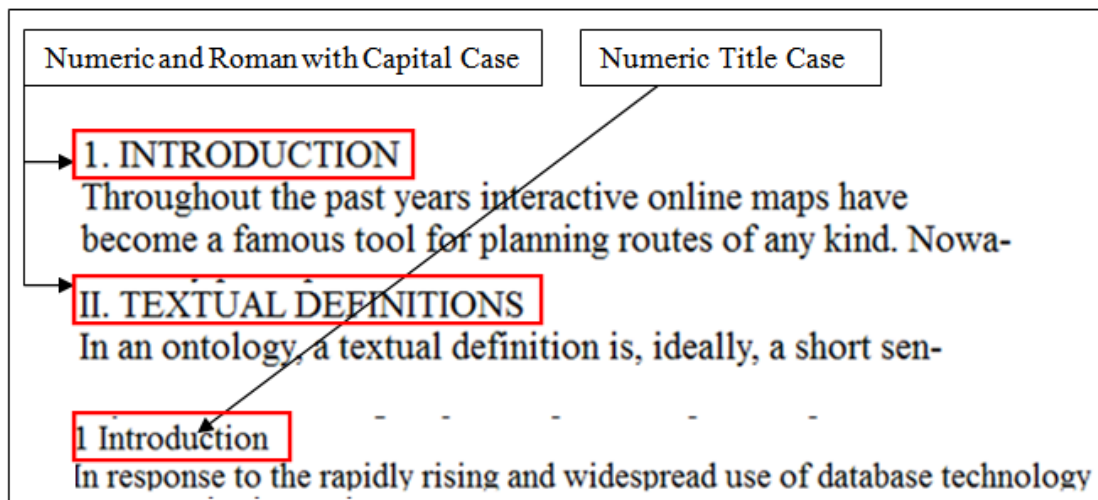
This Table shows that two heading cases such as ‘Numeric with capital case’ and ‘Numeric with title case’ are widely used for heading selection of structural components in research papers. The two formats such as ‘Numeric with capital case’ and ‘Roman with capital case’ are widely observed in the IEEE Journals such as ‘*Journal of Transactional Engineering in Health and Medicine*’, ‘*Journal for Computing*’ etc and ACM standard Journals template that also follows the ‘Numeric with capital case’ for heading selection.

In the analysis of ‘first level section heading, the three formats of section heading in PDF documents are considered as shown in Table 4.3. The first format ‘Numeric with capital case of section heading consists of section ‘heading number (1) and heading name INTROUDCTION in capital case. The second format ‘Roman with capital case is denoted by ‘roman heading number (II) and ‘uppercase name RELATED WORK. The third format of section heading is represented by ‘numeric

heading number (1) along with ‘heading name Related Work in title case. The extraction of all headings formats from the XML document is not completely possible. Therefore, for the first level section heading analysis, two formats are considered such as XML document and plain-text in this research thesis. Let us see the scenario of both formats in below figure 4.3.



(a)



(b)

FIGURE 4.3: Analysis of section headings in both XML and Plain-text formats
a) Snapshot of first level section headings in XML format b)Snapshot of first level section headings in plain-text format

In Figure 4.3(a), the various formats of first level section headings are highlighted in XML formats of PDF documents. The PDFx tool properly assigns the <h1> tag to section heading in both cases Numeric title case and Numeric capital case after conversion of PDF document. While in the roman capital case, the PDFx tool does not assign any tag. This analysis shows that the XML format is better

for Numeric Title and Capital cases of section headings. The roman title case might not be detected from the XML format. Therefore, in this case, we are using the Plain-text format of PDF documents in our analysis. The snapshot of Plain-text format of section headings is given in Figure 4.3(b). The Plain-text format is also suitable for the extraction of section headings in numeric and roman with capital cases. The numeric with title case is not properly extracted due to the commonality in content.

Section Heading Extractor

The section heading extractor function in our proposed rule-based approach has been designed to extract the heading labels of structural components. This function gets two formats of a research paper as inputs, such as PDFfile and XML file. The function also contains three functions ,i.e., “Section_Heading_Recognizer”, “Section_Heading_Refiner”, and “Section_Heading_Splitter”. The inputs of first function are PDFfile and XMLfile. This function returns the set of section labels and store in “sectionHeadingArr” array. These headings are further refined by the “section_Heading_Refiner and its return the refined set of section labels”. The refined set of section labels are finally classified and structured by the “section_Heading_Splitter”. The section heading extractor exploits the heuristic and rules which exist in the form of regular expressions. All the regular expressions are verified over the content of both XML and Plain-text formats in “EDITpad Pro 74” tool³ and then it is used in java code.

³<https://www.editpadpro.com/>


```

1: function SECTION_HEADING_EXTRACTOR(PDFFILE, XMLFILE)
2:   sectionNo := “ ”
3:   sectionName := “ ”
4:   plaintext := PDFbox(PDFfile)
5:   sectionHeadingArr [ ] := Section_Heading_Recognizer(plaintext, XMLfile)
6:   sectionHeadingArr := Section_Heading_Refiner(sectionHeadingArr)
7:   i := 0
8:   While i < sectionHeadingArr.length
9:     HeadingArr[ ] := Section_Heading_Splitter(sectionHeadingArr(i))
10:    sectionNo := HeadingArr[0]
11:    sectionName := HeadingArr[1]
12:    stored (sectionNo, sectionName)
13:    i := i + 1
14:   End loop
15: end function

1: function SECTION_HEADING_RECOGNIZER(PLAINTEXT, XMLTEXT)
2:   HeadingArr[ ] := Numeric_Capital(plaintext)
3:   IF (HeadingArr.length < 3)
4:     HeadingArr := Roman_Capital(plaintext)
5:
6:   IF (HeadingArr.length < 3)
7:     HeadingArr := Capital_Case(plaintext)
8:
9:   IF (HeadingArr.length < 3)
10:    HeadingArr:= Numeric_Title_Sentence_Case(plaintext)
11:
12:   IF (HeadingArr.length < 3)
13:    HeadingArr := XML_Heading(XMLtext)
14:   return HeadingArr;
15: end function

```

Section Heading Recognizer

The section heading recognizer has the ability to identify “Numeric Capital Case”, “Roman Capital case”, “Capital_Case”, and “Numeric_Title_Sentence_Case” headings in the processed content of XML or plain-text format. The following regular expressions are built to extract different types of headings.

Regular Expression (1) for Numeric Capital Case:

$\backslash n \backslash d + \backslash . ? \backslash s * [\backslash p \{ Lu \} : 0 - 9 \backslash s \& -] *$

The symbol newline ‘ $\backslash n$ ’ occurred at the start of the section heading. The $\backslash d +$ symbol shows one or more than one occurrences of digits. The dot symbol is optional after the digit. The symbol $\backslash s *$ represents zero or more than zero spaces

after digits and dot symbols. Usually, the section heading may be contains the symbols, such as capital alphabets, numbers, dash sign, colon, and & sign. The part `[-\p{Lu}:0-9\s&-]*` of regular expression is used to represents such types of symbols in the label of section heading.

Regular Expression (2) for Roman Capital Case:

`(\n[IVX]* \s([\p{Lu}0-9\u2019&-/*\}s?)*\r\n|REFERENCES)`

The symbols `\n[IVX]*` are used to represent newline and roman characters in the start of roman capital heading. These symbols `\p{Lu}` are used to extract the capital letters in section label while 0-9 denote the numeric symbols in heading. The unicode character `\u2019` is used for right single quotation mark.

Let us see the scenario of the ‘Roman with capital case’ which is identified from the Plain-text format by using regular expression (2) as shown in Figure 4.4. This regular expression extracts the highlighted headings along with carriage return characters (`\r\n`). The regular expression has been verified in the ‘EDIT pro 7 tool’. The carriage return can be remove by using some pre-processing on extracted headings. Finally, different rules are defined in the form of regular expressions to detect the other cases of section headings along with pre-processing steps. The roman numbers are replaced with numeric numbers, such as ‘I,’II with ‘1,’2.

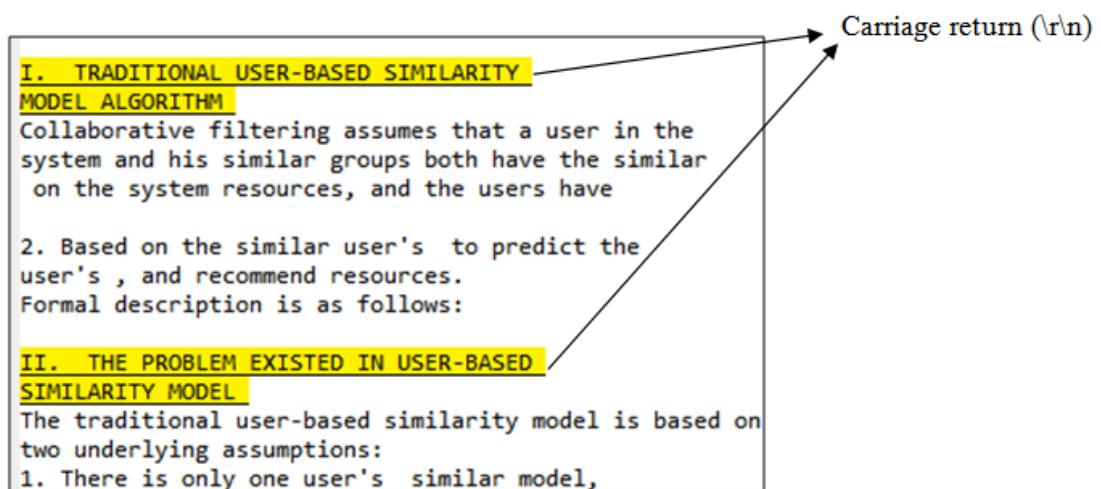


FIGURE 4.4: Roman with capital case detection

Regular Expression (3): Numeric Title Sentence Case:

`\n\d\.\?\s+([A-Z0-9]{1,2}[a-z:\.,-]* \s)*[\x20-\x7E]`

The symbols such as `\n\d\.` are used in regular expression to represent “newline, digit, and dot” characters. The characters “{1,2}” in this pattern `[A-Z0-9]{1,2}` show one or two occurrences of either capital alphabets or numbers. It represents the start character of each word which is in capital form while the pattern `[a-z:\.,-]*` are used to extract the lowercase letter, dot character, colon, comma, and dash sign in the section heading. This pattern `[\x20-\x7E]` are used to remove non-ascii characters from section heading.

Section Heading Refiner

The section heading refiner is used to remove the wrong patterns and additional characters from the output of section heading recognizer. The input of this function is a set of section heading labels of structural components with extra characters such as carriage return and newline. This function returns the refined set of section heading labels as output.

Section Heading Spitter

Finally, the splitter separates the structured elements such as ‘section number and ‘section title as mentioned in Figure 4.6 from the first level section heading. The input of this function is the heading label of section and in a result the function returns two outputs ,i.e., section number and section name.

The functionality of section heading recognizer, section heading refiner, and splitter is explained in given scenario. In Figure 4.5, the PDF document is parsed into XML document by the PDFx tool. This tool represents the section heading by tags `<h1>` and `<h2>` due to different formats of section heading represented by rectangle in PDF document, such as ‘1. INTRODUCTION, ‘2. Background, ‘3. Visualization Approach, ‘4. Implementation, ‘5. Case Study, ‘6 CONCLUSION AND FURTHER WORK, and ‘7. REFERENCES. However, most of the time the `<h2>` tag is used to represent the second level heading, such as ‘2.2, ‘2.3, ‘4.1, ‘4.2, and ‘4.3. To solve the various formats problem with section heading, first we extracted all patterns of `<h1>` and `<h2>` by section heading recognizer. Second, the output of section heading recognizer is transferred to section heading refiner. The refiner removes the second level headings by using pre-processing and

also removes some additional characters such as ‘>, and ‘</h1> or ‘</h2> with section headings.

```

<h2 class="DoCO:SectionTitle" id="6" confidence="possible" page="1" column="1">General Terms Keywords</h2>
<h1 class="DoCO:SectionTitle" id="8" page="1" column="1">1. INTRODUCTION</h1>
<h2 class="DoCO:SectionTitle" id="17" confidence="possible" page="1" column="2">2. Background 2.1 Collaborative Writing Activities in Education</h2>
<h2 class="DoCO:SectionTitle" id="25" confidence="possible" page="1" column="2">2.2 Visualization Approaches for Collaborative Writings</h2>
<h2 class="DoCO:SectionTitle" id="32" confidence="possible" page="2" column="1">2.3 Representing Mental Models as Graphs</h2>
<h2 class="DoCO:SectionTitle" id="39" confidence="possible" page="2" column="1">3. Visualization Approach</h2>
<h1 class="DoCO:SectionTitle" id="45" confidence="possible" page="2" column="2">4. Implementations</h1>
<h2 class="DoCO:SectionTitle" id="47" page="2" column="2">4.1 Extracting Concept Networks from Texts</h2>
<h2 class="DoCO:SectionTitle" id="66" page="3" column="2">4.2 Networks from Different Revisions</h2>
<h2 class="DoCO:SectionTitle" id="69" page="3" column="2">4.3 Quantitative Characterization of Contributors</h2>
<h2 class="DoCO:SectionTitle" id="79" confidence="possible" page="4" column="2">5. Case Study</h2>
<h2 class="DoCO:SectionTitle" id="82" confidence="possible" page="5" column="1">Author Color EVC NBC EBC</h2>
<h1 class="DoCO:SectionTitle" id="85" page="5" column="1">6. CONCLUSION AND FURTHER WORK</h1>
<h1 class="DoCO:SectionTitle" id="87" page="5" column="1">7. REFERENCES</h1>
    
```

FIGURE 4.5: Section heading recognition in XML document by section heading recognizer

Finally, the refiner generates the accurate section heading. The output of refiner is further processed by the splitter to produce the structured elements such as ‘section Number and ‘section Title of each section heading in a research paper as shown in Figure 4.6.

Section Heading	Section-Number	Section-Title
1. INTRODUCTION	1	INTRODUCTION
2. Background	2	Background
3. Visualization Approach	3	Visualization Approach
4. Implementation	4	Implementation
5. Case Study	5	Case Study
6. CONCLUSION AND FURTHER WORK	6	CONCLUSION AND FURTHER WORK
7. REFERENCES	7	REFERENCES

FIGURE 4.6: Section heading conversion into structured elements

Module 2: Structural components boundary identification

The structural components of research papers have the body of text under specific heading labels. In second module “Structural components boundary identification”, the start and end byte address of each structural component body is identified by using the extracted heading labels in structural components offset

dataset. These start and end addresses are then stored in offset dataset for next phase. Figure 4.7 presents the structure of the research paper “He et al, 2010. Context-aware citation recommendation. In Proceedings of the 19th international conference on World Wide Web (pp. 421-430). ACM”. The structure is further divided into different structural components. The heading labels of each component are represented by numeric with capital case like “1. INTROUDCTION”, “2. RELATED WORK”, “3. PROBLEM DEFINITION AND ARCHITECTURE”, “4. THE CONDIDATE SET”, “5. MODELING CONTEXT-BASED CITATION RELEVANCE”, “6. EXPERIMENTS”, and “7. CONCLUSIONS”. The content boundary is denoted by “S” and “E” symbols of each structural component. “S” and “E” represent start and end addresses of text body in structural components. All these information of a concerned paper are stored in structural component offset for using in the next phase and has been shown in Table 4.4.

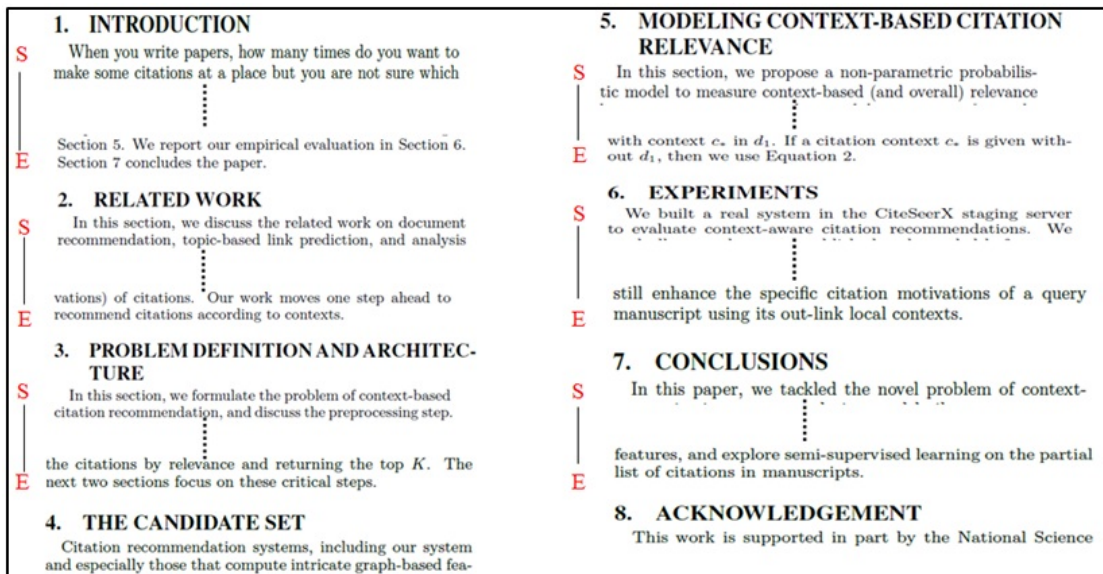


FIGURE 4.7: Structure of a research paper

In initial experiment of the training dataset, the structure of structural component offset dataset has been obtained automatically for a research paper in Figure 4.7. The structural component offset dataset as given in Table 4.4. This dataset holds information such as “Paperid”, “Heading labels”, “Content boundary information, start and end bytes addresses of the text body”, and “Heading page” of structural components. For example, Table 4.4 represents the heading

labels of structural components “1 INTRODUCTION”, “2 RELATED WORK”, “3 PROBLEM DEFINITION AND ARCHITECTURE”, “4 THE CANDIDATE SET”, “5 MODELING CONTEXT-BASED CITATION RELEVANCE”, “6 EXPERIMENTS”, and “7 CONCLUSIONS” of a research paper “S1” as shown in Figure 4.7. The content of the first structural component “INTRODUCTION of “S1” in the research paper is denoted by start and end byte addresses ‘1379’ and ‘8100’. The “Heading page information” holds the page number “1” of that page on which the heading of first structural component of “S1” occurred. All such information for a specific paper has been shown in Table 4.4.

TABLE 4.4: Structural components offset dataset of a research paper

Paperid	Heading	Start	End	Hpage
S1	1 INTRODUCTION	1379	8100	1
S1	2 RELATED WORK	8101	14324	2
S1	3 PROBLEM DEFINITION AND ARCHITECTURE	14325	17853	3
S1	4 THE CANDIDATE SET	17854	21641	4
S1	5 MODELING CONTEXT-BASED CITATIOIN RELEVANCE	21642	28982	5
S1	6 EXPERIMENTS	28983	45424	6
S1	7 CONCLUSIONS	45425	46274	9

Module 3: *Heading page identifier*

The heading page identifier has been designed in first phase to identify the page numbers on which the heading labels of the structural components occurred. The page numbers of structural components headings are used in the module (C) “Page and structural components based Analysis” of the “Structural components mapping and splitting phase” as shown in Figure 4.1.

4.2.3 Structural component splitting and mapping phase

The second phase of proposed architecture is the “structural components splitting and mapping phase as highlighted in Figure 4.1. This phase has been designed to map each of structural components in research papers on generic sections as shown in Table 4.7. It consists of three modules (A) Document structure splitting

& integration (B) Structural components mapping on generic sections and (C) Generic sections integration. In the first module, the structural components of research papers are divided and integrated by using structural components offset dataset as shown in Table 4.4. The second module has been designed to map the structural components on generic sections. The last module is the generic sections integration that are used to integrate the generic sections.

Module (A): Document structure splitting and integration

In module (A), splitting and integration of structural components of research papers is performed using structural components offset dataset, the splitting and integration of two components “Related work and “Methodology”, as shown in Figure 4.8. The “Related work” component has three sub-components such as “2.1”, “2.2” and “2.3”. While the “Methodology” component has two sub-components such as “3.1”, and “3.2”. Therefore, in the integration process, all the sub-components are combined with main structural component to make a compound structural component. Figure 4.8 shows the integration of structural components ‘2’ and ‘3’ with their sub-components by using red rounded rectangle and green rounded rectangle respectively.

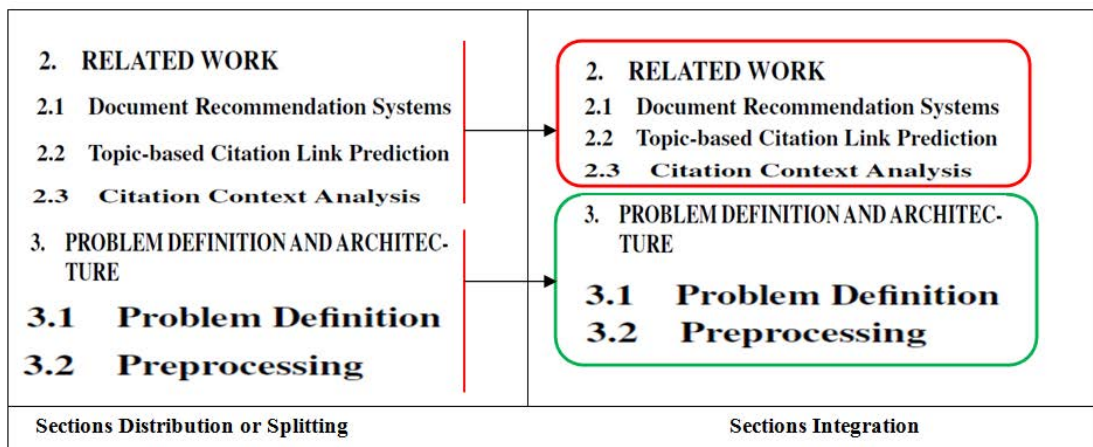


FIGURE 4.8: Document structure splitting and integration

Module (B): Structural components mapping on generic sections

In this module, structural components are analyzed and mapped on the generic sections. This module consists of four sub-modules (I) Section Headings labels

based analysis (II) In-text patterns based analysis (III) Pages and structural components based analysis, and (IV) Rule based algorithm. The decisions of first three sub-modules are recombined to make the final decision in sub-module (IV). Based on the final decision, the structural components are mapped on the generic sections.

Sub-module 1: *Section headings labels based analysis*

In the first sub-module, structural components of research papers are mapped on generic sections by using pre-defined keywords and stemming words that are developed over training dataset of 211 research papers as mentioned in Table 4.5. In Table 4.5, the section heading labels of structural components have been highlighted under their respective generic headings such as Introduction, Literature, Methodology, Results, Discussion and Conclusions. The generic sections are denoted by “INTR”, “LITR”, “MET” , “RES”, “DISC”, and “CON” respectively. The key and stemming words are retrieved after the detailed analysis of 1,220 section heading labels in 211 research papers.

TABLE 4.5: Key and Stemming words selection over training dataset of 211 research papers for heading label based analysis

Sec#	Generic Section	Heading Labels	Keywords	Stemming words
1	INTRODUCTION (INTR)	Introduction, Introduction and Background, Introduction and related work, Introduction & Motivation	INTRODUCTION	Intro
2	LITERATURE (LITR)	Related Work, Related Works, Introduction and related work, Related work and Discussion, Previous studies, Background & Related work, Related Works and Conclusions, Related Background, Prior work	Related, Previous, Prior	Relate, Previou, Prio
3	METHODOLOGY (MET)	Background, Overview, System Overview, The proposed approach, our approach, implementation, System implementation, System and service implementation, Methodology, Research Method, Research objective & Methodology, Problem Definition, System Architecture, Simulation	Background, Overview, Proposed, Approach, Implementation, Method, Definition, Architecture, Simulation	Backgr, Overvi, Propos, Appro, Implement, Simula
4	RESULTS (RES)	Result, Results, Results map, Evaluation, Experimental Evaluation, Evaluation & Results, Experiments, Experimental setup, Analysis, Experiment and Analysis	Result, Results, Evaluation, Experiments, Experimental, Analysis	Result, Evalua, Analy, Experime
5	DISCUSSION (DISC)	Discussion, Result and Discussion, Discussion and Conclusion	Discussion	Discuss
6	CONCLUSIONS (CON)	Future work, Future works, Concluding Remarks, Conclusions and Future work, Conclusion and Future plan, Conclusion and Direction of Future research, Conclusion and Future study, Summary, Summary and Conclusion, Limitations and Future work, Final Remarks	Future, Concluding, Conclusions, Conclusion, Summary, Final	Futur, Conclu, Summa, Final

In Table 4.6, the structural components heading labels are mapped on the generic sections. The total number of section heading labels is 1,220 that were extracted from the training dataset of 211 research papers during data preparation phase (section 4.2.2). In the heading based analysis, 56% section headings of structural components are mapped over the generic sections using the stemming words of respective generic sections. The remaining 44% unmapped section heading labels are mapped by using two other proposed methods as discussed in next modules.

TABLE 4.6: Generic sections identification based on stemming words in 211 training dataset of research papers

	Papers	Number of Heading labels	Mapped section heading labels on generic section	Unmapped section heading labels on generic section
Training Dataset	211	1,220	56%	44%

In Table 4.7, the structural components of a research paper have been mapped on the generic sections based on section heading labels based analysis, as shown in Figure 4.7. Table 4.7 shows that the structural components “1 Introduction”, “2 Related work”, “6 Experiments” and “7 Conclusions” have been mapped on generic sections “INTRODUCTION”, “LITERATURE”, “RESULTS and DISCUSSION” respectively. The structural components “3 Problem definition and architecture”, “4 The Candidate Set”, and “5 Modeling content-based citation relevance” have not been mapped onto any generic section such as “METHODOLOGY”. As in the start of main section, it has been discussed that most of the authors represent the methodology section in research papers with different number of structural components and structural headings.

TABLE 4.7: Structural components mapping on generic sections

Paperid	Section heading labels	Start	End	GS.ID	Generic Section
S1	1 INTRODUCTION	1379	8100	1	INTRODUCTION
S1	2 RELATED WORK	8101	14324	2	LITERATURE
S1	3 PROBLEM DEFINITION AND ARCHITECTURE	14325	17853		unmapped
S1	4 THE CANDIDATE SET	17854	21641		unmapped
S1	5 MODELING CONTEXT-BASED CITATION RELEVANCE	21642	28982		unmapped
S1	6 EXPERIMENTS	28983	45424	4	RESULTS
S1	7 CONCLUSIONS	45425	46274	6	CONCLUSION

The function “Keyword_Mapping” has been written for section mapping. This function is also used in RBA (Rule Based Algorithm) as given in section 4.2.4. The input parameters for this block of code are section heading label and stemmed words dataset. The output of this function is Generic section id which is return to RBA algorithm. The pseudocode of “Keyword_Mapping” function is shown below. This function uses the predefined stemword_dataset to map the candidate section heading label on ILMRaD structure.

```

1: function KEYWORD_MAPPING(HEADINGLABEL, STEM_DATASET)
2:   i := 0
3:   GSID := 0
4:   While stem_dataset (i) != null
5:     IF stem_dataset(i) == headinglabel Then
6:       GSID := headingid(i)
7:       Break
8:     End IF
9:     i := i + 1
10:  End loop
11:  return GSID
12: end function

```

Sub-module 2: *In-text patterns based analysis*

In the previous sub-module, the “section heading labels based analysis” is conducted to map the structural components on generic sections using the key and stemming words dataset. However, some of the components of a research paper in table 4.7 did not map in the sub-module “section heading labels based analysis”. Hence, the sub-module “In-text pattern based analysis” has been included in the second phase of “Generic sections/ILMRaD structure identification” phase in the

proposed architecture as shown in 4.1. In this part, mapping of structural components on generic sections is further evaluated by using in-text patterns. The structure of research papers contains regular in-text patterns “Citation-Anchors”, “Figure”, “Table”, “First person plural pronoun”, and “Algorithm” which might be beneficial to identify the unmapped sections labels.

The authors of research papers use citations to support their research work in the research papers. The “Citation-Anchors” patterns are used to represent the citations in the text of research papers. These patterns have been observed mostly in the “Introduction”, “Related Work” sections [81]. Hence, “Citation-Anchors” can be helpful in the identification of generic sections, “Related Work”. The identification of citation-anchors has been comprehensively discussed in chapter 5. For the short view, here in Figure 4.9, the patterns ”Citation-Anchors” have been highlighted from the research papers. The numeric citation-anchors are represented in red circle. While the string citation-anchors are denoted by red oval shape. The regular expression 1 has been built to access the frequency of numeric and string citation-anchors patterns from the text of citing documents.

Regular Expression 1:

```

\\[[1-9][0-9]*\\|\\s*([1-9][0-9\\u2013]*\\s*[;,|-|\\u2013](\\s|\\))*)+[1-9][0-9]*\\s*
(-[1-9][0-9]*)?\\|\\|[[1-9][0-9]*-|\\u2013][1-9][0-9]*\\s*\\|\\|([A-Za-z][A-Za-z+\\}.\\s]*[0-9]2[0-9]*|(\\s)?)*\\

```

\\[[1-9][0-9]*\\]: This part of regular expression represents the citation-anchors of one or more than one digits such as [1], [22] in text of citing document. The ‘*’ sign shows zero or more than zero occurrences of second digit position and onward. The \\u2013 encode character is used to represent the hyphen character. The ‘\\s’ character shows the space occurrence in citation-anchor. The ‘?’ symbol shows the zero or one occurrence of any character in regular expression. The pipe ‘|’ sign is used to combine more than one regular expressions.

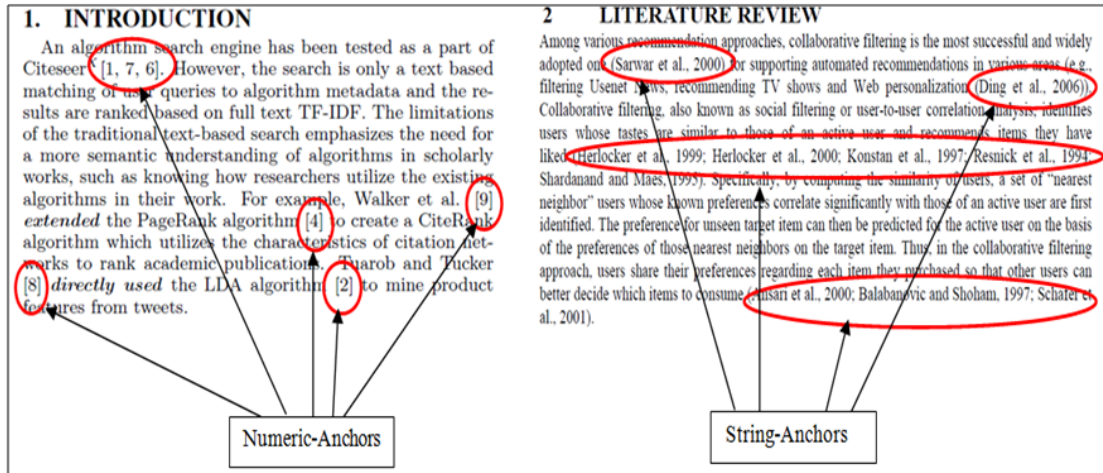


FIGURE 4.9: Snapshot of “citation-anchor patterns” from research papers

The second pattern is the “Figure” that shows the trends and features of the research work in the research papers. According to Nair & Nair [23] “Figure” pattern is the essential part of well presented scientific papers. This has been highlighted by many researchers [82, 83] that majority of “Figures” are used in the result section. Therefore, the pattern “Figure” and numeric literals “1” will be searched in each structural component of the research paper and the component have more occurrences of the “Figures” will be considered and marked as the “Result” section. Figure 4.10 shows snapshot of only four occurrences of “Figure” pattern out of eight observed patterns in the result section of IEEE standard research paper “Cai et al, 2014; Typicality-Based Collaborative Filtering Recommendation; IEEE Transactions on Knowledge and Data Engineering”. The occurrences of “Figure” pattern in this research paper show the importance of it in the result section.

Regular expression 2: $[f[F]ig[a-zA-Z\s\.\,]*\d[A-Za-z,()\s]*\r$

The regular expression 2 is exploited to count the frequency of “Figure” pattern in text of citing documents. This pattern of regular expression $[f[F]ig[a-zA-Z\s\.\,]*\d$ is used to represent the patterns like ‘Fig 1 or fig 1’, ‘Fig 1. or fig 1.’, and ‘Figure 1 or Figure 1.’. The remaining part of regular expression $[A-Za-z,()\s]*\r$ is built to extract the text of the figure caption.

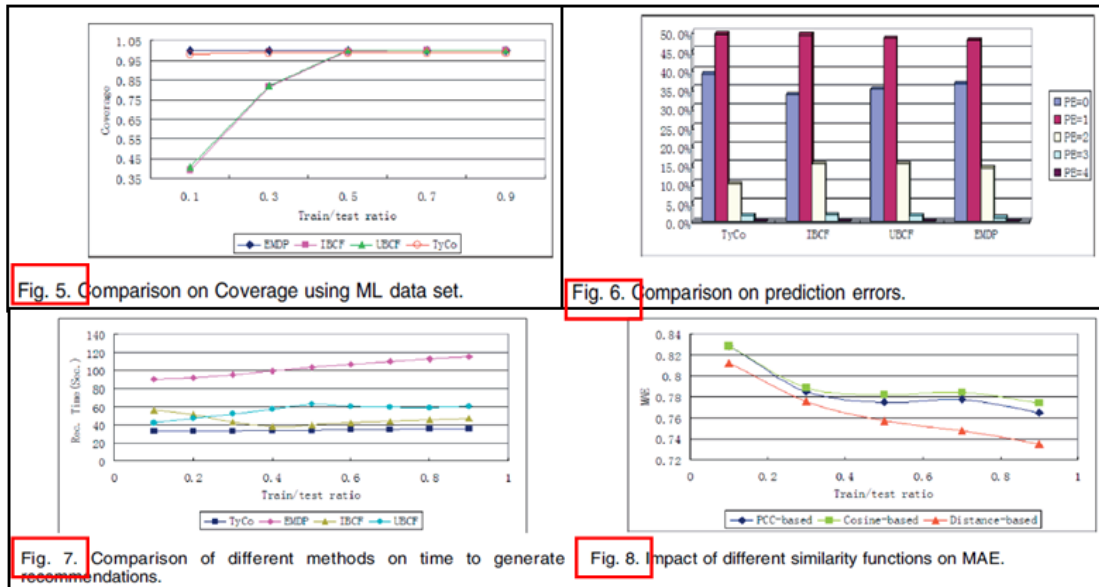


FIGURE 4.10: Snapshots of “Figure patterns” from a researcher paper

The “Table” patterns are other important patterns like “Figure” pattern that are widely used in the “Methodology” and “Result” sections of research papers [23, 84]. This “table” pattern is exploited in the “Methodology” and “Result” sections of research papers to provide the complete details in statistical form about the new method for the understanding of it in simple way. The snapshot of the “Table” pattern has been taken from “Methodology” and “Result” sections of a research paper “Cai et al, 2014; Typicality-Based Collaborative Filtering Recommendation; IEEE Transactions on Knowledge and Data Engineering” as shown in Figure 4.11 with red rectangle. The patterns “Tables” and “Figure” might be a good indicator to detect the “results” sections. These two patterns “Figure” and “Table” along with captions are extracted in our published work [85] with average F-score of 77%. The regular expression 3 is used to count the occurrences of “Table” pattern in citing documents body.

Regular Expression 3: $[t|T]ab[a-zA-Z\backslash\cdot]^*\backslash d[A-Za-z,()\backslash\cdot]^*\backslash r$

The regular expression 3 is exploited to count the frequency of “Table” pattern in text of citing documents. This pattern of regular expression $[t|T]ab[a-zA-Z\backslash\cdot]^*\backslash d$ is used to represent the patterns like ‘Tab 1 or tab 1’, ‘Tab 1. or tab 1.’, and ‘Table

1 or Table 1.’. The remaining part of regular expression $[A-Za-z,()\s]*\r$ is built to extract the text of the table caption.

TABLE 1							TABLE 2						
Sensitivity of n on MAE with Different Train/Test Ratios							Sensitivity of n on Coverage with Different Train/Test Ra						
	γ	$\chi=0.1$	$\chi=0.3$	$\chi=0.5$	$\chi=0.7$	$\chi=0.9$		γ	$\chi=0.1$	$\chi=0.3$	$\chi=0.5$	$\chi=0.7$	$\chi=0.9$
n=5	0.8	0.8106	0.771	0.7478	0.7436	0.7361	n=5	0.8	0.9401	0.965	0.9711	0.9699	0.9773
n=10	0.7	0.8115	0.7771	0.7546	0.7451	0.7394	n=10	0.7	0.9637	0.9794	0.9795	0.9792	0.9847
n=15	0.6	0.8117	0.7774	0.7563	0.7502	0.7393	n=15	0.6	0.9764	0.9874	0.9891	0.9895	0.9896
n=20	0.6	0.8125	0.7757	0.7568	0.7481	0.735	n=20	0.6	0.9774	0.9877	0.9889	0.9862	0.986
n=25	0.5	0.8136	0.777	0.7576	0.7515	0.739	n=25	0.5	0.9798	0.9909	0.9918	0.9923	0.9934
n=30	0.5	0.8129	0.7726	0.7536	0.7438	0.7349	n=30	0.5	0.9739	0.9882	0.9902	0.986	0.9882
AVG		0.8121	0.7751	0.7544	0.747	0.7373	AVG		0.9685	0.9831	0.9851	0.9838	0.9865

TABLE 3					TABLE 4				
Improvement of TyCo for Other Methods with Sparse Training Data on MAE and Coverage					Comparison with State-of-the-Art Methods on MAE				
	method	$\chi = 0.1$	$\chi = 0.2$	$\chi = 0.3$	Training Set	Methods	Given5	Given10	Given 15
MAE	IBCF	9.89%	12.65%	11.69%	ML100	SCBPCC	0.874	0.845	0.839
	UBCF	11.77%	13.56%	11.55%		WLR	0.915	0.875	0.890
	EMDP	17.56%	16.17%	14.92%		CBT	0.840	0.802	0.786
	Random Guess	94.72%	97.93%	101.01%		SVD++	0.925	0.911	0.916
	TyCo	0.830	0.799	0.777		TyCo	0.830	0.799	0.777
Coverage	IBCF	59.78%	44.98%	17.11%	ML200	SCBPCC	0.871	0.833	0.828
	UBCF	58.78%	44.29%	16.75%		WLR	0.941	0.903	0.883
	EMDP	-2.31%	-1.444%	-1.24%		CBT	0.839	0.800	0.784
	Random Guess	-2.31%	-1.444%	-1.24%		SVD++	0.881	0.815	0.812
	TyCo	0.830	0.775	0.775		TyCo	0.830	0.775	0.775

FIGURE 4.11: Snapshot of “Table pattern from a research paper”

The patterns such as “First person plural pronoun” are widely repeated especially in the “Methodology” section. The occurrences of such patterns have been highlighted in the snapshot which has been taken from paper “Building a Search Engine for Algorithms”. The regular expression 4 is developed to count the frequency of such patterns as shown in Figure 4.12.

Regular Expression 4: $\backslash s(\text{we}|\text{our}|\text{for us})\backslash s([A-Za-z]*\backslash s)\{2\}$

This part of regular expression $\backslash s(\text{we}|\text{our}|\text{for us})\backslash s$ contains different types of patterns to represent the existence of first person pronoun in the text of citing document. These patterns are ‘we’ or ‘our’ or ‘for us’.

in a document, these captions usually serve the purpose of being anchors which can be referred to by context in the running text. Here we present three algorithms for detecting pseudocodes in scholarly documents: a rule based method (PC-RB), a machine learning based Machine Learning Based Method (PC-ML)

The PC-RB method gives a high precision rate, however it still suffers from a low coverage resulting in a poor recall. We found that 25.8% of pseudocodes in our test data set do not have accompanied captions. These pseudocodes would remain undetected using the PC-RB method. To get around this issue, we propose a machine learning based (PCML) method aiming to directly detect the presence of pseudocode contents (instead of their captions). Our motivation originated from the observation that most pseudocodes are written in a sparse manner, resulting in sparse regions, we call them sparse boxes, in documents. The PC-ML method first detects and extracts these sparse boxes, then classifies

FIGURE 4.12: Snapshot of “First person plural pronoun” patterns from a research paper

The last pattern is the “Algorithm”. A significant number of research papers in computer science and other domains consists of “Algorithm” patterns that provide short description for a solving a wide variety of computational tasks [86]. It can be represented by other words such as “Pseudo code”, “Flowchart” along with linked caption and algorithm number. This algorithm number is then used to identify the algorithm in the running text of the scientific document [86, 87]. They developed the algorithm search engine based on the “Algorithm” pattern. The algorithm is the procedure to identify the method of any problem in automatic way. Therefore, most of the time this pattern is used in the methodology section of the research works because the authors provide details about the implementation of new method in the “Methodology” section. Hence, it can also be used with phrases, “we devised algorithm”, “In proposed algorithm”, for the identification of “Methodology” section in academic research papers. In Figure 4.13, the snapshot of “Algorithm” patterns is shown with red circle. The regular expression 5 is used to count the occurrences of “Algorithm” pattern in the citing document.

Regular Expression 5: $([w|W]e|our|(L|l)et|the)[A-Za-z0-9\s]^*(a|A)lgorithm$

This regular expression is used to represent different occurrences of ‘Algorithm’ pattern in the text of citing document. These patterns are ‘we algorithm’, ‘We algorithm’, ‘our algorithm’, ‘Let see the algorithm’, ‘the algorithm’, and ‘algorithm or Algorithm’.

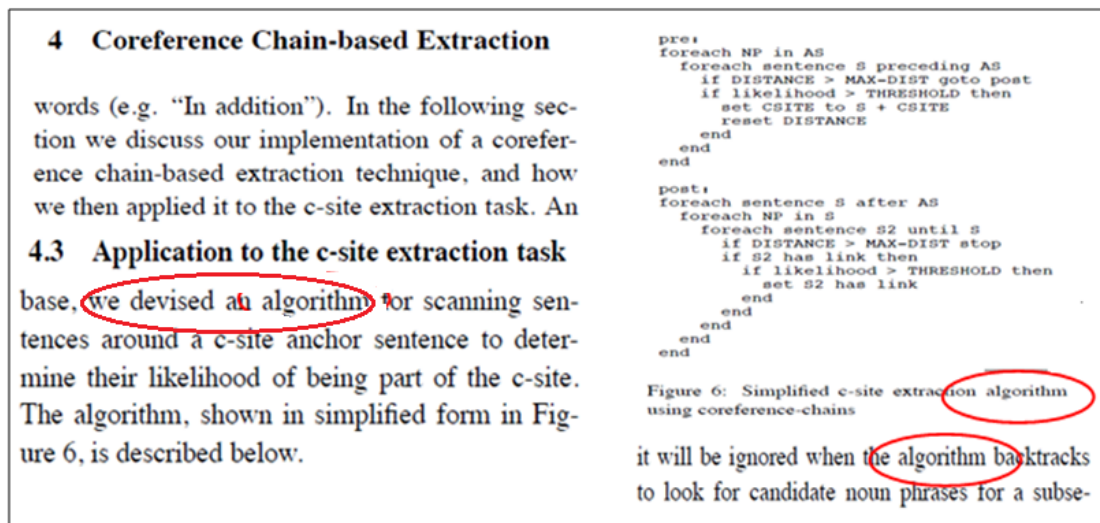


FIGURE 4.13: Snapshot of “Algorithm” pattern from a research paper

The patterns such as “First person plural pronoun”, and “Algorithm” are mostly exploited in the “Methodology” section of the scientific research papers.

The pseudocode of “Intext_Patterns_Mapping function is given below. This function has been used in Rule-Based_Algorithm as shown in sub-section 4.2.4. The section-heading-number, section-heading-label, section-body, total number of structural components are used as input in intext-patterns-mapping function. The ‘intext-patterns-mapping function will map the candidate structural component on generic section with the help of ‘RuleBasedDecision’ function by using different patterns, such as section-heading-number, total number of structural components, intextcitationfrequency, figurefrequency, tablefrequency, pronounfrequency, total_pages, and heading-page. Finally, the ‘intext-patterns-mapping’ function will return the generic id of mapped section as output.

```

1: function INTEXT_PATTERNS_MAPPING(SHNO, SHLAB, SBODY, TOTSEC)
2:   Shno → Section Heading number
3:   Shlab → Section Heading Label
4:   Sbody → Section body
5:   TotSec → Total Sections
6:   intextcitationfrequency := getIntextCitation(Sbody)
7:   figurefrequency := getFigures(Sbody)
8:   tablefrequency := getTables(Sbody)
9:   pronounfrequency := getPronoun(Sbody)
10:  total_pages := getPages()
11:  hp := sectionHeadingPage(Shlab)
12:  GSID := RuleBasedDecision(Shno, TotSec, intextcitationfrequency, figure-
    frequency, tablefrequency, pronounfrequency, total_pages, hp)
13:  return GSID
14: end function

```

In RuleBasedDecision function, the generic sections such as “Introduction, Literature, Methodology, Results, Discussion and Conclusion” are denoted by ‘1’, ‘2’, ‘3’, ‘4’, ‘5’, and ‘6’ respectively.

Sub-module 3: *Page and Structural Component Based Analysis*

The third sub-module is the last part of module (B) “Structural components mapping on the generic sections” as shown in Figure 4.1. All research papers contain different number of pages such as ‘2’, ‘3’, ‘4’, and ‘5’. The structural components, “Abstract”, “Introduction”, “Related Work”, “Methodology”, “Results”, “Discussion”, “Conclusion”, “Future Work”, and “Acknowledgement”, “References”

```

1: function RULEBASEDDECISION(SHNO, LASTSEC, ITCF, FIGF, TABF,
  PROF, TP, SHP)
2:   Shno → Section Heading number
3:   KastSec → Last Section
4:   ITCF → Intext Citation Frequency
5:   FigF → Figure Frquency
6:   TabF → Table Frquency
7:   ProF → Pronoun Frquency
8:   TP → Total pages
9:   SHP → Section Heading page
10:
11:   start := TP/3
12:   end := TP - start
13:   If Shno = 1 Then
14:     GSID := 1
15:   End If
16:   If Shno = LastSec Then
17:     GSID := 6
18:   End If
19:   If Shno = 2 and ITCF > ProF Then
20:     GSID := 2
21:   Else If Shno = 2 and ProF > ITCF Then
22:     GSID := 3
23:   End If
24:   If (SHP >= end and ITCF > 5) and (FigF = 0 and TabF = 0) Then
25:     GSID := 2
26:   End If
27:   If (Shno > 2 and hp < end) and ProF > 0) Then
28:     GSID := 3
29:     flag := 1
30:   End If
31:   If (Shno > 2 and hp < end) and ProF > 0 and (FigF >0 and TabF >
0)Then
32:     GSID := 4
33:     flag := 1
34:   End If
35:   If (Shno > 2 and hp < end) and flag = 0) Then
36:     GSID := 3
37:   End If
38:   If (hp >= end and hno < LastSec) Then
39:     GSID := 5
40:   End If
41:   return GSID
42: end function

```

are distributed over pages in research papers. The number of structural components varies in research papers based on the number of pages. While in most of cases, the sequence of these structural components does not change in research papers. However, in our research work, according to ILMRaD format research papers are organized by four basic generic sections: “INTRODUCTION”, “LITERATURE”, “METHODOLOGY”, “RESULTS and DISCUSSION”. The ILMRaD structure does not consider the three structural components “Abstract”, “Acknowledgment”, and “References”. But most of times, these four generic sections are represented by different structural components in such sequence, “Introduction”, “Related Work”, “Methodology”, “Results”, “Discussion”, “Conclusion”, and “Future Work” [79]. Therefore, the above sequence of structural components in research papers can be represented by the sequence of generic sections. For example, in Figure 4.14 the sequence of structural components of a research paper (from Figure 4.7) is represented by the sequence of generic sections. The sequence of structural components is represented by red dotted round rectangle and the sequence of generic sections is represented by green solid round rectangle. The sequence pattern of generic sections for structural components in a research paper is “I L M M M R C”.

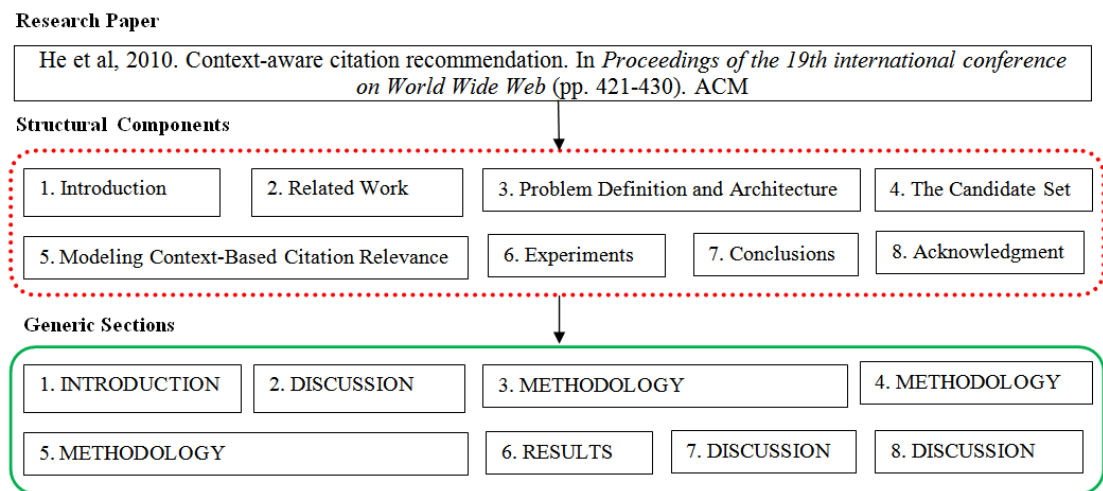


FIGURE 4.14: Structural components of a research paper mapped on generic Sections

The sub-module (3) “page and structural component based analysis” is used for the mapping of structural components of research papers on generic sections. The

mapping process of sections is performed by using predefined dataset of sequence patterns of generic sections. The sequence patterns of generic sections in research papers was identified by analyzing the sequence of structural components. Initially the sequence patterns of generic sections are prepared based on structural components in training dataset and then these patterns are stored in the generic sections dataset for future use. Therefore, the generic sections dataset has been included in the second phase of proposed architecture as shown in Figure 4.1. In this sub-module, the research papers corpus are classified into different groups based on the number of pages. Then, the research papers with same number of pages in each group are further classified into sub-groups based on the number of structural components. For the initial experiment of “Pages and structural components based analysis”, some of the papers having 4, 6 and 8 pages are shown in Table 4.8. This dataset contains information such as ‘PaperId (P#)’, ‘Paper Title’, ‘Total pages (TP)’, and ‘Structural components (SC)’. The research papers were selected from the corpus of 211 research papers.

The dataset in table 4.8 is classified based on pages and structural components as shown in Figure 4.15. In first step of classification, the research papers dataset at root node is classified into three branches based on number of pages. Each of three branches shows the subset of original dataset that consists of research papers with the same number of pages, the middle branch in Figure 4.15 contains 12 research papers of ‘4’ pages. In second step of classification, each subset of the second level is further classified into third level subsets based on structural components. Each subset of third level consists of research papers with the same pages and structural components, the middle subset at second level in Figure 4.15 that contains four subsets in third level. One of the four subsets contains research papers with the same number of pages and structural components. The third level subsets of research papers in Figure 4.15 are further analyzed for the sequence patterns of generic sections based on structural components sequence as shown in Figure 4.15. The sequence patterns of generic sections are stored in the generic sections dataset. This dataset is used for the identification of generic sections in testing dataset.

TABLE 4.8: Training dataset for pages and structural components based analysis

P#	Paper Title	TP	SC
1	A Case Study on How to Manage the Theft of Information	4	4
2	A Similarity Measure for Motion Stream Segmentation and Recognition	6	5
3	A Flexible 3D Slicer for Voxelization Using Graphics Hardware	3	5
4	A Survey of Collaborative Information Seeking Practices of Academic Researchers	4	6
5	Towards Content-Based Relevance Ranking for Video Search	4	5
6	An Architectural Style for High-Performance Asymmetrical Parallel Computations	4	4
7	A WEIGHTED RANKING ALGORITHM FOR FACET-BASED COMPONENT RETRIEVAL SYSTEM	6	7
8	An empirical comparison of supervised machine learning techniques in bioinformatics	4	6
9	Measuring Cohesion of Packages in Ada95	6	7
10	An Integrated Environment to Visually Construct 3D Animations	4	4
11	Building a Research Library for the History of the Web	8	6
12	Catenaccio: Interactive Information Retrieval System through Drawing	4	7
13	A Geometric Constraint Library for 3D Graphical Applications	8	8
14	A Coupling and Cohesion Measures for Evaluation of Component Reusability	4	7
15	Unwanted Traffic in 3G Networks	4	4
16	Easy Language Extension with Meta-AspectJ	4	5
17	Distance Measures for MPEG-7-based Retrieval	8	6
18	Real-world Oriented Information Sharing Using Social Networks	4	4

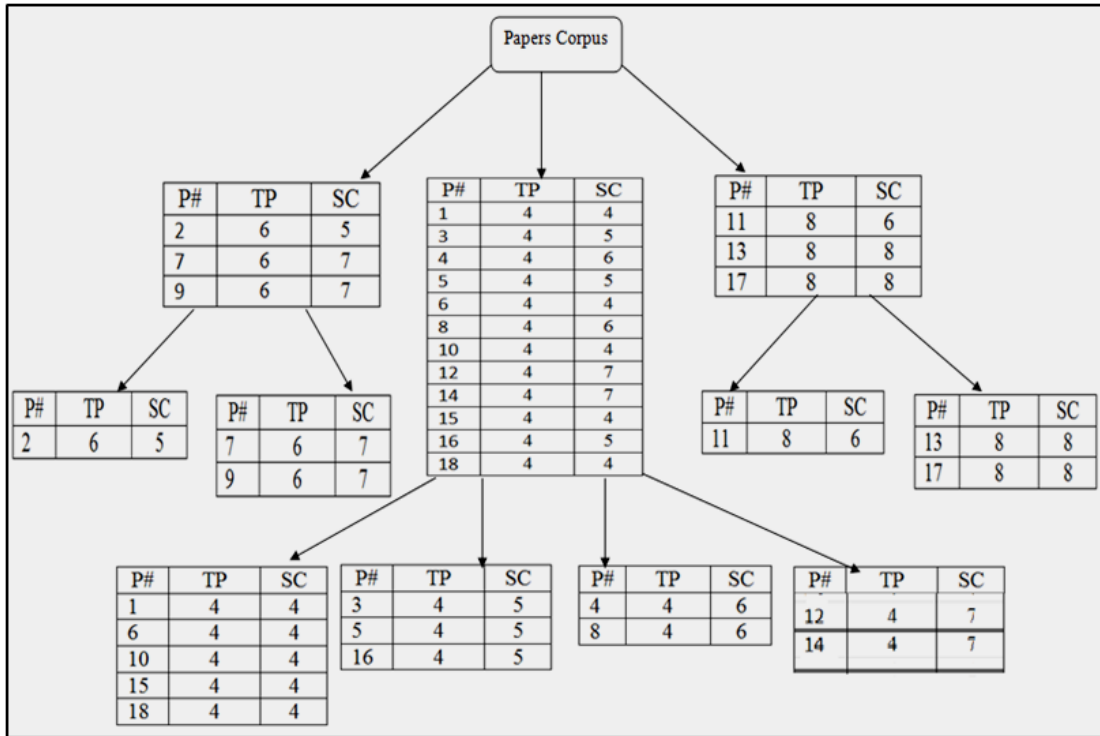


FIGURE 4.15: Training dataset classification based on pages and structural components

After the classification of training dataset, twelve research papers of 4 pages are selected from the training dataset for the analysis of sequence patterns in generic sections. The set of twelve research papers are varying based on structural components. The first subset contains five research papers of 4 structural components such as (P# = 1, 6, 10, 15, and 18). The second subset contains three research papers of 5 structural components such as (P# = 3, 5, and 16). The third subset consists of two research papers with 6 structural components such as (P# = 4, 8). The last subset also consists of two research papers with 7 structural components like (P# = 12, 14) Now, each research paper in four subsets is further analyzed for the sequence patterns of generic sections based on the structural components sequence. For example, the second research paper (P#=6) “An Architectural Style for High-Performance Asymmetrical Parallel Computations” in first subset in Table 4.9 contains four structural components such as “Introduction”, “Motivation”, “A Novel Protocol”, and “Discussions”. Based on the sequence pattern of structural components in the research paper (P#6), the sequence of generic sections is manually identified such as “Introduction”, “Literature”, “Methodology”, and

“Conclusion” (I, L, M, C) as shown in Table 4.9. The same process is repeated for the rest of the research papers in each subset. The sequence patterns of generic sections in five research papers such as (P#=1 {I, L, R, C}, P#=6 {I, L, M, D}, P#=10 {I, M, L, C}, P#=15 {I, M, M, C}, P#=18 {I, M, R, C}) obtained as shown in Table 4.9. The occurrences column shows the frequency of particular sequence pattern in sequence dataset. The 'N' is the total number of patterns in sequence dataset which can be calculated by the values of occurrences column in Table 4.9.

TABLE 4.9: Sequence patterns of Generic Sections in first subset of 4 pages Research Papers

Papers group	P#	Structural Components	Sequence Patterns of Generic Sections	Occurrences
4 Pages	1	4	I, L, R, C	1
	6		I, L, M, D	1
	10		I, M, L, C	1
	15		I, M, M, C	1
	18		I, M, R, C	3

The sequence patterns of generic sections in above subset of research papers is used to create the position frequency matrix (PFM) as has been used by Roderic and Pape [84, 88]. Here, it is created by counting the occurrences of each generic section at each position in five sequence patterns of generic sections. In this matrix, columns are represented by the number of structural components and rows are represented by the number of generic sections such as ‘I’, ‘L’, ‘M’, and ‘C’. The structure of PFM is given in Table 4.10 with frequencies of generic sections in the subset of sequence patterns.

TABLE 4.10: Position frequency matrix (M1)

	1	2	3	4
I	7	0	0	0
L	0	2	1	0
M	0	5	2	0
R	0	0	4	0
D	0	0	0	1
C	0	0	0	6

The frequency of each generic section at each position in a given set of five sequence patterns is calculated by Equation 4.1. The symbol ‘X’ represents the set of generic sequence patterns shown in the fourth column of Table 4.9. The symbol ‘N’ denotes the total number of sequence patterns in ‘X’ and can be calculated by using the “occurrences” column in Table 4.9. For example in 4 pages case, ‘N’ is ‘5’. The ‘gs’ is the set of generic sections (I, L, M, R, D, C) whereas ‘sp’ stands for sequence patterns while ‘p’ represents the position of the each generic section in the sequence patterns in set ‘X’. The value of ‘sp’ will be considered within the range of ‘1 to N’. The value of ‘p’ will be considered within the range of ‘1 to l’. The symbol ‘l’ shows the length of sequence patterns which will be constant for all sequence patterns of generic sections in each subset. The result of Equation 4.1 will be stored in ‘M₁’ Position Frequency Matrix. ‘I’ is the indicator function which will return either 1 or 0 value.

$$M_{1(gs,p)} = \sum_{sp=1}^N I(X_{(sp,p)} = gs) \quad \begin{cases} 1, I(a = gs) \\ 0, otherwise \end{cases} \quad (4.1)$$

The values of ‘M₁’ matrix do not exist in normalized form. Hence, Equation 4.2 is used for normalization of ‘M₁’ matrix. By this Equation, each non-zero value of ‘M₁’ matrix is divided by the total number of sequence patterns in set ‘X’.

$$M_{2(r,c)} = \frac{M_{1(r,c)}}{N} \quad IF \quad M_{1(r,c)} \neq 0 \quad (4.2)$$

The result of Equation 4.2 is stored in Matrix ‘M₂’ which is shown in table 4.11. The ‘M₂’ matrix is called position probability matrix (PPM).

TABLE 4.11: Position probability matrix (M₂)

	1	2	3	4
I	1	0	0	0
L	0	0.3	0.1	0
M	0	0.7	0.3	0
R	0	0	0.6	0
D	0	0	0	0.1
C	0	0	0	0.9

Finally, we will find the probability of each sequence patterns in set ‘X’ based on position probability matrix (PPM). The probability of each sequence pattern will be stored in the sequence probability matrix (SPM). The probability of each element in SPM will be calculated by the Equation 4.3.

$$M_{3(s,1)} = \prod_{j=1}^L M_2(X_{(s,j)}, j) \quad M_2(X_{(s,j)}, j) \neq 0 \quad (4.3)$$

In Equation 4.3, the symbol ‘S’ denotes the sequence pattern of generic sections in set ‘X’. The symbol ‘L’ is the length of sequence pattern. It varies based on the number of structural components in the subset of research papers. The ‘M₂’ is the position probability matrix in Equation 4.2 that will be exploited for the calculation of each sequence pattern probability in set ‘X’. Let us take the sequence ‘S’ = I, M, R, C from the set ‘X’. The Probability of ‘S’ can be calculated by multiplying the relevant probabilities of each generic section at each position in matrix ‘M₂’.

Sequence (S) = I, M, R, C

Position Probability (PP) = 1, 0.7, 0.6, 0.9

(Sequence Probability) P (S|M₂) = 1 × 0.7 × 0.6 × 0.9 = 0.378

In the same way, the sequence probabilities of all sequence patterns of subsets in Table 4.9 can be calculated by Equation 4.3. The sequence probabilities of unique sequence patterns of first subset in Table 4.9 have been shown in Table 4.12. The Table 4.12 shows that the sequence pattern such as “I,M,R,C” has the highest probability (0.378) in the subset of five research papers with four pages and four structural components. Based on this highest probability sequence, the new research paper with 4 pages and 4 structural components can be mapped on generic sections.

TABLE 4.12: Sequence patterns with probabilities

Sequence Patterns	Position Probabilities	M3 with Sequence Probabilities
I, L, R, C	1, 0.3, 0.6, 0.9	$1 \times 0.3 \times 0.6 \times 0.9 = 0.162$
I, L, M, D	1, 0.3, 0.3, 0.1	$1 \times 0.3 \times 0.3 \times 0.1 = 0.009$
I, M, L, C	1, 0.7, 0.1, 0.9	$1 \times 0.7 \times 0.1 \times 0.9 = 0.063$
I, M, M, C	1, 0.7, 0.3, 0.9	$1 \times 0.7 \times 0.3 \times 0.9 = 0.189$
I, M, R, C	1, 0.7, 0.6, 0.9	$1 \times 0.7 \times 0.6 \times 0.9 = \mathbf{0.378}$

In Figure 4.16, the sequences probabilities have been shown for the sequences in seven research papers which consist of 4 pages and different structural components such as 4, 5.

P#	SC	Sequence Patterns of generic sections	Occurrence	PFM (M ₁)	PPM(M ₂)	SPM (M ₃)																																																																																				
1 6 10 15 18	4	(S1) I, L, R, C (S2) I, L, M, D (S3) I, M, L, C (S4) I, M, M, C (S5) I, M, R, C	1 1 1 1 3	<table border="1"> <tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th></tr> <tr><th>I</th><td>7</td><td>0</td><td>0</td><td>0</td></tr> <tr><th>L</th><td>0</td><td>2</td><td>1</td><td>0</td></tr> <tr><th>M</th><td>0</td><td>5</td><td>2</td><td>0</td></tr> <tr><th>R</th><td>0</td><td>0</td><td>4</td><td>0</td></tr> <tr><th>D</th><td>0</td><td>0</td><td>0</td><td>1</td></tr> <tr><th>C</th><td>0</td><td>0</td><td>0</td><td>6</td></tr> </table>		1	2	3	4	I	7	0	0	0	L	0	2	1	0	M	0	5	2	0	R	0	0	4	0	D	0	0	0	1	C	0	0	0	6	<table border="1"> <tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th></tr> <tr><th>I</th><td>1</td><td>0</td><td>0</td><td>0</td></tr> <tr><th>L</th><td>0</td><td>0.3</td><td>0.1</td><td>0</td></tr> <tr><th>M</th><td>0</td><td>0.7</td><td>0.3</td><td>0</td></tr> <tr><th>R</th><td>0</td><td>0</td><td>0.6</td><td>0</td></tr> <tr><th>D</th><td>0</td><td>0</td><td>0</td><td>0.1</td></tr> <tr><th>C</th><td>0</td><td>0</td><td>0</td><td>0.9</td></tr> </table>		1	2	3	4	I	1	0	0	0	L	0	0.3	0.1	0	M	0	0.7	0.3	0	R	0	0	0.6	0	D	0	0	0	0.1	C	0	0	0	0.9	(S1) I, L, R, C = 0.162 (S2) I, L, M, D = 0.009 (S3) I, M, L, C = 0.063 (S4) I, M, M, C = 0.189 (S5) I, M, R, C = 0.378														
	1	2	3	4																																																																																						
I	7	0	0	0																																																																																						
L	0	2	1	0																																																																																						
M	0	5	2	0																																																																																						
R	0	0	4	0																																																																																						
D	0	0	0	1																																																																																						
C	0	0	0	6																																																																																						
	1	2	3	4																																																																																						
I	1	0	0	0																																																																																						
L	0	0.3	0.1	0																																																																																						
M	0	0.7	0.3	0																																																																																						
R	0	0	0.6	0																																																																																						
D	0	0	0	0.1																																																																																						
C	0	0	0	0.9																																																																																						
3 5 16	5	(S1) I, L, M, R, C (S2) I, M, M, R, C (S3) I, M, M, M, C	7 1 1	<table border="1"> <tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th></tr> <tr><th>I</th><td>9</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><th>L</th><td>0</td><td>7</td><td>1</td><td>0</td><td>0</td></tr> <tr><th>M</th><td>0</td><td>2</td><td>9</td><td>1</td><td>0</td></tr> <tr><th>R</th><td>0</td><td>0</td><td>0</td><td>8</td><td>0</td></tr> <tr><th>D</th><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><th>C</th><td>0</td><td>0</td><td>0</td><td>0</td><td>9</td></tr> </table>		1	2	3	4	5	I	9	0	0	0	0	L	0	7	1	0	0	M	0	2	9	1	0	R	0	0	0	8	0	D	0	0	0	0	0	C	0	0	0	0	9	<table border="1"> <tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th></tr> <tr><th>I</th><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><th>L</th><td>0</td><td>0.8</td><td>1</td><td>0</td><td>0</td></tr> <tr><th>M</th><td>0</td><td>0.2</td><td>1.0</td><td>0.1</td><td>0</td></tr> <tr><th>R</th><td>0</td><td>0</td><td>0</td><td>0.8</td><td>0</td></tr> <tr><th>D</th><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><th>C</th><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td></tr> </table>		1	2	3	4	5	I	1	0	0	0	0	L	0	0.8	1	0	0	M	0	0.2	1.0	0.1	0	R	0	0	0	0.8	0	D	0	0	0	0	0	C	0	0	0	0	1	(S1) I, L, M, R, C = 0.691 (S2) I, L, M, M, C = 0.024 (S3) I, L, M, M, C = 0.197
	1	2	3	4	5																																																																																					
I	9	0	0	0	0																																																																																					
L	0	7	1	0	0																																																																																					
M	0	2	9	1	0																																																																																					
R	0	0	0	8	0																																																																																					
D	0	0	0	0	0																																																																																					
C	0	0	0	0	9																																																																																					
	1	2	3	4	5																																																																																					
I	1	0	0	0	0																																																																																					
L	0	0.8	1	0	0																																																																																					
M	0	0.2	1.0	0.1	0																																																																																					
R	0	0	0	0.8	0																																																																																					
D	0	0	0	0	0																																																																																					
C	0	0	0	0	1																																																																																					

FIGURE 4.16: Page and structural component based analysis for research papers with four pages

In PSCA_Analysis_Mapping algorithm, the citing document is used as input while it returns the generic section id as output. This algorithm consists of three important functions, such as create_PFM, create_PPM, and create_SPM. These functions are used to finally find the most frequent sequence of generic sections in

the sequences dataset for section mapping. The functionality of these functions is discussed as above.

```

1: function PSCA_ANALYSIS_MAPPING(CITINGDOCUMENT)
2:   pages := getPages(Citingdocument)
3:   sections := getSections(Citingdocument)
4:   PFM [ ][ ] := create_PFM (sections) // PFM—Position Frequency Matrix
5:   sequences [ ] = get_Patterns (pages, sections)
6:   PFM := populatePFM (sequences)
7:   PPM [ ][ ] := create_PPM (sections) // PPM—Position Probability Matrix
8:   SPM [ ] := create_SPM (sequences, PPM) // SPM—Sequence Probability
   Matrix
9:   getFrequentSequence [ ] := get_Frequent_Pattern(SPM)
10:  GSID [ ] := convertGSID (getFrequencySequence)
11:  return GSID
12: end function

```

4.2.4 Rule Based Algorithm (RBA) for generic section identification

In third sub-module, the “Rule-based algorithm” is developed based on the three proposed methods for structural components mapping on generic sections in research papers. These three methods are “Section heading labels based Analysis”, “Intext pattern based analysis” and “Pages and structural components based analysis”. After mapping process, each method generates individual sequence pattern of generic sections for the structural components of the candidate research paper. Now the problem is to select the best sequence pattern of generic sections out of three patterns. To solve this problem, we present the “Rule based algorithm”, as shown in the architecture of Figure 4.1. This algorithm analyzes the results of three methods based on some predefined rules for the selection of best sequence pattern of generic sections in research papers. Each method will generate different types of result for same structural components as shown in Figure 4.17.

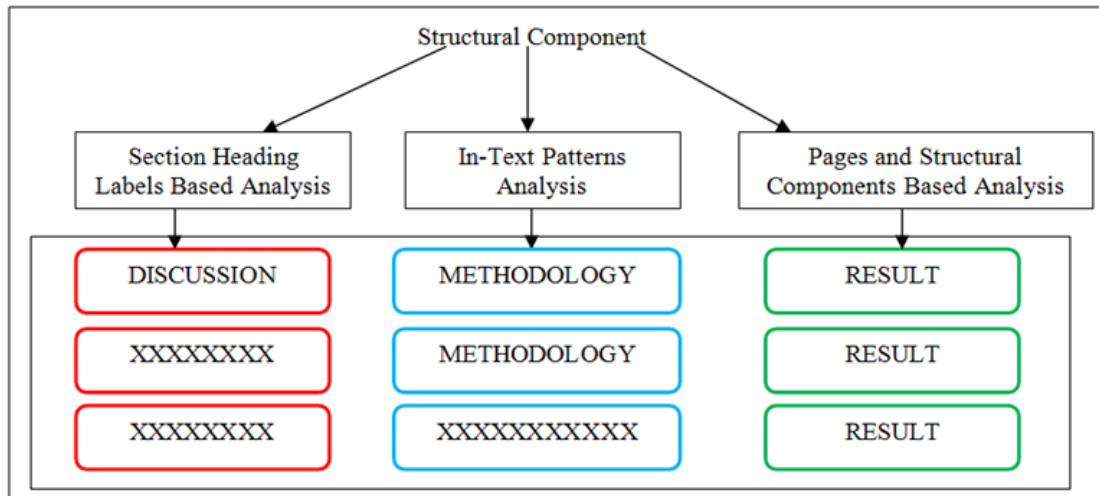


FIGURE 4.17: Proposed methods for section mapping

Our proposed algorithm will first prefer the result of “Section heading labels based Analysis” method. If the “Section heading labels based analysis” does not yield any result, then the proposed algorithm will decide the result of “Intext pattern Analysis” method. If the second method also fails to provide any result, then finally the result of “Pages and Structural components based Analysis” will be considered for final decision of mapping for particular structural component. This is unlikely that any of the three modules does not provide a result; the module “Pages and Structural components based Analysis” is guaranteed to provide an answer. The pseudo-code for “Rule Based Algorithm” as given below. The function ‘Rule-Based-Algorithm’ gets the citing document as input and gives the generic section id as output for final section mapping.

```

1: procedure RULE_BASED_ALGORITHM(CD)
2:   CD → Citing-document
3:   Stem_Word_Set := get_Stemwords_Dataset()
4:   SC := get_Structural_Components(CD) //SC → Structural component
5:   PSCB-GSID [ ] := PSCB_Analysis_Mapping(CD)
6:   For i := 1 To SC.length
7:     KM-GSID := Keyword_Mapping(SC(i). heading_title, Stem_Word_Set)
8:     IPM-GSID := Intext_Pattern_Mapping(SC(i).hno, SC(i).label,
9:     SC(i).textbody, SC.length)
10:    PSCBGSID := PSCBMapping (PSCB-GSID [i])
11:   If KM-GSID != 0 Then
12:     SC_Final_result := KM-GSID
13:   End If
14:
15:   If KM-GSID == 0 && IPM-GSID !=0 Then
16:     SC_Final_result := IPM-GSID
17:   End If
18:
19:   If KM-GSID == 0 && IPM-GSID == 0 Then
20:     SC_Final_result := PSCBGSID
21:   End If
22:
23:   Mapped_Structural_Components(SC, SC_Final_result)
24:   End loop
25: end procedure

```

After the decision of “Rule based algorithm”, the final pattern of generic sections will be integrated with structural components of research papers by using generic section integrator. The generic section integrator will finally store the result in the structured dataset as shown in table 4.7.

4.2.5 Generic section evaluation phase

First, in the training step, we have selected the training dataset of 211 research papers for the preparation of our proposed approach. This dataset contained 1,220 section heading labels. Now in the testing and evaluation step, two annotated section labels datasets are selected for the evaluation of our proposed approach. The first dataset consisted of 279 citing documents. From this dataset, 150 unique citing documents are selected for our experiments with 850 sections. The second dataset consisted of 500 research papers. After analyzing the documents of 500 research papers, only 300 documents are selected for our experiments. These 300 documents consisted of 1600 sections heading labels. The statistics of training and testing datasets are given in Table 4.13

TABLE 4.13: Training and testing datasets for generic section identification task

Datasets	Citing documents	Number of Section heading labels
Training set	211	1220
Testing dataset1	150	850
Testing dataset2	300	1600

Our technique is compared with the state-of-the-art technique [28] over both sets of testing data. First both approaches are implemented over testing dataset1 and then the results of proposed approach are compared with state-of-the-art on 50 randomly selected research papers out of 150 papers with 304 sections. For the proper analysis of generic section identification, the individual confusion matrix for each technique is prepared over both test datasets. The confusion matrix of proposed approach is constructed over testing dataset1 as given in Table 4.15 and in the same way the confusion matrix of state-of-the-art technique is constructed over testing dataset1 as shown in Table 4.14

TABLE 4.14: Confusion matrix of proposed approach for 50 papers in testing dataset1

Predicted as	INTR	LITR	MET	RES	DISC	CON
Introduction	49	1	3	0	0	0
Literature	0	38	4	0	0	0
Methodology	1	4	105	8	0	0
Results	0	1	5	21	1	0
Discussions	1	0	0	1	22	0
Conclusions	0	0	0	0	1	38

TABLE 4.15: Confusion matrix of State-of-the-art approach for 50 papers in testing set1

Predicted as	INTR	LITR	MET	RES	DISC	CON
Introduction	43	0	2	0	0	0
Literature	0	39	3	0	0	0
Methodology	0	4	104	15	0	0
Results	0	0	6	20	13	0
Discussions	0	0	2	0	14	3
Conclusions	0	0	3	0	0	33

Both approaches are evaluated over the confusion matrix using Precision, Recall, and F-score. The Precision, Recall, and F-score can be measured by using Equations 4.4, 4.5, and 4.6 respectively.

$$Precision = \frac{TruePositive(TP)}{TruePositive(TP) + FalsePositive(FP)} \quad (4.4)$$

$$Recall = \frac{TruePositive(TP)}{TruePositive(TP) + FalseNegative(FN)} \quad (4.5)$$

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.6}$$

Let us demonstrate the procedure of finding Precision, Recall, and F-score of “Introduction” section by using the values of confusion matrix as shown in Table 4.15. The ‘TP’ (True positive) values of introduction is ‘49’ while the ‘FP’ (False positive) values can be calculated by adding the values on Y-axis under the “INTR” column i.e. ‘0, 1, 0, 1, 0’. The ‘FN’ (False negative) values can be calculated by adding the values in front of ‘Introduction’ section on X-axis i.e. ‘1, 3, 0, 0, 0’. The recall value of Introduction is calculated by Equation 4.5. Similarly, the precision of Introduction can be calculated by using Equation 4.4. Finally the F-score is calculated by using Equation 4.6.

$$Recall = \frac{49}{49+1+3+0+0+0} = 0.924$$

$$Precision = \frac{49}{49+0+1+0+1+0} = 0.96$$

$$F-score = 2 \times \frac{0.96 \times 0.924}{0.96+0.924} = 0.94$$

Similarly, the precision, recall, and F-score of other sections can be determined using confusion matrix values of proposed approach over testing dataset1 as shown in Table 4.16

TABLE 4.16: statistical data of proposed approach over testing dataset1

Sections	Total	Correct	Incorrect	Precision	Recall	F-Score
INT	53	49	4	0.960784	0.924528302	0.942307692
LITR	42	38	4	0.863636	0.904761905	0.88372093
MET	118	105	13	0.897436	0.889830508	0.893617021
RES	28	21	7	0.7	0.75	0.724137931
DISC	24	22	2	0.916667	0.916666667	0.916666667
CON	39	38	1	1	0.974358974	0.987012987
Aggregate Score	304	273	31	0.88975	0.893357726	0.891552158

The statistical data analysis of state-of-the-art technique is shown in Table 4.17. This data also shows the precision, recall and F-score over 304 sections. The statistical result shows that the proposed approach performed better than the

state-of-the-art technique. The F-score of proposed approach is 0.89 while the F-score of state-of-the-art technique is 0.80.

TABLE 4.17: Statistical data of state-of-the-art technique over testing dataset1

Sections	Total	Correct	Incorrect	Precision	Recall	F-Score
INT	45	38	7	1	0.955556	0.977273
LITR	42	32	10	0.906977	0.928571	0.917647
MET	123	104	19	0.866667	0.845528	0.855967
RES	39	16	23	0.571429	0.512821	0.540541
DISC	19	11	8	0.518519	0.736842	0.608696
CON	36	33	3	0.916667	0.916667	0.916667
Aggregate Score	304	234	70	0.79671	0.816	0.80624

For the comprehensive analysis, both approaches were again evaluated over the testing dataset2 which consists of 300 papers with 1600 sections. The statistical data of proposed approach and state-of-the-art approach is given in Table 4.18 and in Table 4.19 respectively over the testing dataset2. This statistical data is prepared for a sample of 50 research papers out of 300 papers in testing dataset2. The F-score of proposed approach is 0.95 and the F-score of state-of-the-art technique is 0.82. The second analysis also shows that the proposed approach is better than the state-of-the-art technique.

TABLE 4.18: Statistical data of proposed technique over testing dataset2

Sections	Total	Correct	Incorrect	Precision	Recall	F-Score
INT	50	50	0	0.980392	1	0.990099
LITR	44	39	5	0.975	0.886363	0.928571
MET	80	71	9	0.934210	0.87654321	0.90445859
RES	47	46	1	0.836363	0.978723	0.90196784
DISC	6	5	1	1	1	1
CON	46	46	0	1	1	1
Aggregate Score	273	257	16	0.95432772	0.956938375	0.955631265

TABLE 4.19: Statistical data of state-of-the-art technique over testing dataset2

Sections	Total	Correct	Incorrect	Precision	Recall	F-Score
INT	50	50	0	1	1	1
LITR	36	32	4	0.820513	0.888889	0.853333
MET	87	76	11	0.915663	0.873563	0.894118
RES	47	30	17	0.857143	0.638298	0.731707
DISC	7	5	2	0.25	0.714286	0.37037
CON	46	46	0	1	1	1
Aggregate Score	273	239	34	0.80722	0.852506	0.829245

In Figure 4.18, the comparison of both proposed and state-of-the-art approaches has been shown over both testing datasets. The graph shows that the precision, recall, and F-score of proposed approach is higher than the state-of-the-art [28].

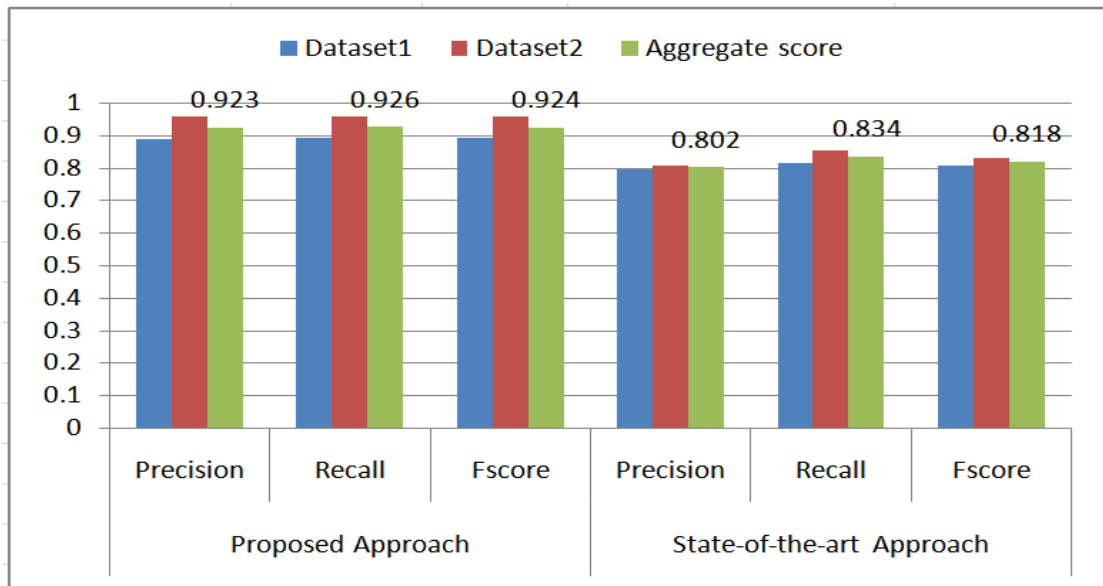


FIGURE 4.18: Aggregated precision, recall, and F-score of generic section identification for both approaches

4.3 Summary

In this chapter, we proposed, implemented and evaluated a novel approach for section mapping. Furthermore, in the evaluation process, two annotated testing

datasets were selected with 150 and 300 citing documents respectively. The proposed technique was evaluated based on well-known measures of precision, recall and F-score. The precision and recall values were computed for each standard section, “Introduction”, “Related Work”, “Methodology”, “Results”, “Discussion” and “Conclusion”. For the comparison of proposed approach, the state-of-the-art [28] technique was also applied on the same dataset. The aggregated F-score of proposed approach was 0.92 over the both datasets while the F-score of state-of-the-art technique was 0.81.

The latter approach only considered the keyterms in section labels and the position of sections in the research papers. They have only used the direct matching of section labels with predefined set of section labels using simple rules. In our approach, the patterns, section number, number of citations, number of figures, number of tables, first person plural pronoun, number of pages, and number of structural components were used for the accurate identification of section mapping instead of keyterms. We have used three methods for sections mapping (1) Section Headings labels based analysis (2) In-text patterns based analysis (3) Pages and structural components based analysis. Finally, the rules and heuristic based algorithm “Rule based algorithm” take decision for final section mapping using three methods of section mapping.

For the section wise co-citation analysis, three modules have been developed. This chapter covered the functioning of the first module and this becomes one of the key contributions of this thesis. The next chapter describes the module of “in-text citation patterns and frequencies identification” while chapter 6 discusses “section wise co-citation analysis (SWCA)” in detail.

Chapter 5

In-Text Citation Patterns Identification

Note: The parts of this chapter have been published in Journals^{1 2}

In chapter 3, the detailed architecture of our proposed approach has been introduced. This chapter has been written over the second problem in the proposed architecture. If “in-text co-citation patterns and frequencies identification” problem has been solved, it will not only help us to develop Section Wise Co-citation Analysis (SWCA) system but it will also be helpful in improving state-of-the-art in other domains and application scenarios, ranking of authors, journals, institutions, and organizations. Sometimes, documents cite a reference many times in their full-text which is further used in many application scenarios, such as (1) finding relationship between cited and citing papers [34] (2) identifying influential cited paper from a set of references in a citing paper [6] (3) identification of suitable citation functions [26], and (4) study of in-text citations in different logical sections of papers to conclude different findings [18].

¹Ahmad, R., Afzal, M. T., & Qadir, M. A. (2017). Pattern Analysis of Citation-anchors in Citing documents for Accurate Identification of In-text Citations. *IEEE Access*, 5:5819- 5828. [Impact Factor: 3.244]

²Ahmad, R. & Afzal, M.T. (2018), CAD: an algorithm for citation-anchors detection in research papers. *Scientometrics*. Published online 29th September 2018. <https://doi.org/10.1007/s11192-018-2920-6>

This chapter proceeds as follows. Section 5.2 highlights the real issues of identification task of in-text citation-anchors. In section 5.4 the proposed taxonomy of citation-anchor is discussed. The methodology adopted for the experiments is explained in section 5.5. The dataset, evaluation metrics and results are presented and discussed in section 5.6.

5.1 Overview of Basic Terminology

In Figure 5.1, the difference between reference string, citation-tags, in-text citation and citation-anchors are highlighted. The reference string is the set of alphabetical, numerical and special characters symbols which are included in the reference section of a citing document to represent the link to the cited document. This type of link is called citation of a cited document. Each reference string is identified by unique key in a reference section which is called citation-tag as shown in small red circle in Figure 5.1. When a cited document is cited in the text of a citing document it is called in-text citation. The in-text citation is represented by the identifier which is called citation-anchor as shown in large green circle in Figure 5.1. The citation anchors may be used more than one time in text of citing document. The in-text citation frequency identification can be affected by the format and style variations between citation-tag and citation-anchor of the same reference string. In experimental analysis, we have found different cases of real scenarios which are not solved by the direct matching [18] of citation-tag with citation-anchor. These citation-anchors are used with different style and formatting as discussed in following section with issues.

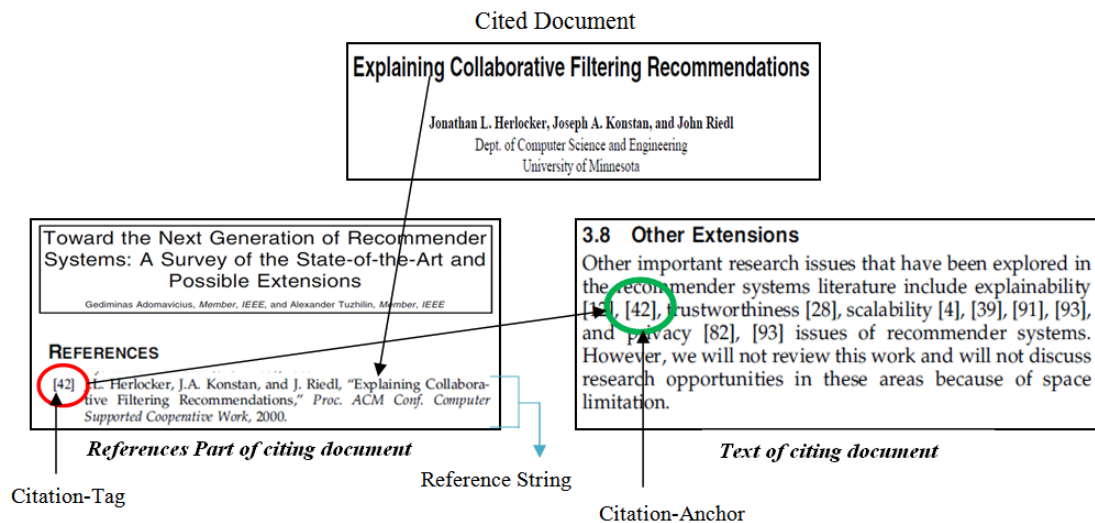


FIGURE 5.1: Reference string, citation-tag and citation-anchor relationship

5.2 Pattern Analysis and Issues of Citation-Anchor

After the critical analysis of citation-anchors in text of citing documents, it is concluded that there are two types of citation-anchors that are used in citing documents (1) Numeric citation-anchors and (2) String citation-anchors. The numeric citation-anchors are detected by the numeric citation-tags while the string citation-anchors are extracted by using string citation-tags. In this section, we have highlighted the key issues with both numeric and string citation-anchors during matching with numeric citation-tags and strings citation-tags respectively.

5.2.1 Numeric citation-tags problems

In the numeric citation-tags problems, the frequency of citation reduces due to the different style of numeric, such as citation-anchor, multiple-anchor, range-anchor, and compound-anchor.

Multiple-anchor Problem

The real snapshot of numeric citation-tag mapping on multiple citation-anchor is shown in Figure 5.2. In this scenario, a numeric citation-tag does not exactly match with multiple citation-anchors due to the inclusion of more than one citation, such

as “28, 26, and 38”. If we try to find an exact match between [25, 28, 26, 38] in the text of the citing document with [25] in the references, the search will fail, hence the in-text citation count for this paper will be incorrect.

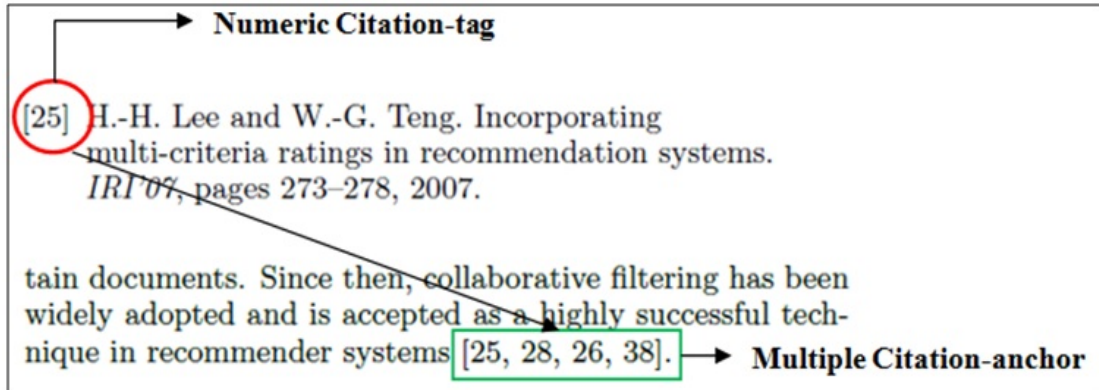


FIGURE 5.2: Mapping of numeric citation-tag on multiple citation-anchors

Range-anchor problem

In pattern analysis of citation-anchors, it is observed that significant numbers of citations are represented in text of citing documents by range citation-anchors. The range citations are denoted by the sign, such as “-” or “[]-[]”. In Figure 5.3, the real snapshot shows that numeric citation-tag does not properly match with the range citation-anchors, such as “[2]-[4], (6-8), [4-6]”. If we try to find all the in-text citations for paper [3] using exact matching, we will miss the citation which has been included in the range style.

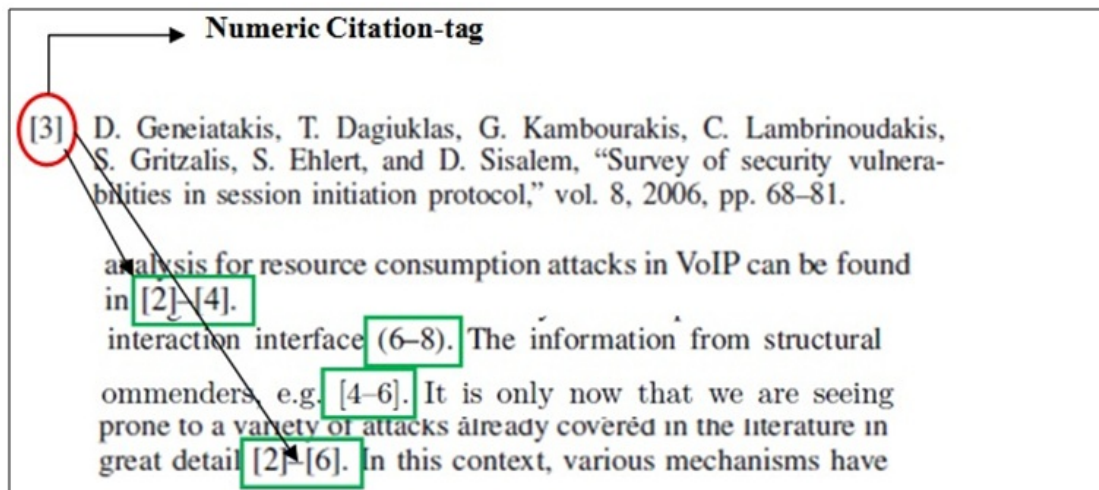


FIGURE 5.3: Mapping of numeric citation-tag on range citation-anchors

The range numeric notation raised the mathematical ambiguity problems during the identification of citation-anchors in text of citing document. The snapshot of these problems are highlighted in Figure 5.4. The red rectangle in the Figure 5.4(a) shows the numeric citation-tag while the red circle and black rectangle in Figure 5.4(b) shows the valid and invalid occurrences of numeric citation-anchors in content of citing research papers for the same citation-tag. This wrong identification of equation number as in-text citation-anchor occurred due to the direct mapping of numeric citation-tag value in content of citing documents.

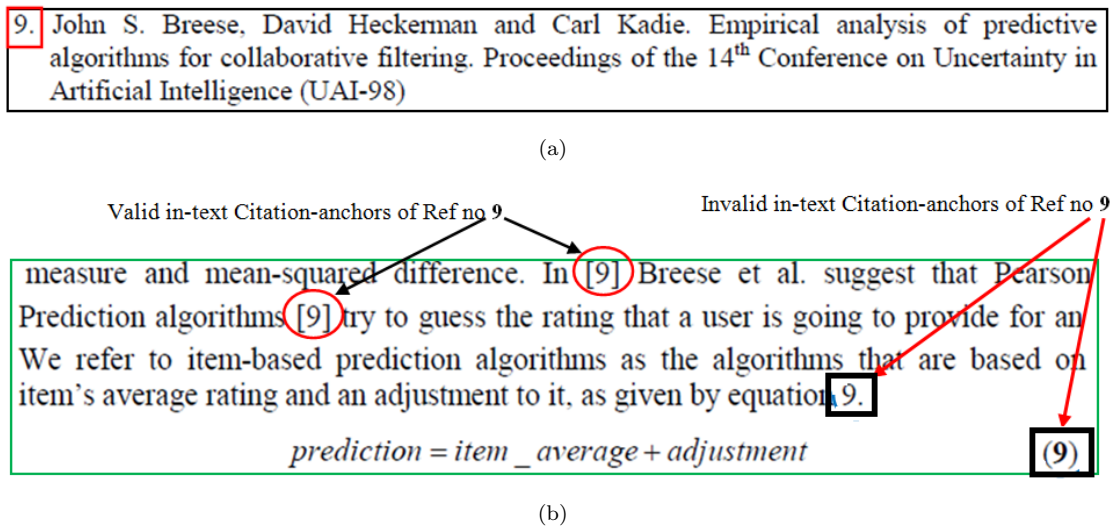


FIGURE 5.4: Incorrect citation-anchor due to mathematical ambiguity. a) Snapshot of reference or citation string with numeric-tag b) Content snapshot with valid and invalid citation-anchors for numeric citation-tag

Compound-anchor problem

In compound-anchor problem, the frequency of numeric citation-tag reduces due to the compound citation-anchors in the text of citing documents. The compound citation-anchors “[1-7, 44, 88]” are constructed by the combination of range-citation “1-7” and multiple citations “44, 88” as shown in Figure 5.5.

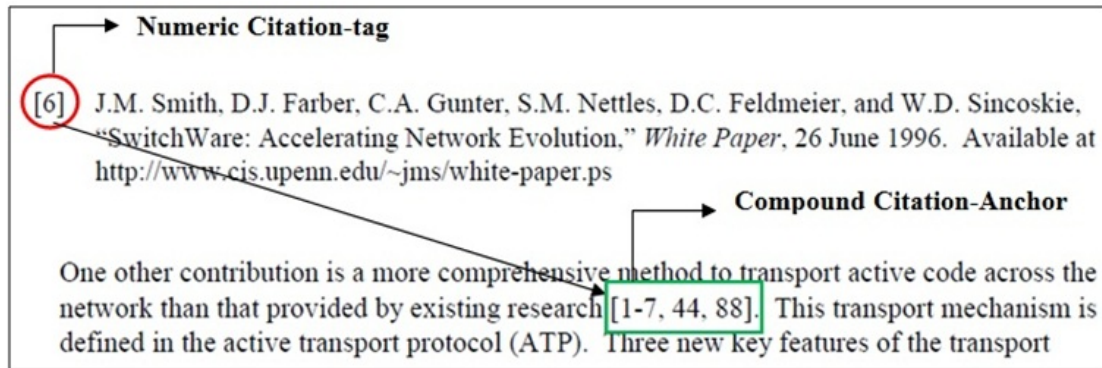


FIGURE 5.5: Citation-tag mapping with compound citation-anchor

5.2.2 String-tags problems

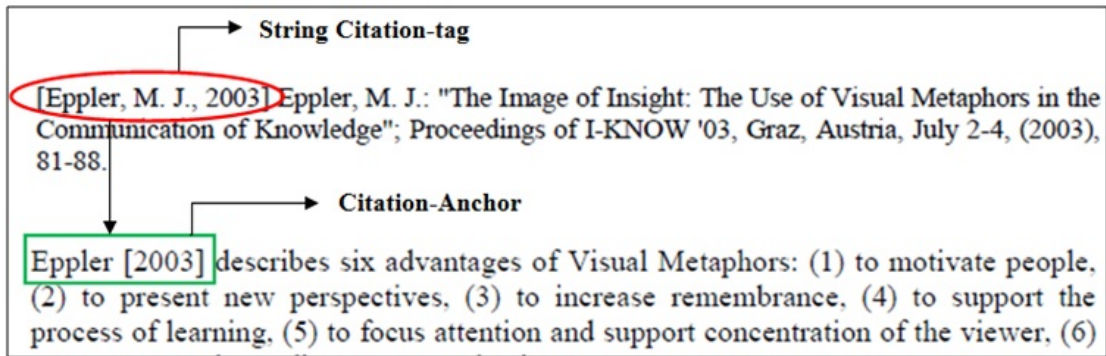
In string-tag problems, the frequency of citation reduces due to a number of problems that are highlighted below with real snapshots.

Format problems

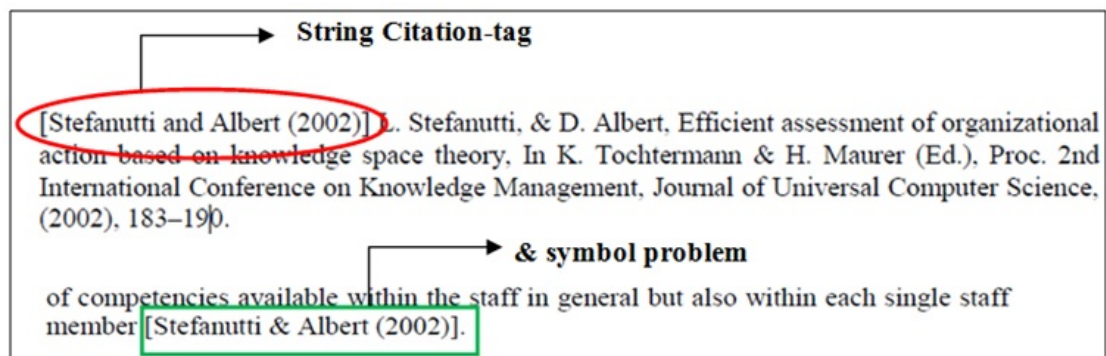
In pattern analysis of string citation-anchors, we observed different format related problems. Some of the real snapshots of these problems are highlighted in Figure 5.6. These problems were detected during the pattern searching of one author, two authors and multiple authors' anchors in text of citing documents. All these problems cannot be detected by exact matching and finally will reduce the frequency of in text citations.

Hyphen with carriage return and line feed problem

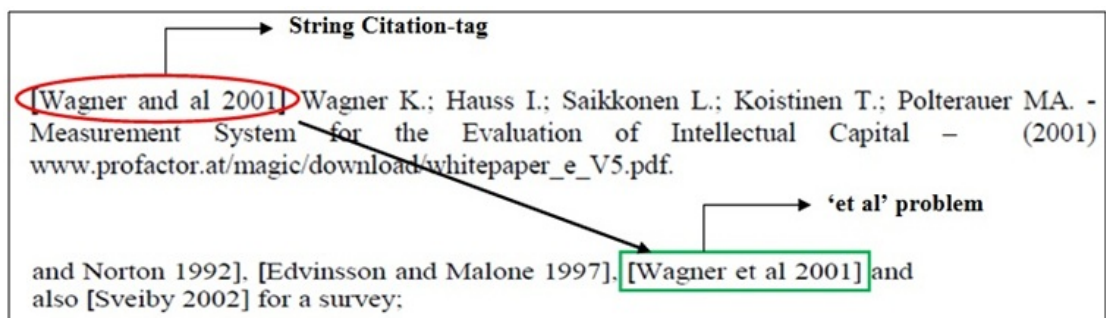
Generally, the research papers are prepared by editing software MS Word and LaTeX. These editing tools automatically add some extra characters such as hyphen, carriage return and linefeed in the text of research paper or other documents. These characters mostly occur with citation-anchors in the research paper. The pattern identification of citation-anchors by different autonomous tools are missed in exact matching [18] due to the inclusion of these extra characters as mentioned in Figure 5.7.



(a)



(b)



(c)

FIGURE 5.6: Format problems with one author, two authors and multiple authors anchor' cases a) One-author case b) "&" symbol problem in two-authors case c) "et al" problem in multiple authors case

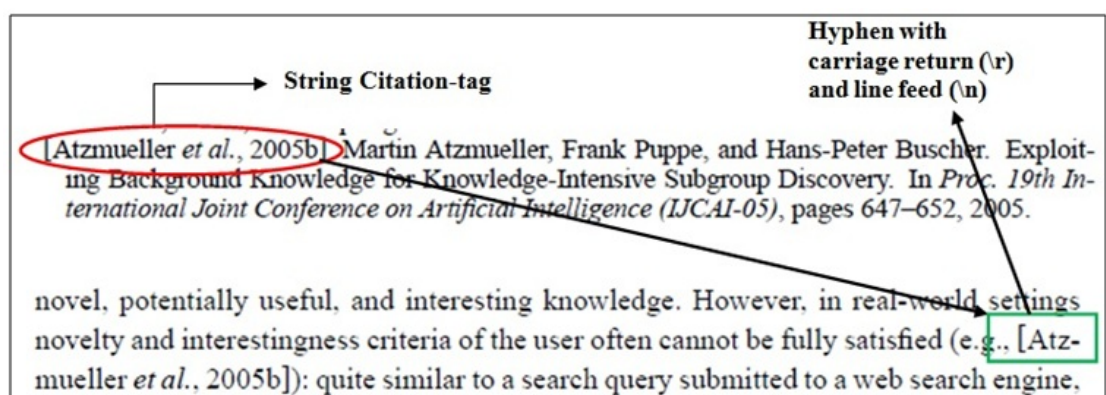
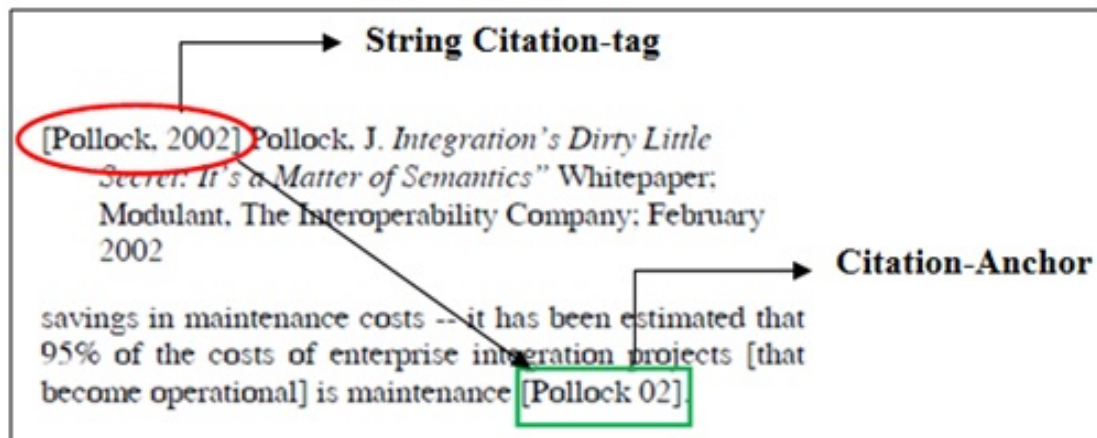


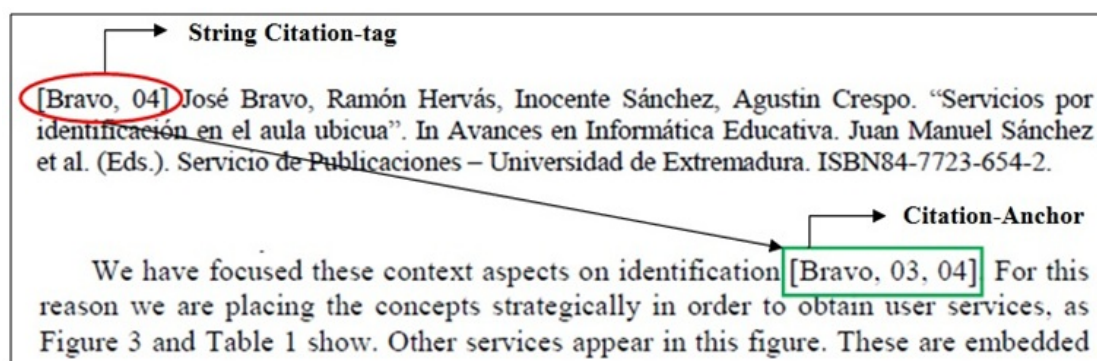
FIGURE 5.7: Carriage return and line feed problem

Year related problems

Usually, the string citation-anchors of the cited documents are constructed by the metadata (authornames, year) in text of citing documents. In the preparation of research papers, authors do not follow the same year format in citation-tags and citation-anchors as shown in Figure 5.8(a). Therefore, the occurrence of citations are missed by automatic tool in text of citing document due to the format variation in publication year, such as “Pollock, 2002” and “Pollock 02”. In the same way, mostly authors cite more than one papers of the same author with different years in single citation-anchor, such as “[Bravo 03, 04]”. By the inclusion of extra year in the citation-anchors, the citation-tag such as “[Bravo 04]” does not exactly match with the citation anchor as mentioned in Figure 5.8(b).



(a)



(b)

FIGURE 5.8: Year related problems a) Year format problem b) Year inclusion problem

Space character problem

In the pattern analysis of citation-anchors, often frequency of citations in text of citing document reduces due to lack of proper spacing in the citation-anchors. Hence, the citation-tags do not match exactly with citation-anchors as shown in Figure 5.9.

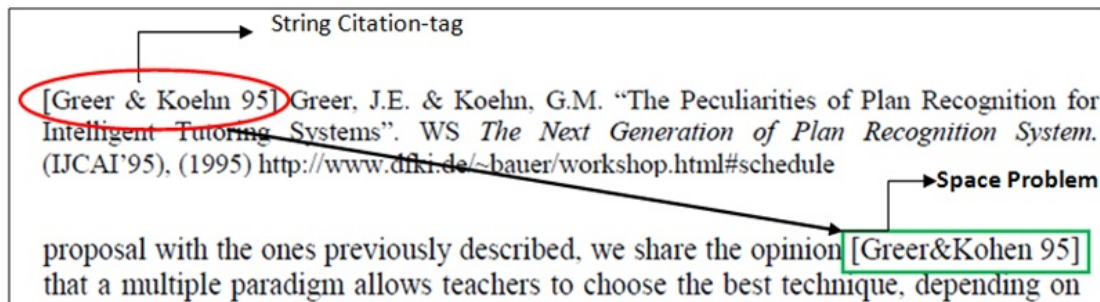


FIGURE 5.9: Citation-anchor with space character problem

Citation-anchor with POS problem

In the citations representation process in text of citing document, the authors also indicate the citation-anchors along with part-of-speech (POS), such as “rank scoring criteria”. These additional characters among the author name and publication year cause the reduction of citations frequency in text of citing document. The real snapshot of research paper is given in Figure 5.10.

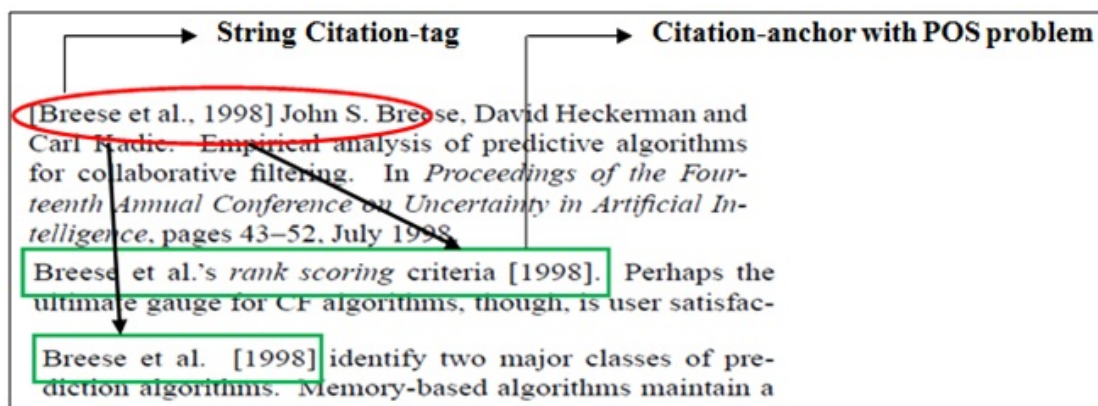


FIGURE 5.10: Citation-anchor with POS (part-of-speech) problem

Reference string without citation-tag Problem

In state-of-the-art technique [18], the pattern and frequency identification of citation-anchors depend on the citation-tag. In previous approach, the citation-tags are

detected from the reference string of cited document. Then the citation-tags are matched with citation-anchors in text of citing document. In the paper construction phase, most of the authors present the reference string of cited documents without citation-tags as shown in Figure 5.11. This type of citation-anchors detection fails due to the lack of citation-tag.

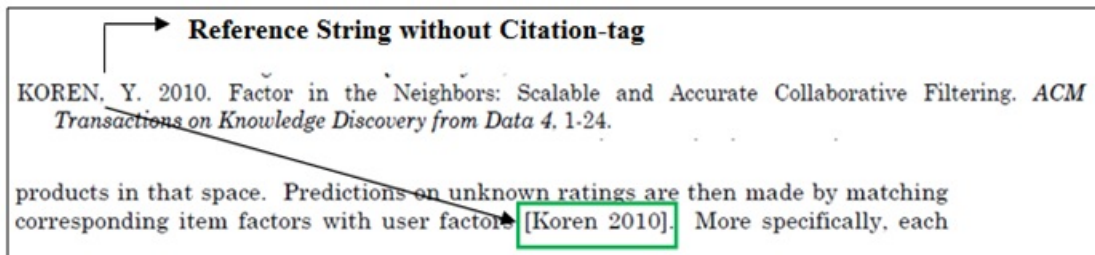


FIGURE 5.11: Reference-string without citation-tag problem

Commonality in Contents

According to Shahid et al [18], some authors use very common citation-tags. For example, reference or citation string shown in Figure 5.12 represents a citation-tag ‘[N]’ in red circle. Here, the contemporary systems will only use the character ‘N’ as a citation-tag. These kinds of citation-tags are very sensitive as ‘N’ is common character which may occur many times in the full text of citing paper and will result in inaccurate calculation of in-text citation frequencies.

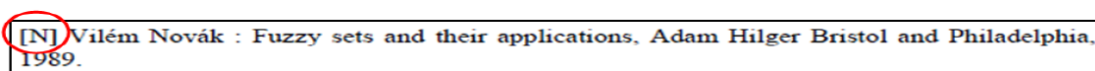


FIGURE 5.12: Common character as Citation-anchor

Reference string with superscript citation-anchor

The superscript is one of citation-anchor formats that is used in different Journals like Nature and Science etc. The cases of superscript format are also analyzed for in-text citation-anchors analysis. One such case is shown in Figure 5.13.

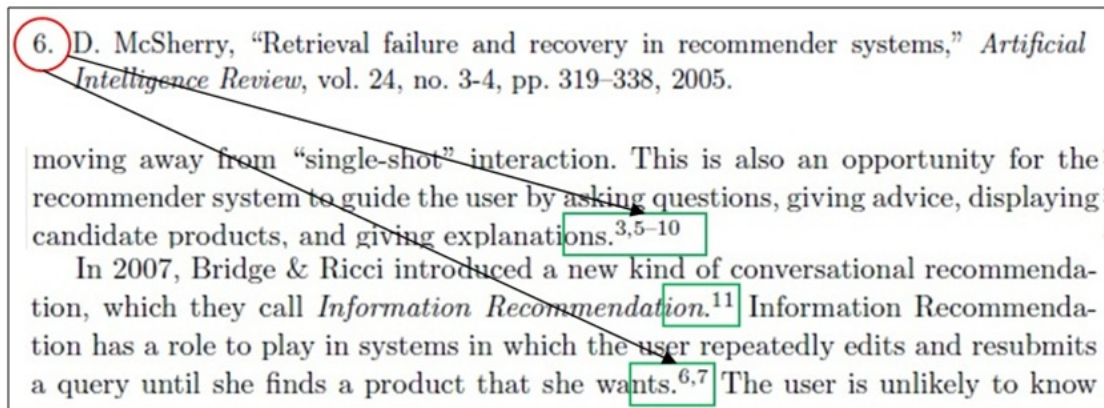


FIGURE 5.13: Reference-string with superscript citation-anchor problem

5.3 Exploratory Analysis of GROBID AND CER-MINE Tools

In the recent study [56], the CERMINE [57] and GROBID [58] are declared best tools for the extraction of metadata and structure from reference strings of citations in the research papers. Therefore, the proposed approach is also evaluated and compared with these two tools in this research work. The manual analysis of CERMINE and GROBID tools are conducted by their online web services available at link² and link³ respectively. During analysis of these tools, the occurrences and frequency of the patterns of citation-anchors are seen and calculated from the research papers in parsed format XML. In the experimental analysis of CERMINE and GROBID tools, we have found different cases of real scenarios which reduces the frequency of citation-anchors in-text of citing document. Some of the cases have been shown as below.

²<http://cermine.ceon.pl/index.html>

³<http://cloud.science-miner.com/grobid/>

5.3.1 String Citation-anchor with Bracket problem

In Figure 5.14, the string citation-anchor with bracket problem is shown. Due to this problem, the identification and frequency of citation-anchor in text of research paper is reduced by CERMINE tool as shown in Figure 5.14(b). The string citation-tag is highlighted in PDF text and XML formats in Figure 5.14(a). Though, the GROBID tool is better performed against this problem. The snapshots in Figure 5.14 are captured from the paper with titled “Collaborative Filtering by Personality Diagnosis: A Hybrid Memory-and Model-Based Approach”.

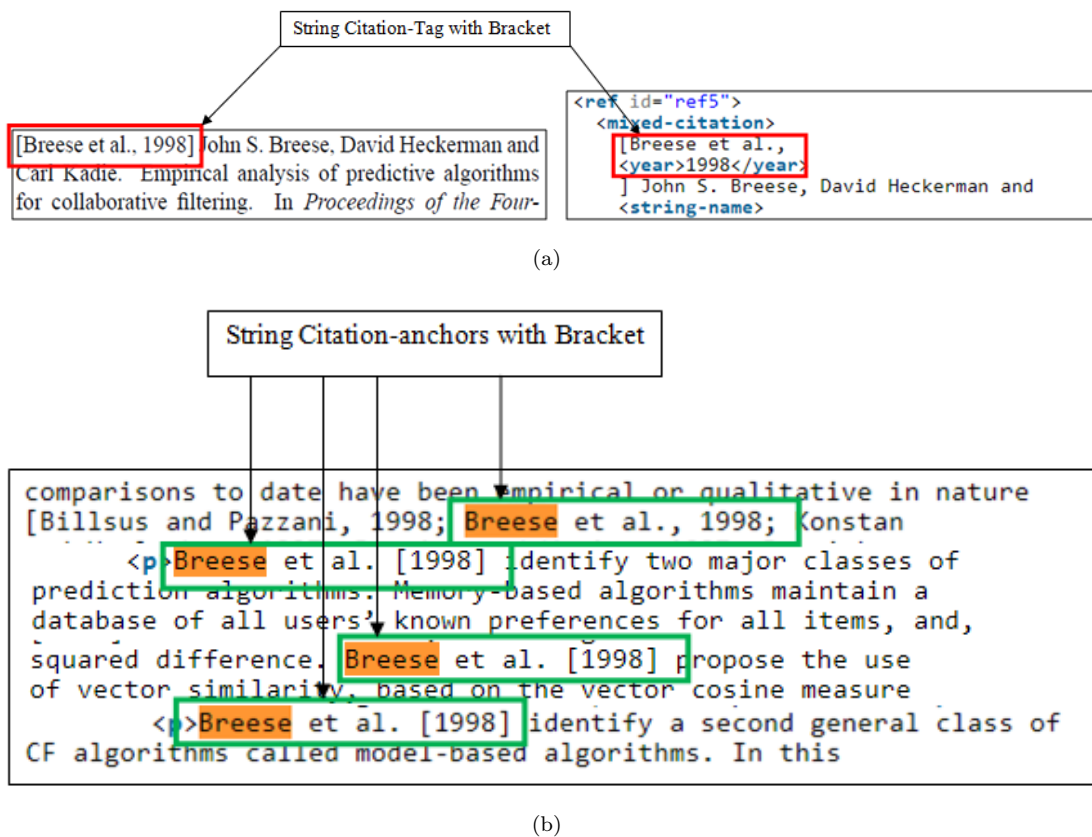


FIGURE 5.14: CERMINE tool with String Citation-anchor with Bracket Problem a) Reference String with String Citation-tag with Bracket in Text and XML formats b)The Missed String Citation-anchors

5.3.2 Citations with Same Author and Year problem

Sometimes, the authors cite more than one citations of same first author published in same year in the citing document as shown in Figure 5.15. In the present of such type of citations, CERMINE tool assigned the wrong reference id to citation-anchors as shown in Figure 5.15(b). The GROBID tool also suffered due to the same problem as given in Figure 5.16.

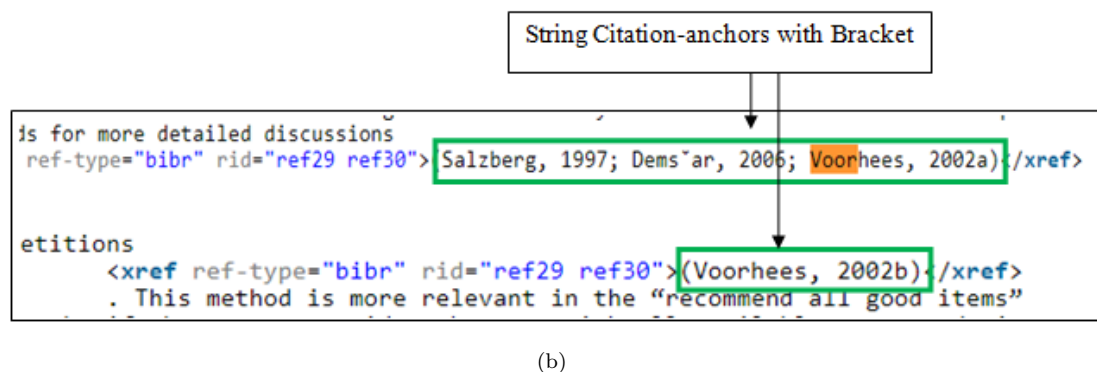
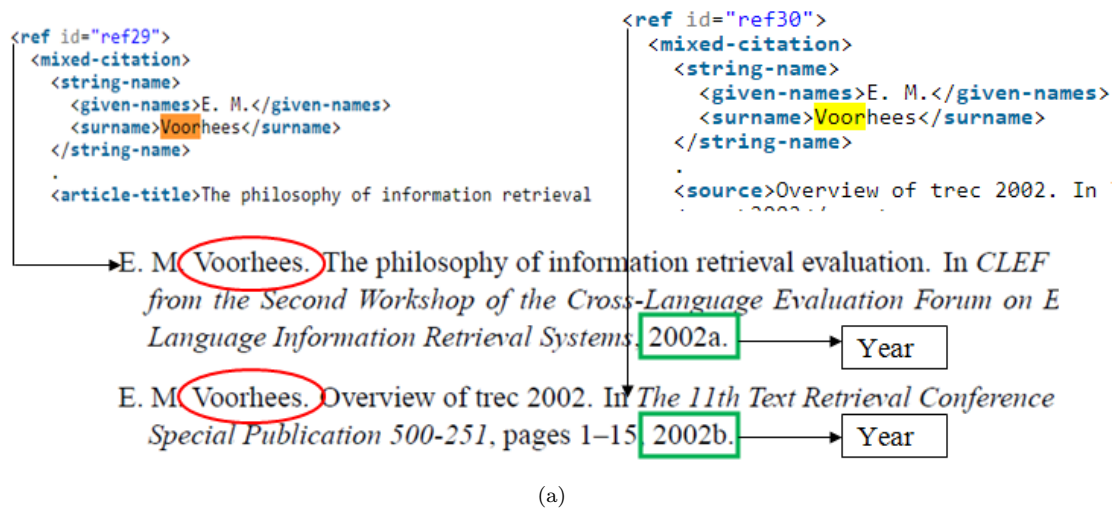


FIGURE 5.15: Citations with Same Author and Same Year Problem a) Reference String in Text and XML formats b) CERMINE tool Assigned the Wrong Reference ID to Citation-anchors

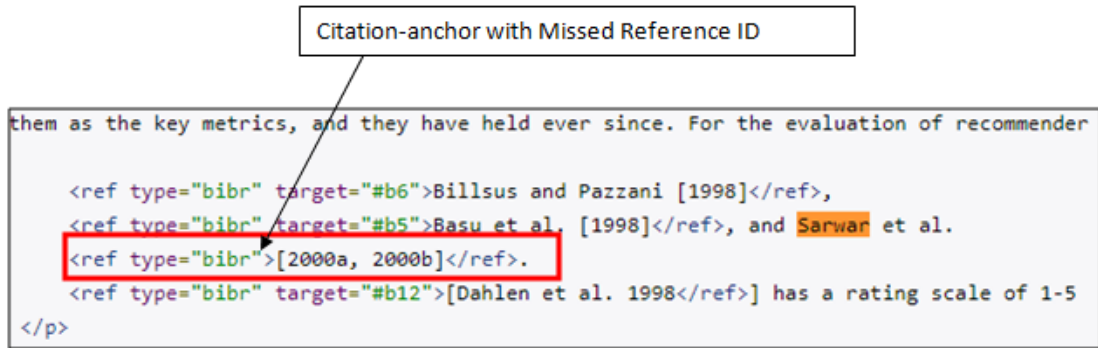


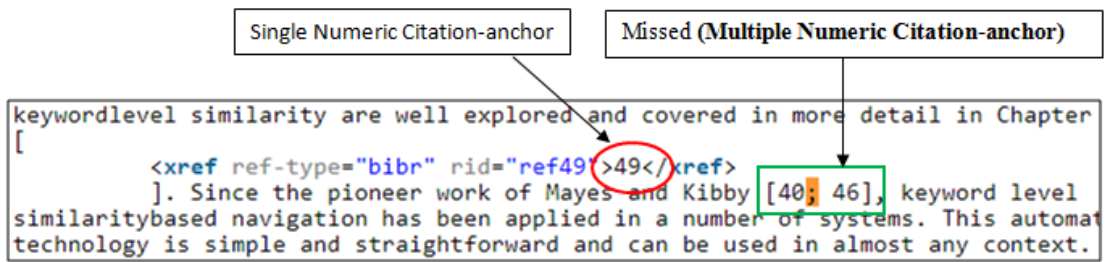
FIGURE 5.16: Missed Citation-anchors with GROBID tool due to Same Author and Year Problem

5.3.3 Multiple Numeric Citation-anchor with Semicolon Problem

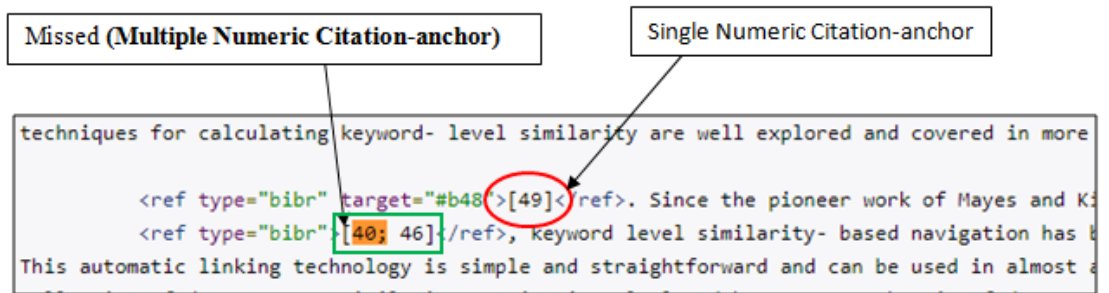
In Figure 5.17, the semicolon problem is shown with multiple numeric citation-anchor. Due to this problem, Both CERMINE and GROBID tools suffered in the identification of patterns and frequency of citation-anchors as shown in Figure 5.17(a) and in Figure 5.17(b) respectively.

5.3.4 CERMINE and GROBID tools Effectuated with Year Inclusion Problem

The snapshots in Figure 5.18 are taken from the book with titled “Recommender Systems for Learning”. The format of citation-anchor such as “Burke (2000; 2007)” is not detected by both CERMINE and GROBID tools during the analysis step. The CERMINE tool missed reference id of both citations anchors such as “Burke 2000” and “Burke 2007” while the GROBID tool missed only the reference id of “burke 2007” citation anchor.



(a)



(b)

FIGURE 5.17: Multiple Numeric Citation-anchor with Semicolon Problem a) CERMINE: Missed Multiple Numeric Citation-anchor b) GROBID: Missed Multiple Numeric Citation-anchor

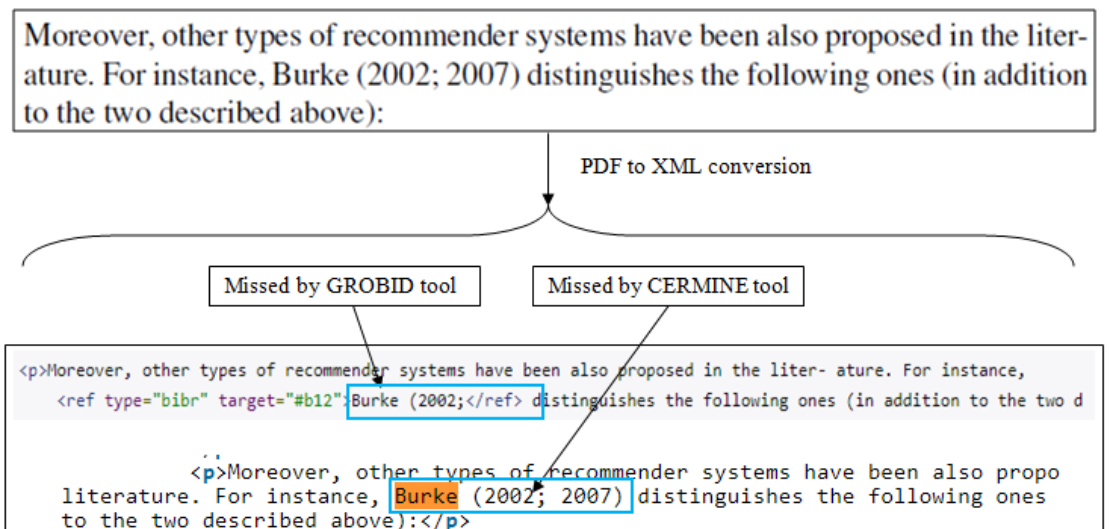


FIGURE 5.18: Missed Citation-anchors with Year Inclusion Problem

5.4 Proposed taxonomy of citation-anchor

The patterns identification of citation-anchors depends on its various styles and formats. The detailed literature review revealed that there was no classification of citation-anchors. Therefore, the taxonomy of citation-anchors was built using a comprehensive procedure. This procedure consists of (1) study of existing state of the art techniques such as: “Giles et al (Giles et al., 1998)”; “Bergmark (Bergmark, 2000)”; “Shahid et al (Shahid et al., 2014)” (2) analysis of standard citation formats APA⁴, MLA⁵ (Garcia, 2010), AMA⁶, and CBE⁷ and (3) experimentation on papers belonging to different domains, such as computer science, medical and biology etc.

The citation-anchor taxonomy contains various types of citation-anchors. For understanding, the proposed taxonomy has been classified into two branches based on their format and style (1) Numeric citation-anchors and (2) String citation-anchors.

Numeric citation-anchors

The numeric citation-anchors were found in two formats, such as plain format and superscript format. Therefore, the numeric category is classified into two sub-categories ,i.e., plain format and superscript format. Each category of numeric-anchors has four sub-parts “Single-anchor”, “Multiple-anchor”, “Range-anchor” and “Compound anchor”. The single anchor is used to represent only one cited paper in the text of the citing document, such as “[3]”, “[1] [2] [3]”. In the multiple anchors, more than one paper is cited in citation-anchor, such as “[1, 2, 3, 4]”. The range anchors consist of range of cited documents, such as “[1-5], [1]-[5]”. The compound type of citation anchor is the combination of either single anchor or multiple anchor and range anchor like “[1-5, 7]”, “[1-5, 4, 6, 9]”. For superscript format, the citation anchor is mentioned as a superscript with the citation text. The format is going to be one of the four as mentioned before.

⁴www.apa.org. American Psychological Association

⁵Modern Language Association

⁶www.lib.jmu.edu/citation/amaguide.pdf American Medical Association.

⁷https://www.libraries.psu.edu/psul/lls/students/cse_citation.html Council of Science Editor

String Citation-anchors

The string citation anchors have different variations. For the ease of understanding, these anchors are classified into four sub-parts “Single-anchor”, “Short-anchor”, “Compound-anchor” and “Parts-of-speech-anchor”.

The single tags are prepared by the use of First author ‘Last name’ and year of publication. These tags have been further classified into two sub-categories based on year ,i.e., “Author with year”, and “Authors without year”. In the ‘author without year’ of the single-anchor, the authors are shown without year like single author “Swets”, two author “Sinha and Swearingen” and more than two authors “Amento et al”. The research papers are written either by one author, two or more than two authors. Based on the number of authors, we have further divided the “Author with year” category into three classes: one author, two authors, and multiple authors.

- One author citation-anchor is used with year in different style “Yao [1995], [Swets, 1995], Swets [1963, 1969] and Harter 1996”.
- The two authors citation-anchor with year has noticed in different variations “Balabanovic and Shohan 1997”, “Billsus and Pazzani [1998]”, “[Sinha & Swearingen 2002]”, “Swearingen and Sinha[2002, 2001]” and “[Wexelblat and Maes 1999]”.
- The citation-anchor with multiple authors has exploited in text of citing document with year in different variations “Amento et al[1999, 2003]”, “Bailey et al. 2001”, “Basu et al. [1998]”, “[Konstan et al. 1997]”, “[Sarwar et al, 2000a, Sarwar et al. 2000b]” and “Sarwar et al. [2000a, 2000b]”.

The short-anchor is the second type of string variation category. It is made by the combination of first character of author names, special symbols (‘+’, ‘*’) and the last two digits of the year ‘Good+98, SkkR*01, Unfo98’.

The third variation of string citation-anchor is compound-anchor. The compound citation-anchor is prepared by the citation of more than one cited document “[bill-sus and Pazzani, 1998; Basu et al, 1998; Basilico and Hofmann, 2004]”.

The fourth variation of string citation-anchor is parts-of-speech-anchor that consists of author name, part of speech and year “Turpin & Hersh’s study of search engines [2001]”.

This taxonomy can be exploited by an automatic program to identify citation-anchors accurately. Currently, citation-anchor taxonomy looks like depicted in [Figure 5.19](#).

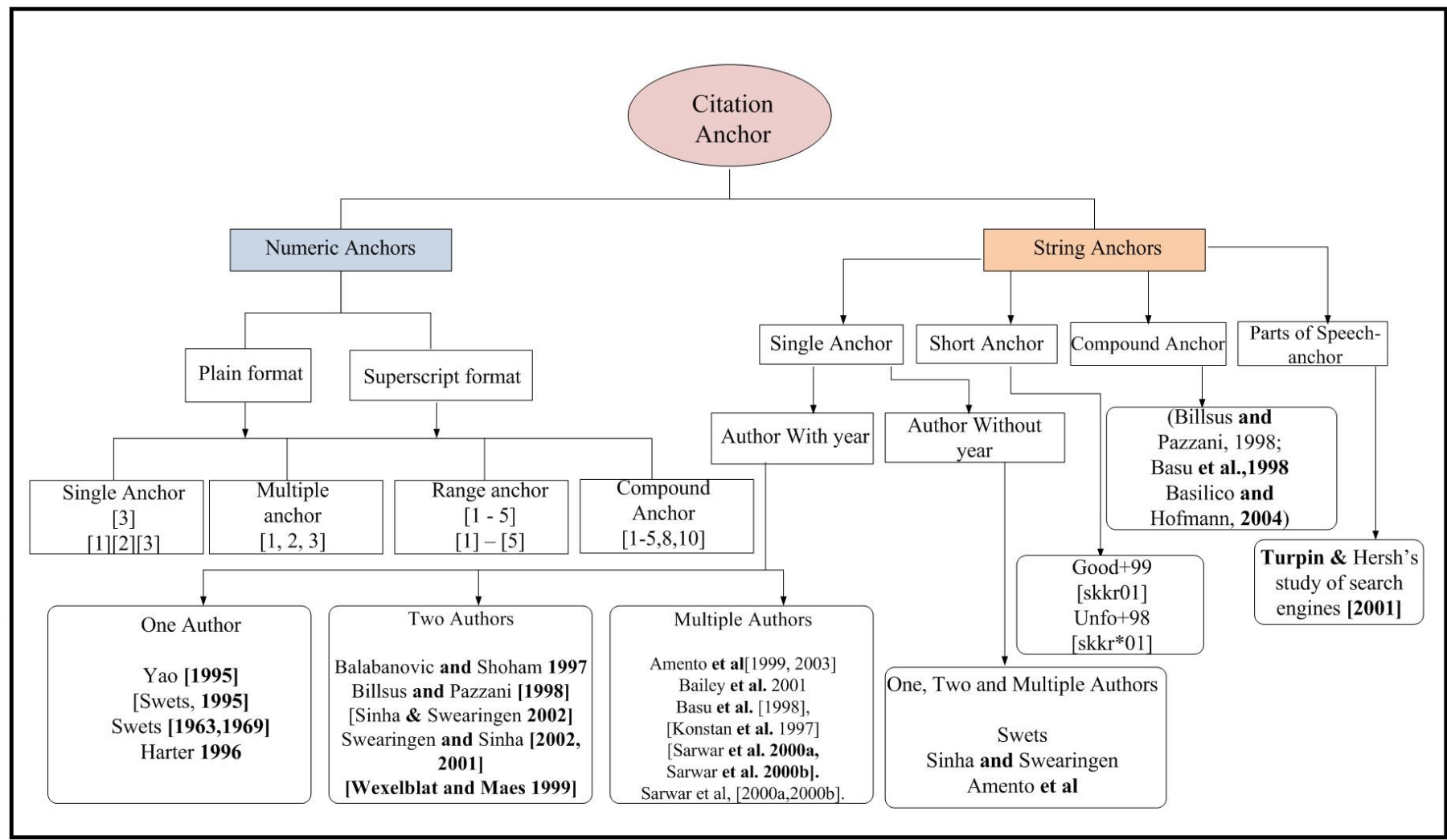


FIGURE 5.19: Citation-anchor taxonomy

5.5 Proposed Architecture for In-Text Citation Patterns and Frequencies Identification Approach

The proposed approach architecture for in-text citation-anchors detection consists of two phases. The first phase is the ‘data preparation’ phase and second phase is the ‘automatic pattern detection of citation-anchors’ phase. The detailed architecture of our proposed system is given in Figure 5.20.

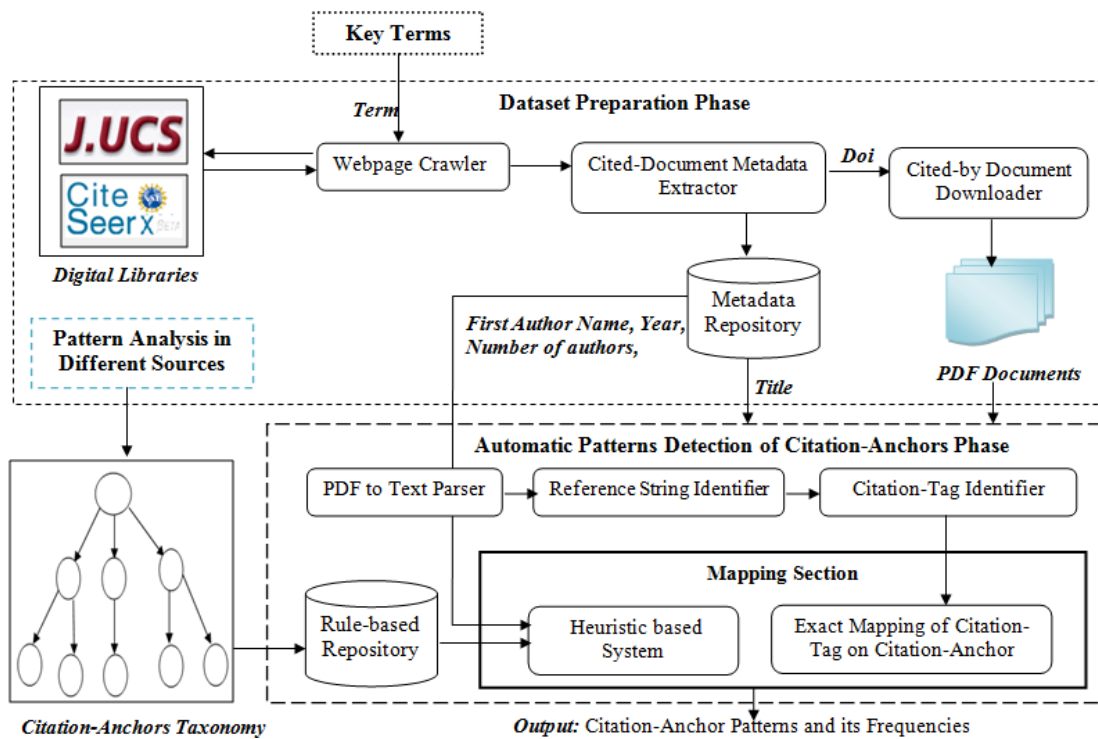


FIGURE 5.20: Proposed architecture for citation anchor detection

5.5.1 Data preparation phase

In this phase, we constructed the dataset for our experimental analysis. The dataset consisted of metadata of two types of documents: cited-documents, and citing documents. The data preparation phase consisted of three sub-components: webpage crawler, cited-document metadata extractor, and citing document downloader.

Webpage crawler

The webpage crawler is a program which systematically browses the selected digital libraries J.UCS and CiteSeer, for the purpose of webpage indexing. Each webpage consists of number of links of cited documents. This program selects the WebPages of cited documents automatically based on set of diversified key-terms as shown in Table 5.1.

TABLE 5.1: Key-Terms for the selection of cited documents

KeyTerms
Recommender System
Information Visualization
Datamining
Web-based Knowledge Discovery
Ontology
Wireless Network
Semantic Web
Distributing Computing
Software Engineering
Information Retrieval

Cited-document metadata extractor

The indexed webpage is further processed by the metadata extractor of cited and citing documents. The extractor program decomposes the link into required metadata informations ‘Title’, ‘Author Names’, ‘Year’ and ‘number of citing documents’. Furthermore, ‘citation-id (cid)’, ‘First-Author’ and ‘number of authors’ information are extracted from ‘citing documents’ and ‘Author Names’ metadata

respectively. Finally, the collected metadata in Figure 5.21 is stored in the metadata repository. For this analysis, we have also prepared the set of citing documents (PDF files) for each cited-document. The extractor exploits the ‘citation-id (cid) and ‘number of citing documents’ to extract the (digital object identifier) ‘DOI’ of each citing document.



FIGURE 5.21: Metadata of cited and citing documents

Citing document downloader

The collection of PDF files for ‘citing documents’ is downloaded by using ‘DOI’ metadata because each document is uniquely represented in World Wide Web (WWW) by unique ‘DOI’. For example, the DOI (<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.167.7612>) denotes the document with title “Item-based Collaborative Filtering Recommendation Algorithms (2001)”.

The function “documentMetadata.Extractor.Downloader” is built to extract the metadata of cited documents, such as Title, Citations, Authors, Venue, PublishedYear, and Doi. In this function, first we have created the URLs of research papers based on keyterms selected from computer science domain. The URLs are further used to get the Webpages that consist the links of cited documents. We have used the DOM parser to extract the tags of different research papers links. Each tag contain the required metadata as mentioned earlier. The metadata of citing documents are also extracted by this function, such as title and doi. The PDFfile downloader uses the Doi of citing documents to download their PDF files of research papers.

```
1: function DOCUMENTMETADATA_EXTRACTOR_DOWNLOADER
2:   Keyterm := getKeyTerm()
3:   Url := createUrlPath(Keyterm)
4:   Content := getWebpageContent(Url)
5:   Tags[ ] := getTags_DOMparser(content)
6:   For i = 0 To Tags.length
7:     //Metadata Extraction
8:     Title := getTitle(Tags[i])
9:     Citations := getCitation(Tags[i])
10:    Authors := getAuthors(Tags[i])
11:    Venue := getVenue(Tags[i])
12:    PublishedYear := getPublishedYear(Tags[i])
13:    Doi := getDoi(Tags[i])
14:    StoreMetadata (Title, Citations, Authors, Venue, PublishedYear, Doi)
15:    PDFfile := PDFfile_Downloader(Doi)
16:  End ForLoop
17: end function
```

5.5.2 Automatic pattern detection of citation-anchors phase

The second phase consists of four key components for the pattern identification of citation-anchors in text of citing documents. These components are (1) PDF to Text Parser (2) Reference String Identifier (3) Citation-Tag Identifier and (4) Mapping Section. The details of each component are discussed below. In the end of this subsection, we have mentioned the algorithm for automatic pattern detection of citation-anchors.

PDF to Text parser

The direct pattern recognition from PDF documents is very tedious task due to the unavailability of proper tool. Hence, the PDF to Text parser component is designed to convert the PDF document into plain-text format. The proposed

parser utilizes the Java PDFbox library for conversion of PDF documents into plain-text.

Reference string identifier

The reference string is the portion of text in the references section of citing documents which represents the citation of each cited document as mentioned in Figure 5.22. The reference string identifier extracts the reference string of cited document from the citing documents using its metadata, such as “Title: Explaining Collaborative Filtering Recommendation”, “First Author Name: (Herlocker), and “Year: 2000”. The reference string identifier uses these metadata information in regular expression.

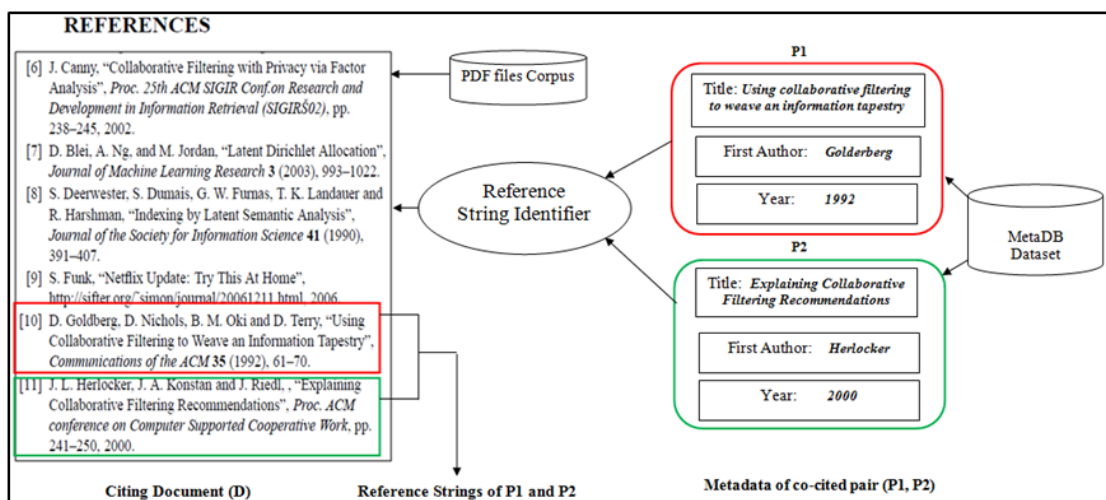


FIGURE 5.22: Reference string extraction

Citation-tag identifier

Citation-tag is the unique identifier which is used at the start of each reference strings. It is shown in red small circle in Figure 5.23. The citation-tag identifier component is added to identify and extract the various patterns of citation-tag from reference strings by using different regular expression. For example the numeric citation-tag such as ‘1’, ‘1.’, ‘[1]’, ‘[23]’ etc can be extract by using regular expression “ \n?((\ (| \ [] ? \ [1-9] [0-9] * \ (| \) ? | [0-9] {1,3} \) . ? ” from any reference string with numeric citation-tag. Furthermore, these citation-tags are used in mapping section to detect the different patterns of citation-anchor as

discussed in Figure 5.19. The citation-anchor in large green circle is highlighted in Figure 5.1.

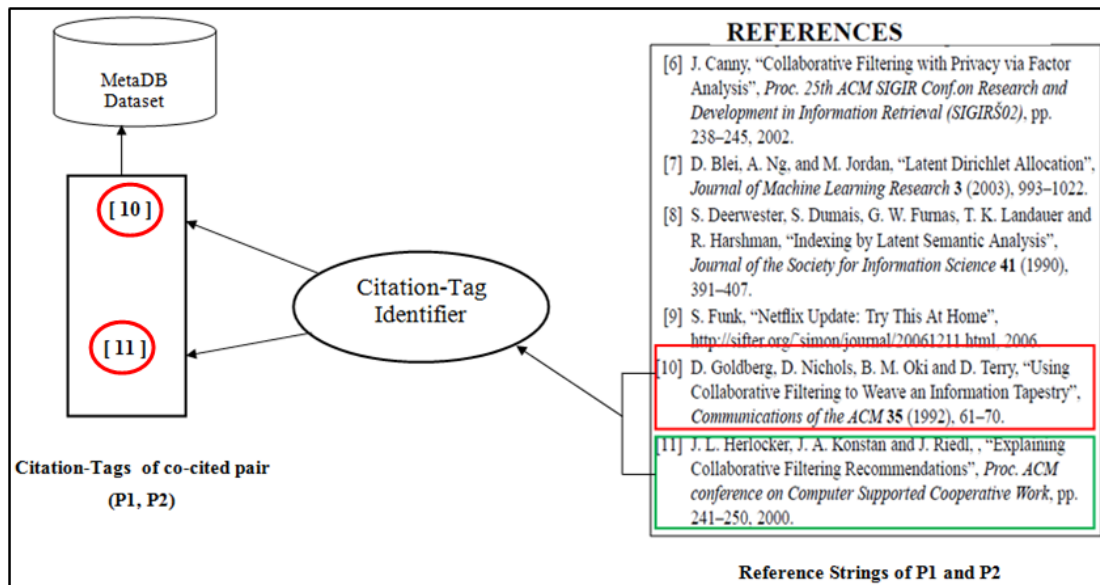


FIGURE 5.23: Numeric citation-tag extraction

Section mapping

The section mapping is the component of proposed architecture in Figure 5.20. In this component, the patterns identification and extraction of different citation-anchors as in Figure 5.19 are performed by using two types of methods (1) Exact mapping of citation-tag on citation-anchor, and (2) Heuristic based system. The latter approach [18] is based on only exact mapping method, while the proposed approach combines exact mapping and heuristic based methods. In the exact mapping method, the extracted citation-tags are exactly mapped with patterns of citation-anchors in text of citing document. This method is beneficial when the format of both citation-tags and citation-anchors are similar. All those cases in section 5.2 could not be properly detected by the exact mapping method due to the variation between citation-tags and citation-anchors. Therefore, the heuristic based system is added in our proposed system. This system utilizes different pre-defined rules and metadata ‘First name of author’, ‘number of authors’ and ‘publication year’ that are stored in rule-based repository and metadata repository respectively. The rule-based repository is constructed based on the proposed citation anchor taxonomy (CAT) shown in Figure 5.19.

The “Intext_Citation_Frequency_Identification” function is developed to detect the patterns of citation-anchors from the text of citing document. The function ‘getCitingDocument()’ get the PDFfile of citing document. Then, the PDFfile is converted into plaintext by using PDFbox java library. The Metadata such as ‘Firstauthor’, Title, and Year have been used to extract the citation-tag of the cited document. The citationtag and PDFfile have been passed as inputs to CAD algorithm. The CAD algorithm is also mentioned as below.

```
1: function INTEXT_CITATION_FREQUENCY_IDENTIFICATION
2:   PDFfile = getCitingDocument( )
3:   PlainText = PDFboxJavaLibrary(PDFfile)
4:   Citeddocument = getCited_Document()
5:   Firstauthor = getMetadata(Citeddocument)
6:   Title = getMetadata(Citeddocument)
7:   Year = getMetadata(Citeddocument)
8:   CitationTag = getCitationTag (Firstauthor, Title, Year, PlainText)
9:   CAD (CitationTag, PDFfile)
10: end function
```

5.5.3 Patterns for citation-anchors identification

In this section, after the deep analysis of randomly selected 3,000 citations out of 17,850 total citations in order to solve the problems related to correct recognition of citation-anchors from the text of citing documents, we propose a two-stage approach. In the first step, regular expressions are devised for matching the patterns of citation-tags and citation-anchors from the text of citing documents. We then use these regular expressions in our rule based Citation Anchor Detection (CAD) algorithm which extracts the patterns and frequencies of citation-anchors from a given document. The regular expressions and the CAD algorithm are discussed below.

Regular Expression

In our experimental study, different patterns are developed for identification of

citation-anchors presented in Figure 5.24. We have divided these regular expressions into three categories. In category (A), the regular expressions are prepared for the identification of numeric-anchors in text of citing document. In category (B), the regular expressions are designed to represent the string citation-anchors. These regular expressions are further divided into two sub-categories, such as ‘B.1 and ‘B.2 based on delimiter symbols with citation-anchors like ‘(author, year) or ‘[author, year]. The regular expressions in category ‘A and ‘B are static to highlight the concerned patterns in ‘Citation_Anchor_Patterns column while in category (C), the dynamic regular expressions are prepared by calling function ‘Dynamic_RegEx_Delimiter as shown in S-CAD Algorithm. All regular expressions are verified with ‘Edit-Pad Pro 7 tool at link⁸ and then executed in java code.

⁸<https://www.editpadpro.com/>

RE_No	Regular Expressions	Citation Anchor Patterns
(A) Regular Expressions for Numeric-Anchors Detection		
1	<code>\[+Tag_Value+]</code>	[4], [66]
2	<code>\[s*([1-9][0-9\u2013-]*s*[z; \u2013](s\)*)+[1-9][0-9]*s*(-[1-9][09]*)?\ [[1-9][0-9]*[- \u2013][1-9][0-9]*s*]</code>	[2, 3, 4], [22; 3 ; 4], [2 - 4], [3] – [6], [4-6, 8], [3, 6-8], [3, 5-8, 10]
3	<code>[1-9][0-9]*[\u2013-][1-9][0-9]*</code>	5 – 6 or 5-67
(B) Regular Expressions for String-Anchors Detection		
(B.1) Bracket Cases		
4	<code>\[[A-Za-z0-9-:&\.\\s,;:\ +^()\]*Tag_First_character[A-Za-z0-9-:&.\ +^s;/\ _]*Tag_Last_character[A-Za-z0-9-:&.\ +^s,^()\]*\]</code>	[sfc+01],[BG*],[DB-Main] ,[bcf+98, bb+99] [Wol94,GV95,SDMH00, SVM01] etc
5	<code>\[[A-Za-z0-9-:&\.\\s,;:\ +^_]*bTag_Value\b[A-Za-z0-9-:&.\ +^s;/_]*\]</code>	[N], [N, P, X], [P, X, N], [X, N, P] , [N, bb+99]
6	<code>\[[sA-Za-z0-9-:&\.\\s,;:\ +^()\]*First_Author_name[0-9-:&+^s,;:\ /[]*Year[sA-Za-z0-9-:&.\ +^s,;:\ /[]]*\]</code>	[sato, 1979], [leuf, 01], [Woitsch,03b],[Allen (1983)] [Christensen (98)], [Hom, R. E., 1998] etc
7	<code>\[[A-Za-z0-9-:&.\ +^s,;:\ /[]]*First_Author_name[A-Za-z-;:\ +^s,;:\ /[]]*(\band\b &)[A-Za-z-;:\ +^s,;:\ /[]]*\bYear\b[A-Za-z0-9-:&.\ +^s,;:\ /[]]*\]</code>	[doignon and falmagne, 199 9], [Goel & Grafman, 1995], [Borland/Inprise 2002b] , [Gr oth and Bowers 01], [Lafra nce and Mullins 2002a] etc
(B.2) Parenthesis Cases		
8	<code>\[[sA-Za-z0-9-:&\.\\s,;:\ +^_]*First_Author_name[0-9-:&+^s,;:\ /[]*Year[sA-Za-z0-9-:&.\ +^s,;:\ /[]]*\]</code>	(bradley 1997) (bradley 1997; catlett, 1995; provost and fawcett, 1997) etc
9	<code>\[[A-Za-z0-9-:&.\ +^s,;:\ /[]]*First_Author_name[A-Za-z-;:\ +^s,;:\ /[]]*(\band\b &)[A-Za-z-;:\ +^s,;:\ /[]]*\bYear\b[A-Za-z0-9-:&.\ +^s,;:\ /[]]*\]</code>	(odonovan and smyth 2005), (burke 1999 thompson and langley 2004 ricci 2002) etc
(C) Dynamic Regular Expression generation by Dynamic RegEx Delimiter procedure		
10	Dynamic_Regular_Expressions generated by Dynamic_RegEx_delimiter procedure as mentioned in Figure 8 (E). This procedure is used to detect the different variants of two-author and multiple-author cases.	[Berendt et al. 2002] , [Neumann, Morgenstern, 1944], [Kantorovich, Vulich, Pinsker, 1959], [Hollman, van Lint, Linnartz and Tolhui-zen 98], OR (Sarwar et al 2000), (Konstan et al 1997, Schafer et al 1999, Sarwar et al, 2001) etc

FIGURE 5.24: Regular expressions for citation-anchors identification

Pseudo code for CAD (Citation Anchor Detection) Algorithm

In this work, we have proposed an algorithm called citation-anchors detection (CAD) Algorithm as below. The citation tag of either query paper or co-cited paper and PDF file of citing document is used as input while citation-anchor patterns and their frequencies are of output in this algorithm.

```

1: function CAD (CT, CD)
   Input: CT: Citation-Tag, CD: Citing-Document
2:   TCD := PDFboxLibrary(CD) // TCD → Text of Citing document
3:   IF CT is Numeric Then //Test for Numeric Tags
4:     Call N-CAD (CT, TCD)
5:   ELSE //Test for String Tags
6:     Call S-CAD (CT, TCD)
7:   ENDIF
8: end function

```

The proposed algorithm consists of two sub-algorithms ,i.e., N-CAD Algorithm and S-CAD Algorithm. The CAD algorithm was prepared based on regular expression as shown in Figure 5.24 and some heuristics that are represented in its different rules. In CAD algorithm, the PDF document parsed into text format TCD (Text Citing Document) using PDFbox Java Library. The calling of one of the two sub-algorithms is based on citation-tag format. The N-CAD algorithm is called for numeric citation-anchors detection and S-CAD algorithm is called for string citation-anchors detection.

In the N-CAD algorithm, citation-tag and text of citing document have used as inputs. The regular expression 1 in line 3 of N-CAD algorithm has been used in the function ‘single-anchor-matching’ at line 5 to detect single numeric citation-anchor ,i.e., ‘[4]’. Similarly, the regular expression 2 in line 4 has been exploited in function ‘Multi_Range_Comp_Anchor_Matching’ to retrieve multiple, range, and compound numeric citation-anchors ,i.e., “[2, 4, 5], [3-4], [4,5, 9-11]”. The lines 8 to 17 in N-CAD algorithm have been constructed to preprocess the output of the function ‘Multi_Range_Comp_Anchor_Matching’. For example, the function

“Convert_range_into_values” has been used to convert the range pattern of citation-anchor “3-5” into the string of citation-anchor values “3 4 5” at line 12. Finally, the lines 18 to 23 have been used to find and store the patterns and frequencies of total numeric citation-anchors in a citing document. All issues related with numeric citation-anchor detection as discussed in Figure 5.2.1 are resolved by N-CAD algorithm.

```

1: function N-CAD (CT, TCD)
   Input: CT: Citation-Tag, TCD: Text of Citing-Document
   Output: Patterns and Frequencies of numeric citation-anchors
2:   SAV :=  $\emptyset$  //String of Citation-anchor values
3:   RE1 := RegExp_No(1) //See Regular expression 1 in Fig 5.24
4:   RE2 := RegExp_No(2) //See Regular expression 2 in Fig 5.24
5:   SNAP := Single-Anchor-Matching(RE1, TCD) //Single Numeric anchor
   pattern
6:   Count1 := Pattern_count(SNAP)
7:   MRCP := Multi_Range_Comp_Anchor_Matching(RE2, TCD)
8:   MRCP(p):= Preprocessing (MRCP (p))
9:   for p = 1 to MRCP.length do// Multiple, Range, and Compound patterns
10:     MRCP(p):= Preprocessing (MRCP (p))
11:     IF MRCP(p).Matcher()== true Then
12:       Values:= Convert_Range_Into_Values(MRCP(p))
13:       SAV:= SAV + “ ” + Replace_Range(MRCP(p), Values)
14:     ELSE
15:       SAV:= SAV + MRCP(p)
16:     ENDIF
17:   end for
18:   MP := Search_Citation_Tag_Value(SAV, CT) //Matched patterns
19:   Count2 := Pattern_count(MP)
20:   Patterns := SNAP + MRCP // add Reg Exp1 and Reg Exp2 patterns
21:   Frequency := Count1 + Count2 //add Frequencies of Expression 1 & 2
22:   Store_Pat_Freq(CT, Patterns, Frequencies)
23: end function

```

All issues in section 5.2.2 related with string citation-anchor detection are resolved by using S-CAD Algorithm. The citation-tag and text of citing document have been used as inputs in S-CAD algorithm. This algorithm will produce the patterns and frequencies of string citation-anchors in a citing document. The lines 2 to 22 have been developed to detect the string citation-anchors which are defined with

bracket, such as “[author, year], [author and author, 2004]”. The regular expressions 4 to 7 as shown in Figure 5.24 are used for the detection of citation-anchors with bracket. The lines 23 to 34 have been used to detect the string citation-anchors with parenthesis, such as “(author, 2000)”. The regular expression 8 and 9 in Figure 5.24 are used to detect the string citation-anchor with parenthesis. The function ‘Dynamic_RegEx_Delimiter in S-CAD Algorithm is also used to generate the dynamic regular expressions for various patterns of one, two and multiple authors cases as shown in ‘C category of Figure 5.24. This function also handles string citation-anchors along with parenthesis and bracket such as “(authors et al, 2000) or [authors et al, 2000]”.

```

1: function S-CAD (CT, TCD)
  Input:CT:Citation-Tag, TCD: Text of Citing-Document
  Output: Patterns and Frequencies of String citation-anchors
2:   IF F_Char(CT) == '[' AND S_Char(CT) is Alphabet== True Then
3:     IF CT not contains space Then
4:       IF CT_lenCT > 1 Then
5:         FC:= F_Character(CT) //bcf+98 FC(First character)→ b
6:         LC:= L_Character(CT.length-1)//bcf+98 LC>Last Character)→ 8
7:         RegEx := RegEx_No(4) //See Reg Exp in Fig 5.24
8:       ELSE
9:         RegEx := RegEx_No(5) //See Reg Exp in Fig 5.24
10:      ENDIF
11:     ELSE
12:       CT_Words:=CT_Split(" ") //CT_Words contains tags values
13:       IF CT_Words.length == 2 Then //Test one Author Case(Author,year)
14:         RegEx := RegEx_No(6) //See Reg Exp in Fig 5.24
15:       ENDIF
16:       IF CT_Words < 4 AND CT_Words contains "and" Then
17:         RegEx := RegEx_No(7) //See Reg Exp in Fig 5.24
18:       ELSE
19:         RegEx := Dynamic_RegEx_delimiter(CT_Words,[''])//Check Procedure
  in Fig 5.17(c)
20:       ENDIF
21:     ENDIF
22:     ELSE
23:       IF FC(CT) == '(' AND SC(CT) is Alphabet== True Then
24:         CT_Words := CT_Split(" ")
25:         IF CT_Words.length == 2 Then //Test One Author Case
26:           RegEx := RegEx_No(8) //See Reg Exp in Fig 5.24
27:         ENDIF
28:         IF CT_Words.length < AND CT_Words contains "and" Then
29:           RegEx:= RegEx_No(9) //See Reg Exp in Fig 5.24
30:         ELSE
31:           RegEx:= Dynamic_RegEx_Delimiter(CT_Words, '(')
32:         ENDIF
33:       ENDIF
34:     ENDIF
35:     Patterns := get_patterns(RegEx, TCD)
36:     Count := get_Frequency(RegEx, Patterns)
37:     Store_Pat_Freq(Patterns,Count)
38: end function

```

```

1: function DYNAMIC_REGEX_DELIMITER (CT_W, DELIMITER)
  Input:CT_W is the set of citation-tag words, delimiter symbols like '[' or '('
  Output: DRE: Dynamic Regular Expressions
2:   DRE:=∅ //Dynamic Regular Expressions
3:   FAN := CT_W(1) //First Author Name
4:   Year := CT_W(CT_W.length) //Publication Year
5:   for word:=1 To CT_W.length-1 do// Multiple, Range, and Compound
  patterns
6:     IF word == 1 Then
7:       DRE:= DRE + “[\sA-Za-z0-9,&.;; + /()-}]* + FAN + “[0-9-&\s,;:[()]*”
8:     ELSE
9:       Two_char:= Substring(CT_W(word), 0, 2) //First two Characters
10:      DRE:= DRE + “\s*(” + Two_char + “[0-9A-Za-z-& + \s,;\.[]*”
11:     ENDIF
12:   end for
13:   DRE:= DRE + Year + “A-Za-z0-9-\s\.,; + (-”
14:   IF delimiter == '[' Then
15:     DRE:= “[” + DRE + “\]”
16:   ELSE
17:     DRE:= “\ (“ + DRE + “\)”
18:   ENDIF
19: end function

```

5.6 Experimental setup

In this section, we present two datasets, evaluation metrics and the experimental results.

5.6.1 Datasets

For the experimental study, two citation based datasets are prepared one from the comprehensive journal of computer science known as Journal of Universal Computer Science (J.UCS) and the other is the largest digital library of Computer Science known as CiteSeer. The J.UCS dataset is taken from the Shahid et al’s work that consists of more than 1,200 citing documents along with 16,000 citations [18]. The references are extracted from the XML format of PDF documents. The XML format is obtained by PDFx [6] online tool at link⁹. Some of approaches like [1] attempted to extract references from text format. The CiteSeer dataset

⁹<http://pdfx.cs.man.ac.uk/>

is prepared in this study from the openly available CiteSeer digital Library that consists of 52 citing documents and 1,850 citations. The statistics of both datasets are shown in Table 5.2. The first dataset consists of 2,258 reference strings with numeric-tag (RS-NT) and 13,742 reference strings with string-tag (RS-ST). The second dataset contains 1,850 references with 1,380 (RS-NT) and 470 (RS-ST). In this way, the total citing documents become 1,252 which consist of 17,850 references. The total citation-anchors in both datasets are 28,550. Further details of citation-anchors is given in Table 5.2. The total data was divided in the following way: more than 3,000 citations out of 17,850 citations were used as training set and the remaining citations were used for testing the proposed approach.

TABLE 5.2: Statistics of Datasets

Datasets	Citing documents	References	RS-NT	RS-ST	Citation-anchors
J.UCS	1,200	16,000	2,258	13,742	25,365
CiteSeerX	52	1,850	1,380	470	3,185
Total	1,252	17,850	3,638	14212	28,550

For evaluation of the proposed approach on diversified data, the CiteSeerX dataset was also prepared. This dataset had 1000 papers selected from the queries mentioned in Table 5.1. For each of the 1000 cited papers, 20 citing papers were added in the dataset making the total of 20,000 citing documents. The dataset consists of citing documents, reference strings of citations with numeric-tags (RS-NT), reference strings of citations with strings-tags (RS-ST) and Citation without citation-tags (C-WT). The statistics of this dataset shown in Table 5.3.

TABLE 5.3: CiteSeerX dataset specifications

Dataset	Citing Documents	Cited Documents	RS-NT	RS-ST	C-WT
CiteSeerX	20,000	1,000	14,000	1,200	4,800

For further comparison and evaluation of our proposed approach with both CER-MINE [56] and GROBID [58] tools, we have prepared the extended dataset from CiteSeer digital library. This dataset consists of 250 cited documents and 5,008

citing documents. Each cited document is analyzed in 20 different citing documents for the identification of citation-anchor. The total 8,134 citation-anchors of 250 cited-documents founded in 5,008 citing documents as shown in Table 5.4.

The accuracy of the proposed approach was checked by manual process. For the manual process, we have distributed the set of 1,000 citations among 3 MS and 2 PhD students in our research lab. Each student have analyzed and annotated 200 citations in citing document to build the gold-standard of citation-anchors frequencies and patterns. Then the result of proposed approach is compared based on gold-standard with state-of-the-art approach [18] and existing online tools i.e GROBID and CERMINE

TABLE 5.4: Statistics of CiteSeerX Extended dataset

Dataset	Citing Documents	Cited Documents	Citation-anchors
CiteSeerX	5,008	250	8,134

5.6.2 Evaluation metrics

The evaluation metrics precision, recall, and F-score measures [89] are widely used in information retrieval community. Here, we define recall, precision and F-score in the context of citation-anchors identification.

The correct number of in-text citations frequency in total retrieved frequency of cited document from the citing document is known as true positive (TP) frequency. The incorrect number of in-text citations frequency in total retrieved frequency of cited document from the citing document is called false positive (FP) frequency. The false negative (FN) frequency is the number of correct citations frequency that can not be identified in citing document during retrieving of in-text citations frequency. Precision is the fraction of retrieved patterns of citation-anchors that are relevant as given in Equation 5.1.

$$Precision = \frac{Matches(TP)}{Matches(TP) + Incorrect(FP)} \quad (5.1)$$

Recall is the fraction of relevant patterns of citation-anchors that are retrieved from each citing document as shown in Equation 5.2.

$$Recall = \frac{Matches(TP)}{Matches(TP) + Missed(FN)} \quad (5.2)$$

F-score is the weighted average of precision and recall. It is calculated by using Equation 5.3 .

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5.3)$$

5.6.3 Results

We have performed comprehensive experiments on both J.UCS dataset and CiteSeerX dataset to show the accuracy and scalability of proposed approach. We compare our method with state-of-the-art technique [18] in every experiment, where the resultant dataset of previous technique is obtained from their authors. In the first experiment, two collections are randomly prepared from J.UCS dataset. The first collection is used as training set of 3,000 citations to build our approach. The second collection of 3,000 is used as testing set to further evaluate the proposed technique. The frequency distribution of in-text citations has been highlighted in J.UCS testing set as shown in Table 5.5. The results of both approaches are evaluated and compared with the manually prepared gold-standard of 3,000 in-text citations. The Table 5.5 shows the performance of both previous and proposed approaches. In Table 5.5, different abbreviations are used such as C-CIT (Correct citations), IC-CIT (Incorrect citations), ZO (Zero occurrences).

The test dataset of 3,000 citations are divided and evaluated into two sets for two different experiments. The precision, recall and F-score of set1, set2 and aggregate of both approaches are shown in Figure 5.25.

TABLE 5.5: Frequency distribution of in-text citations in J.UCS Dataset

In-Text Citation Frequency Range	Gold standard	Shahid et al			Proposed Approach		
		C-CIT	IC-CIT	ZO	C-CIT	IC-CIT	ZO
1–5	2,936	1,284	455		2,893	29	
6–10	52	14	153		49	5	
11–15	8	1	109		5	0	
16–20	4	1	78		4	0	
21–25	0	0	62		0	0	
>25	0	0	510		0	0	
Total	3,000	1,300	1,367	333	2,951	34	15

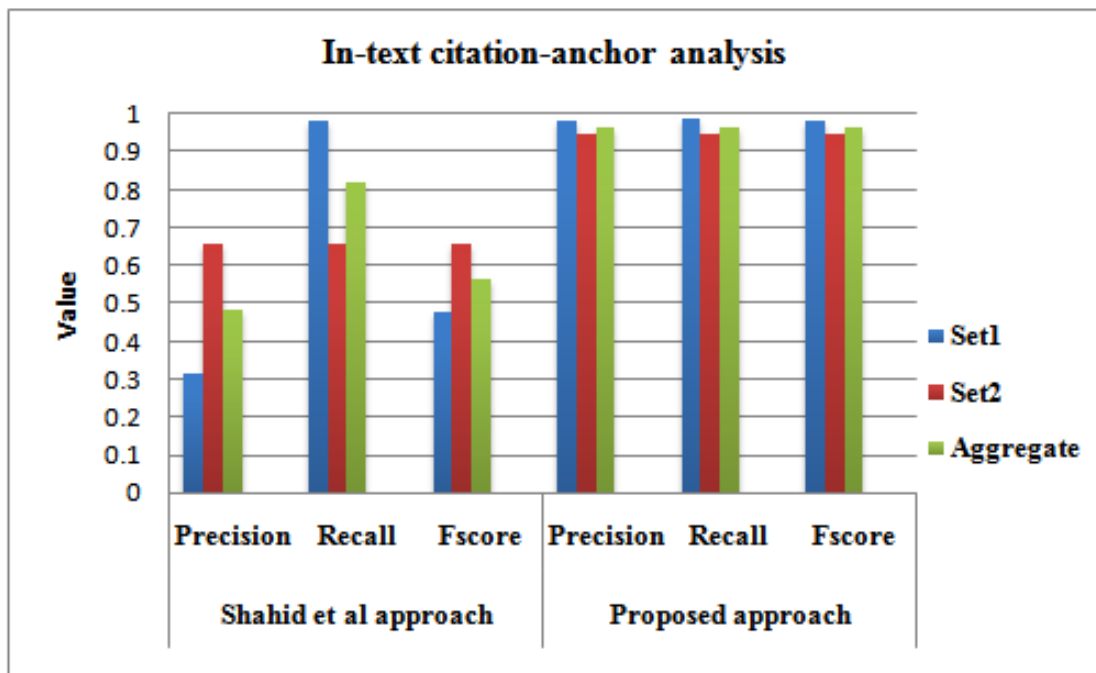


FIGURE 5.25: Precision, Recall, and F-score of both approaches over J.UCS dataset

To check the scalability of previous and proposed approaches over CiteSeerX dataset, we randomly selected 5,000 citing documents out of 20,000 citing documents dataset along with 250 reference strings (metadata) of cited documents. The dataset of 5,000 citing documents were classified into five subsets for different

TABLE 5.6: Frequency distribution of in-text citations in CiteSeerX dataset

In-Text Citation Frequency Range	Gold standard	Shahid et al			Proposed Approach		
		C-CIT	IC-CIT	ZO	C-CIT	IC-CIT	ZO
1–5	3,815	706	693		3,709	14	
6–10	150	23	373		141	11	
11–15	27	2	229		26	1	
16–20	14	1	161		14	0	
21–25	7	0	120		7	1	
>25	0	0	901		0	5	
Total	4,016	732	2,477	807	3,897	62	57

experiments. Each subset consisted of 1,000 citing documents with 50 reference strings of different cited documents. In both techniques, the in-text frequencies of each cited document are manually analyzed across its 20 citing documents. After the detailed analysis of 5,000 documents, we observed 984 documents which were not properly parsed due to image format of PDF file and due to the absence of in-text citations in citing document. From the experiments one can see that proposed approach achieves good accuracy as shown in Table 5.6. It is much more efficient than state-of-the-art approach on CiteSeerX dataset.

The Figure 5.26 shows in-text citations analysis of both approaches over 4,016 citing documents in CiteSeerX dataset. The analysis conducted over five sets of citing documents for different experiments. The aggregate precision, recall, and F-score of five experiments shows that the proposed technique is better performing than state-of-the-art technique over the CiteSeer dataset.

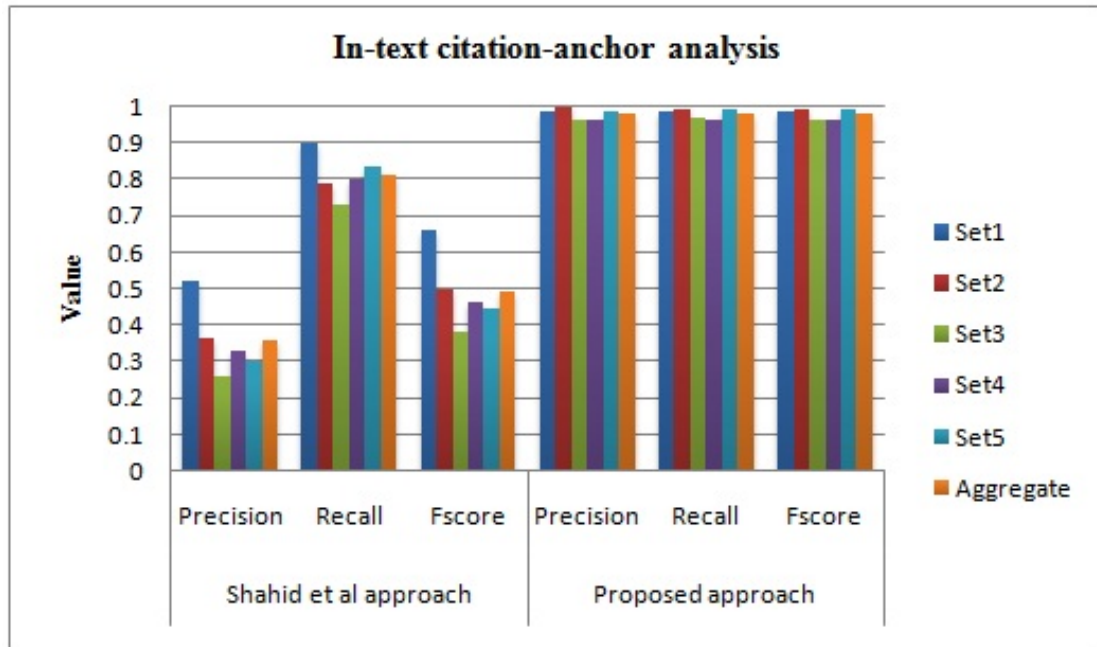


FIGURE 5.26: Precision, Recall, and F-score of both approaches over CiteSeerX dataset

Our proposed algorithm and shahid et al approach is further compared and evaluated with CERMINE and GROBID tools over the extended dataset of CiteSeer that consists of 250 cited documents and 5,008 citing documents. For the analysis, we have randomly selected 1,000 PDF files of citing documents and manually analyzed the occurrences of citation-anchors of 50 citations or cited documents to make standard dataset. The results of our algorithm, shahid et al approach, CERMINE and GROBID tools are compared with the standard dataset as shown in Figure 5.27. Measured with Fscore, our approach (0.99) is best performing than GROBID (0.91) and CERMINE(0.82).

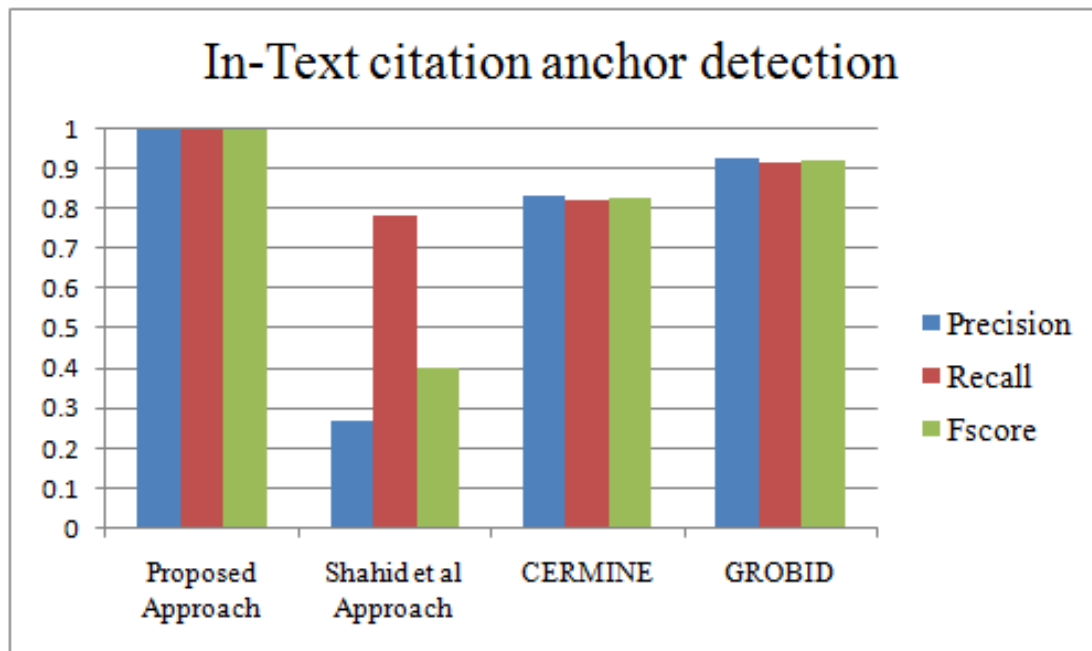


FIGURE 5.27: Comparison of Proposed approach with State-of-the-art Approach and Tools over CiteSeer Dataset

5.7 Summary

The patterns identification of in-text citation-anchor of a cited document is an important problem. Mostly the existing automatic state-of-the-art in-text citation techniques suffer due to problems related to numeric-anchors and string-anchors. The numeric-anchors problems are multiple-anchor, range-anchor and compound-anchor. While the string-anchor problems are due to their various format, hyphen with carriage return and linefeed, year related, space character, part-of-speech, reference string without citation-tag problems etc.

In this chapter, first we proposed citation-anchors taxonomy after the critical analysis of citation-anchors in the citing documents, literature approaches, and well known citation representation formats such as APA, MLA, AMA, and CBE. Secondly, we proposed, implemented and evaluated a novel approach for the identification of in-text citation patterns and frequencies in the citing documents. For the evaluation of proposed approach, two datasets were prepared from openly available J.UCS and CiteSeer sites. The testing set of J.UCS dataset consisted of 3000

citations, While the testing set of CiteSeer dataset consisted of 5000 citations. The state-of-the-art technique was also implemented over the same datasets. The results were compared with the state-of-the-art approach proposed by Shahid et al [18]. Both approaches were evaluated based on well-known measure of precision, recall and F-score. The proposed model has comprehensively outperformed the state-of-the-art approach by scoring average F-score of 0.97 as compared to baseline of 0.58. The state-of-the-art technique used the exact matching of citation-tag with citation-anchor. But the highlighted issues in section 5.2 of in-text citation anchor were not detected with exact matching. Therefore, in our approach different rules and heuristics were developed based on the proposed citation-anchors taxonomy. All these rules were used in heuristic based system as mentioned in Figure 5.20.

This thesis has proposed a new approach which is section wise co-citation analysis. To evaluate this approach, two important tasks had to be completed which becomes two important tasks of this thesis. First task was the identification of sections and mapping them on logical structural components which was successfully done in chapter 4. The second task was the accurate identification of in-text citation frequencies which has been achieved in this chapter. The proposed approach has outperformed the state-of-the-art approach by increasing the F-score from 0.58 to 0.97. In previous chapter, first the generic section identification task was completed. In this chapter, the second task was also completed with the good accuracy. This is the second contribution of our thesis. The third and last contribution will be done in chapter 6. Chapter 6 will evaluate the overall section wise co-citation proposed approach.

Chapter 6

Section Wise Co-citation Analysis

Note: The proposed work “section wise co-citation analysis” has been published in conference ¹

The section wise co-citation analysis phase as shown in Figure 6.1 consists of three main research components. The first two components (1) generic section/ILMRaD structure identification and (2) In-text co-citation patterns and frequencies identification were completed and already discussed in details in chapter 4 and chapter 5 respectively. The first and second components were developed and evaluated on different J.UCS and CiteSeer datasets and were compared with the state-of-the-art approaches. Now we are able to evaluate the proposed approach in proper manner. This research component needs parameters, such as generic sections mapping, in-text co-citation frequencies, and section weights. To evaluate the proposed approach we need the dataset that consists of co-cited document pairs and citing documents.

In section 6.1, we have shown the detailed implementation of section wise co-citation analysis (SWCA) algorithm. Section 6.3 presents the evaluation procedure

¹Ahmad, R., Afzal, M. T., (2015). Research Paper Recommendation by exploiting cocitation occurrences in Generic Sections of Scientific Papers. PhD Symposium at 13th International Conference on Frontiers of Information Technology.

of the state-of-the-art techniques over same dataset and then the rank lists of proposed approach are compared with the rank lists of the state-of-the-art techniques using rank lists based on JSD and cosine similarity as benchmarks.

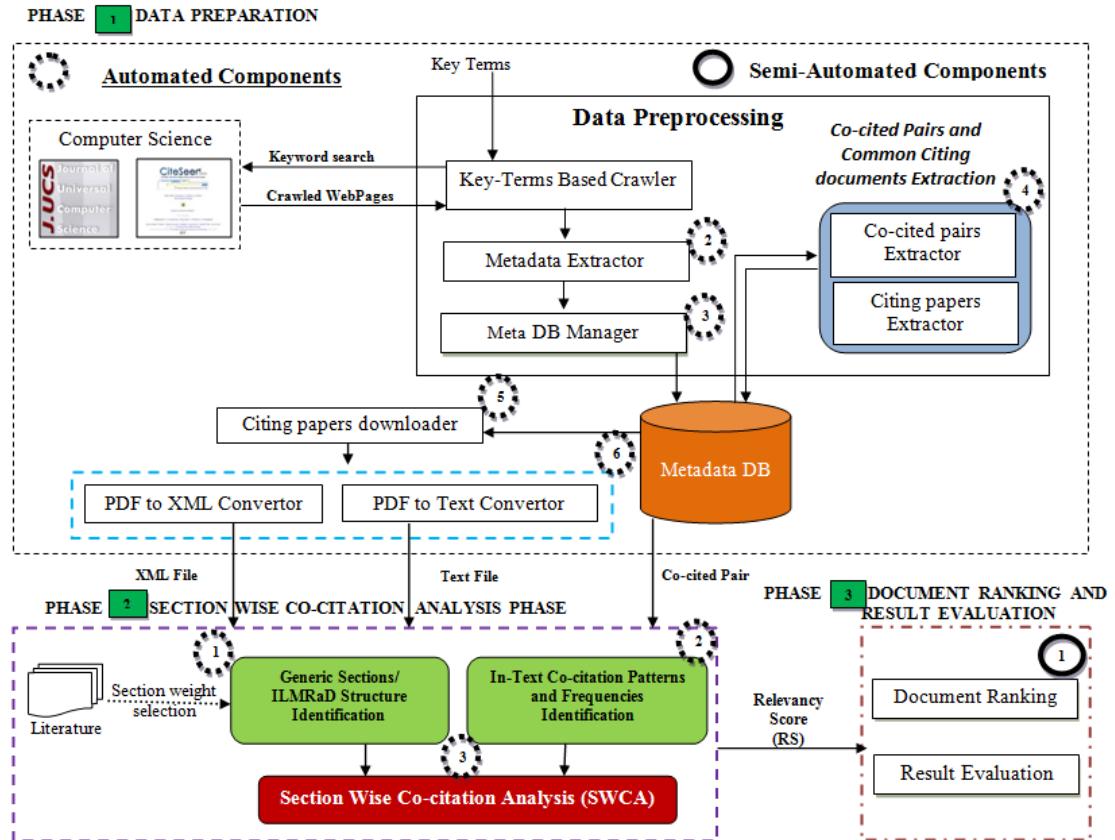


FIGURE 6.1: Proposed architecture for SWCA (Section wise co-citation analysis)

with completed contributions

6.1 SWCA Algorithm

The third and last component of our thesis is section wise co-citation analysis (SWCA) which depends on the first two components (1) Generic section or ILMRaD structure identification and (2) In-text co-citation patterns and frequencies identification. These two components were discussed in detail in chapter 4 and chapter 5. Now in this section, the SWCA algorithm has been discussed in detail to understand the section wise co-citation analysis. This SWCA algorithm consists of several steps: (1) Mapping of structural components on generic sections

(2) Citation-tag identification (3) Citation-anchor patterns and their frequencies identification, and (4) The computation of relevancy score (RS) of co-cited pair. The first step was discussed in detail in chapter 4. The second and third steps were discussed in chapter 5. The fourth step is discussed in subsection 6.1.3.

6.1.1 Dataset

To evaluate the proposed approach (SWCA), we need co-cited pairs and their citing documents. This requires research papers which have been co-cited in other citing documents. Such co-citation approach has been implemented in CiteSeerX². CiteSeerX is the scientific digital library and search engine that provides the access of the literature in the computer and information sciences domain. It has made openly available the metadata of query papers, citations, and co-cited papers which can be easily crawled. In the CiteSeerX citation graph, there are 1,345,249 citing papers and 9,150,279 citations. The total number of links in the graph, i.e. (citing paper - citation), is 25,526,384 [60]. The CiteseerX provides the ‘doi’ of research papers which can be used to download the PDF files of research papers.

In our research work, the dataset is prepared from CiteSeerX because it consists of metadata about the co-cited documents. We need three types of metadata (1) Query paper metadata (2) Co-cited papers metadata, and (3) citing documents metadata of query papers and co-cited papers. The manual preparation of such types of metadata is very difficult task.

In the first step, for the searching of required query papers, the user will enter the keyword in the CiteSeerX search engine. In response of search engine, the webpage of related query papers will be returned to user. Each webpage consists of ten links of query papers. The link of query paper contains the metadata such as “Paper title”, “Author name list”, “year”, “citations or citing documents”. The real snapshot of CiteSeerX site for the query papers is shown in Figure 6.2.

²<http://citeseerx.ist.psu.edu/index>

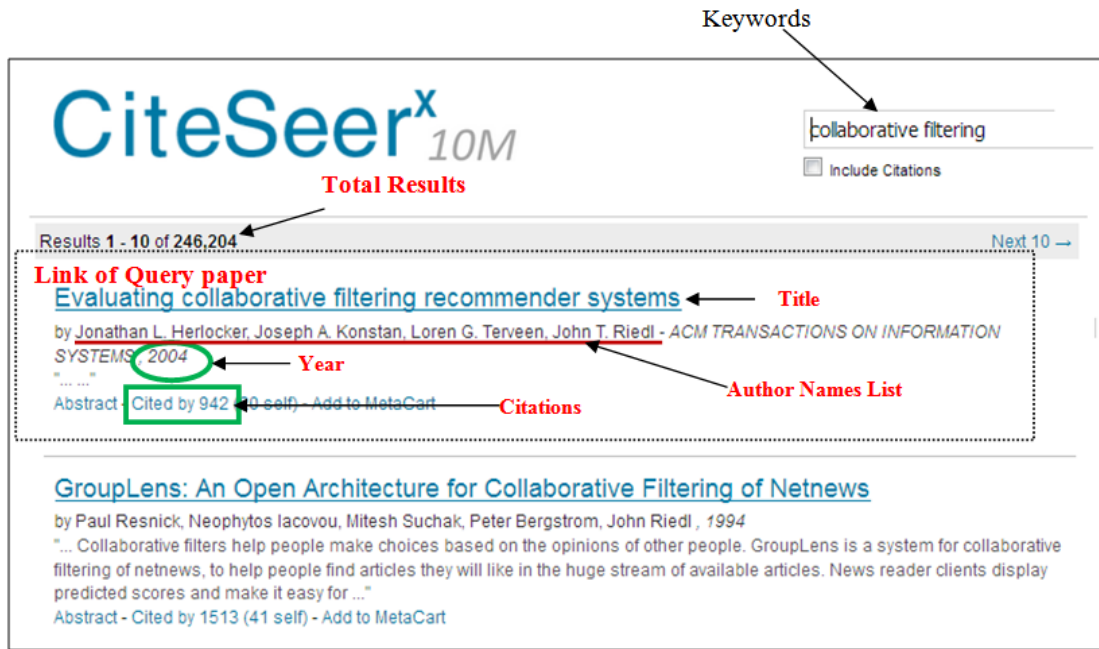


FIGURE 6.2: The real snapshot of query papers from CiteSeerX site

In the second step, after the selection of query paper, the user will need to extract the metadata such as “Title”, “Author Names”, “Number of citations”, “year”, “doi”, and “citation id (id)” of citing papers for each query paper. Let us take the example of query paper “*Evaluating Collaborative Filtering recommender systems (2004)*”. After clicking on the link ‘cited by’ for a query paper, the list of citing papers will appear. This query paper has total 928 citations as shown in Figure 6.3. In this case the user requires the metadata for all 928 citations of the query paper which is very tedious task. The ‘id’ or ‘citation id’ metadata is exploited to find the common citing documents and then the ‘doi’ metadata is used to download the common citing documents.

The screenshot shows the CiteSeerX interface. At the top, the logo 'CiteSeer^x 10M' is visible. Below it, the search results for the query 'J.T.: Evaluating collaborative filtering recommender systems (2004)' are displayed. The authors listed are J L Herlocker, J A Konstan, L G Terveen, and Riedl. The venue is 'ACM Trans. Inf. Syst'. A red box highlights the text 'Total number of citations' with an arrow pointing to the number '928' in the results list. Below the results, a snippet of the paper is shown. A red box highlights the DOI '10.1.1.107.2790' with an arrow pointing to the label 'doi'. Another red box highlights the citation ID '281050' in the HTML snippet, with an arrow pointing to the label 'Citation ID'. The snippet also shows a link to 'Add to MetaCart' and a 'showciting' link with the title 'number of citations'.

FIGURE 6.3: The real snapshot of citations of query paper from CiteSeerX site

In third step, when user clicks on the title of the query paper, the snapshot in Figure 6.5 will appear on the screen. On this screen, under the co-citation tab the list of co-cited documents will be displayed which are co-cited with the query paper based on some common citing documents. The set of co-cited pair is constructed by the query paper and number of co-cited documents. Now the question is that how can a user get the list of common citing documents. In figure 6.5, every co-cited document for each query paper have a number of co-citations such as (11807, 10581, 382, 1481, 1420, 739, 942, 1165, 270) equal to 28797 citations. Now the user needs the metadata of these co-citations for each co-cited document which become a difficult task. Let us consider if a user gets the metadata of citations of a query paper and co-cited papers. Then in the last step, he needs the metadata 'citationid' to get the common citing papers between query papers and co-cited papers as shown in Figure 6.4.

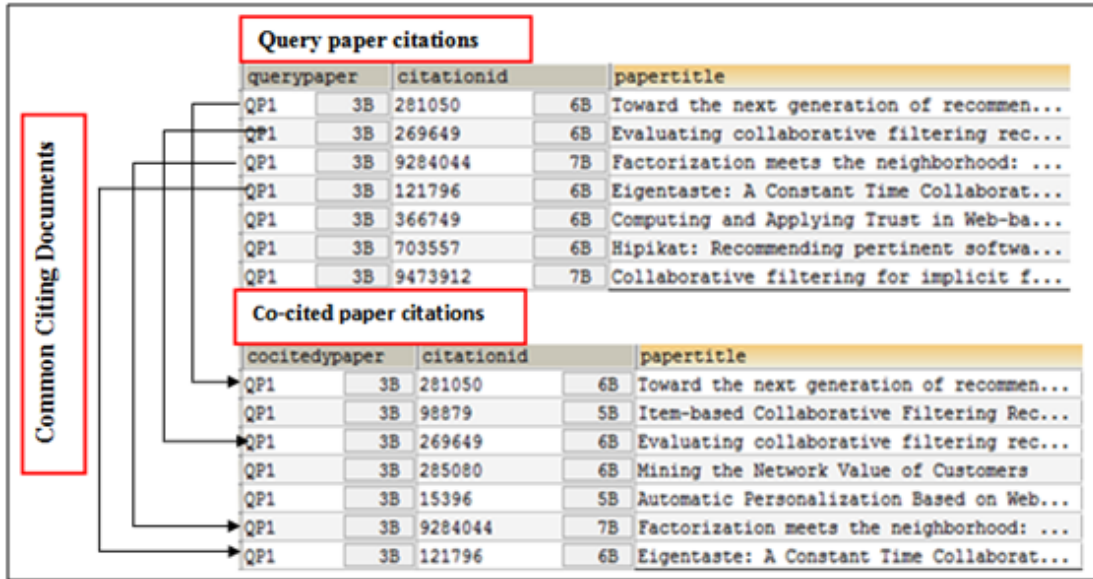


FIGURE 6.4: Visual representation of Equation 3.1

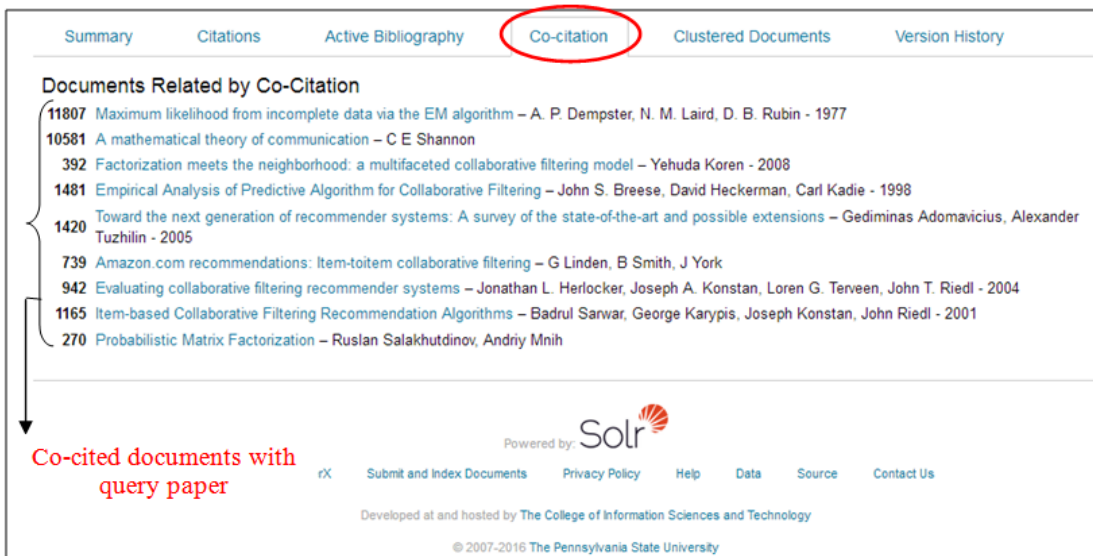


FIGURE 6.5: The real snapshot of co-cited documents with a query paper from CiteSeerX site

Manually, the whole step of data preparation is very difficult job. Hence, we have designed the data preparation phase as shown in Figure 6.1. The whole process of dataset preparation is performed automatically in this phase. For our experiment, we have selected 50 query papers and each query paper have 9 co-cited documents at CiteSeerX site. In this way, the metadata of total 450 co-cited documents are retrieved for 50 query papers. On average, 389 citations were

recorded for each query paper making a total of 19,440 citations for all query papers. Furthermore, after the intersection of query papers citations (19,440) and the co-cited papers citations (1,278,878), we have received 22,943 common citations were recorded for 450 co-cited documents. The set of 22,943 common citations are further analyzed to remove those papers based on the criteria (1) We have only considered the papers of upto 50 pages (2) Those papers are also removed which are not perfectly parsed by PDFx and PDFbox Java library (3) Those papers are also excluded which have no occurrence of co-cited pair. The total number of common citing documents were recorded for 450 pairs that are 11,875. The final dataset is prepared which consists of 50 query papers, 450 co-cited papers, and 11,875 common citations. In all these 11,875 common citations, we need to accurately extract sections and all other processing.

6.1.2 Section Weights Identification

The research papers consist of different generic sections such as “Introduction”, “Literature”, “Methodology”, “Results” and “Discussion” formally recognized as ILMRaD structure. In this research work, the three sections “Results, Discussion and Conclusions” are collectively considered “Result” section. The citations have different meaning in each generic section of scientific papers. For example, co-cited papers in “Methodology/Results” section, most probably, will be more relevant each other than the papers co-cited in the “Introduction” sections. Similarly, co-citations occurring in “Introduction” section, will be considered high relevant than the co-citations occurring in the “Literature” section. Such observations have been pointed out and recognized by different authors [6, 26]. They assigned different weights to generic sections to show their importance in research papers for different tasks. From the above discussion, the following Equation can be constructed:

$$W_{Meth}/W_{Res} > W_{Intr} > W_{Litr} \quad (6.1)$$

The ‘ W_{Meth} ’ and ‘ W_{Res} ’ shows the weights of “Methodology” and “Results” sections respectively. The ‘ W_{Intr} ’ and ‘ W_{Litr} ’ represents the weight of “Introduction”

and “Literature”. Mostly, such weight is represented within the range from (0 to 1) [21]. Boyack et al performed the co-citation proximity analysis across the full-text research documents. They have also assigned static weights of 4, 3, 2, and 1 to different level of co-citation proximities. In our case, the levels of relevance are three as co-cited in “Methodology/Results”, co-cited in “Introduction”, and co-cited in “Literature”. We have used the weights as used by Boyack et al [21] as we want to compare with them.

Motivated from this, we assigned maximum weight of 3 to papers co-cited in the “Methodology/Results” section, the weight of 2 to the co-cited in the “Introduction” section, and the weight of 1 to the papers in the “Literature” section.

6.1.3 Relevancy Score (RS) Calculation

The relevancy score (RS) of co-cited papers is calculated across generic sections of citing document by using in-text citation frequency of co-cited documents and section weights. The concept of the proposed scheme (SWCA) for ranking has been shown using a case scenario. In Table 6.1, we have taken the dataset of five papers. This dataset consists of query paper (qp), co-cited paper (ccp), and citing documents such as cd_1 , cd_2 , and cd_3 .

TABLE 6.1: Dataset of query paper,co-cited paper, and citing documents

Query paper (qp) Title:“Explaining Collaborative Filtering Recommendation”	
Co-cited paper (ccp) Title:“An algorithmic framework for performing collaborative filtering”	
Citing documents (cd)	
cd_1	Personalized recommendation of social software items based on social relations
cd_2	Providing Justifications in Recommender Systems
cd_3	Justified Recommendations based on Content and Rating Data

In Table 6.2, the pair of co-cited papers has been prepared by using the query paper (qp) and co-cited papers (ccp) as shown in Table 6.1. The table consists of one

co-cited pair such as (q_1, ccp_1) . The frequencies of co-cited pair are analyzed across the generic sections of three citing documents (cd_1, cd_2, cd_3) by using the in-text citation patterns and frequencies identification module as discussed in chapter 5.

TABLE 6.2: One co-cited pair of research papers with three citing documents

Pair of Co-cited papers		
Query paper(qp)	Co-cited papers(ccp)	Citing Document(cd)
qp ₁	ccp ₁	cd ₁
qp ₁	ccp ₁	cd ₂
qp ₁	ccp ₁	cd ₃

The frequency of co-cited pair (qp_1, ccp_1) is calculated across the generic sections of citing documents as shown in Table 6.3. 'F(qp)' represents the frequency of query paper in the generic sections of citing document (cd) while 'F(ccp)' denotes the frequency of co-cited documents in the generic sections of citing document.

The relevancy score (RS) of co-cited pair in each citing document is calculated by the Equation 6.2. 'GS' shows the number of generic sections which is four like "Introduction", "Literature", "Methodology", and "Result & Discussion". The parts 'F_{ji}(qp)' and 'F_{ji}(ccp)' are used to find the frequency of query paper and co-cited paper in 'ith' section in 'j' citing document 'cd_j' respectively. Finally the 'Min' function finds the minimum frequency among the frequencies of query paper and co-cited paper and then multiply the minimum frequency with the 'ith' section weight. The last score will show the relevancy score of query paper and co-cited paper in the 'j' citing document such as 1, 7, and 3 as shown in Table 6.3. The same process will be followed for the rest of citing documents of co-cited pair. Let us see the procedure to find the relevancy score of co-cited pair (qp_1, ccp_1) in citing document 'cd₁'. Before calculating the relevancy score, we will identify only those sections which consists of both papers 'qp₁' and 'cc₁'. For example we can see in Table 6.3, the "L = Literature" section contains (1,1) frequencies of both co-cited papers. Now in this case, the minimum frequency 1 will be picked for further processing. The section weight of Literature section is 1. This weight

will be multiplied with the in-text citation frequency of concerned co-cited pair like ($1 * 1 = 1$).

TABLE 6.3: Co-citation frequencies and relevancy score (RS)

Pair set			F(qp)				F(ccp)				Relevancy Score (RS)
qp	ccp	cd	I	L	M	RaD	I	L	M	RaD	
1	1	1	0	1	0	1	2	1	0	0	1
1	1	2	3	2	1	0	2	0	1	0	7
1	1	3	2	3	0	0	1	0	0	0	2
Commulative Relevancy score (CRS):											10

Let us see another scenario in which the co-cited pair is cited in more than one sections, such as ‘Introduction’ and ‘Methodology’ sections. The frequencies of co-cited pair in the ‘Introduction’ and ‘Methodology’ sections of ‘cd₂’ citing document are $I = (3, 2)$ and $M = (1, 1)$ respectively. The minimum frequency of co-cited pair in ‘Introduction’ section is 2 and the minimum frequency of co-cited pair in ‘Methodology’ section is 1. Now the Relevancy score of co-cited papers in ‘cd₂’ is ($I = (3, 2) = 2 \times 2 = 4$) and ($M = (1, 1) = 1 \times 3 = 3$). The score of co-cited papers in “Introduction” and “Methodology” sections is 4 and 3 in citing document ‘cd₂’ respectively. The total relevancy score of co-cited papers in ‘cd₂’ is 7.

$$RS(qp, ccp_x, cd_j) = \sum_{i=1}^{GS} \text{Min}[F_{ji}(qp), F_{ji}(ccp_x)] \times w_i \quad (6.2)$$

The cumulative relevancy score (CRS) of co-cited pair can be computed by using the Equation 6.3. In this Equation ‘cd’ shows the number of citing documents. The relevancy score of co-cited pair is computed against each citing document ‘cd_j’ by using Equation 6.2. The resultant relevancy score is combined to get the cumulative relevancy score 10 as shown in Table 6.3.

$$CRS(qp, ccp_x) = \sum_{j=1}^{cd=N} RS(qp, ccp_x, cd_j) \quad (6.3)$$

In the end of this computation process, we have obtained the final Equation to find the CRS of co-cited pair against ‘N’ citing documents by combining the Equations 6.2, 6.3.

$$CRS(qp, ccp_x) = \sum_{j=1}^{cd=N} \sum_{i=1}^{gs} \text{Min}[F_{ji}(qp), F_{ji}(ccp_x)] \times w_i \quad (6.4)$$

By using Equation 6.4, the cumulative relevancy score of all co-cited documents (ccp_x) with the query paper (qp) are computed across the ‘N’ citing documents. In our experiment, the 9 co-cited documents are selected for a single query paper. Each co-cited pair is analyzed against ‘N’ citing documents. The result of a co-cited pair against ‘N’ citing documents has been shown in Table 6.4.

TABLE 6.4: The Cumulative relevancy score of nine co-cited pairs

Query paper(qp)	Co-cited Papers	Cumulative Relevancy Score (CRS)
qp ₁	1	0
qp ₁	2	0
qp ₁	3	4
qp ₁	4	12
qp ₁	5	21
qp ₁	6	19
qp ₁	7	16
qp ₁	8	27
qp ₁	9	15

6.1.4 Document Ranking

Subsequently the documents are ranked based on the cumulative relevancy score of co-cited pairs as highlighted in Table 6.4. The papers with highest cumulative relevancy score will come on the top of the ranked list. The list of the ‘ccp’ co-cited papers are ranked for the query paper ‘qp’ based on the cumulative relevancy score as shown in Table 6.5. The new rank under the ‘Rank ID’ is generated by the

proposed approach (SWCA). In the next evaluation section, this proposed rank will be compared with state-of-the-art techniques.

TABLE 6.5: The cumulative relevancy score of nine co-cited pairs

Paper Reference ID	Order CRS in Descending	Rank ID
1	27	8
2	21	5
3	19	6
4	16	7
5	15	9
6	12	4
7	4	3
8	0	1
9	0	2

6.1.5 Pseudo code for SWCA algorithm

The pseudo code for the SWCA algorithm is given below. The details of Rule-based Algorithm is given in the end of section [4.2.4](#)

SWCA Algorithm (Co-cited-pairs-Metadatas, Citing-documents)

Co-cited-pairs \rightarrow (qp, ccp_x) x = {1,2,3,...N}

[Query and co-cited papers Metadata (First author, Year, title)]

Citing-documents \rightarrow (cd_N) cd = {1,2,3,...N}

CRS \leftarrow 0

qptag \leftarrow \emptyset

ccptag \leftarrow \emptyset

For cd_j := 1 **To** N **Then** // cd_j:th Citing document

 Rule_Based_Algorithm(cd_j) // gs: Generic sections in section [4.2.4](#)

 gs \leftarrow getGeneric_Section(cd_j)

```

qptag ← getCitationTag(qp.firstauthor, qp.year, qp.title, cdj);
ccptag ← getCitationTag(ccp.firstauthor, ccp. year, ccp.title, cdj);
For i ← 1 To GS := 4
  CAD (qptag, GS[i]) //CAD in section 5.5.3
  CAD (ccptag, GS[i])
  CRS ← CRS + Relevancy_Score(cdj) //cumulative relevancy score
End For Loop
End For Loop
  StoredCRS (CRS)

```

Relevancy_Score (cd)

```

qp_ccp_fr_int ← 0,    qp_ccp_fr_litr ← 0
qp_ccp_fr_met ← 0,    qp_ccp_fr_rd ← 0
qp_ccp_fr_int ← getFrequency(1, cd) // frequency in introduction section
qp_ccp_fr_litr ← getFrequency(2, cd) // frequency in Literature section
qp_ccp_fr_met ← getFrequency(3, cd) // frequency in Methodology section
qp_ccp_fr_rd ← getFrequency(4, cd) // frequency in Result & Discussion section
If qp_ccp_fr_int (0) != 0 && qp_ccp_fr_int (1) != 0 then
  Int_rs := MIN(qp_ccp_fr_int(0), qp_ccp_fr_int(1)) × 2 // weight = 2
End if
If qp_ccp_fr_litr (0) != 0 && qp_ccp_fr_litr (1) != 0 then
  Lit_rs := MIN(qp_ccp_fr_litr(0), qp_ccp_fr_litr(1)) × 1 //weight = 1
End if
If qp_ccp_fr_met (0) != 0 && qp_ccp_fr_met (1) != 0 then //weight = 3
  Met_rs := MIN(qp_ccp_fr_met(0), qp_ccp_fr_met(1)) × 3
End if
If qp_ccp_fr_rd (0) != 0 && qp_ccp_fr_rd (1) != 0 then
  Rd_rs := MIN(qp_ccp_fr_rd(0), qp_ccp_fr_rd(1)) × 3 //weight = 3
End if
RS = Int_rs + Lit_rs + Met_rs + Rd_rs // Relevancy Score

```

6.2 Evaluation

This section presents detailed evaluation of the proposed approach. The first two tasks were evaluated in chapter 4 and 5. For evaluation of SWCA algorithm, we have utilized co-cited pairs dataset from CiteSeerX and have been performed both of the first two mentioned tasks again. Therefore, its evaluation has also been performed in section 6.2.1 and 6.2.2.

6.2.1 Evaluation of generic section identification

In this thesis, an approach was proposed, implemented and evaluated in chapter 4 for mapping of section headings onto logical sections of research papers. However, in this chapter, a new dataset was constructed based on co-citation pairs. Therefore, it becomes important to re-evaluate the working of the section mapping approach on this new dataset. For the evaluation of generic sections identification, total 150 citing documents have selected from the new CiteSeer dataset. The total number of structural components in 150 citing documents are 1,049 that have been extracted by using our proposed approach architecture as discussed in details in chapter 4. The confusion matrix for the section identification is shown in Table 6.6.

TABLE 6.6: Confusion matrix for generic sections identification over 150 papers

Predicted as	INTR	LITR	MET	RES	DISC	CON
Introduction	150	0	0	0	0	0
Literature	0	25	4	2	0	0
Methodology	0	10	339	21	5	0
Results	0	5	22	276	5	0
Discussions	0	0	0	0	41	0
Conclusions	0	0	0	0	0	144

The precision, recall, and F-score of generic section mapping is shown in Figure 6.6. The F-score of our first component over the new set of CiteSeer dataset is 0.90 while the previous F-score of this approach was 0.92 as discussed in chapter 4.2.5.

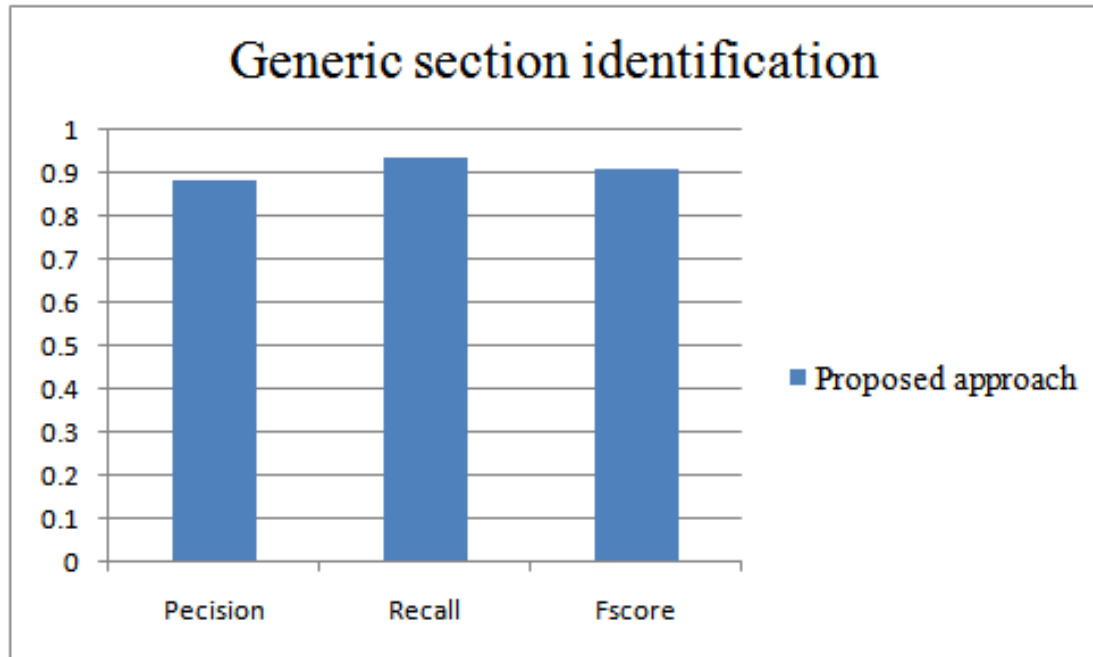


FIGURE 6.6: Precision, Recall, and F-score of generic section Identification over CiteSeer dataset

6.2.2 Evaluation of In-text citation frequency Identification

In this thesis, In-text citation identification approach was proposed, implemented and evaluated in Chapter 5. In this section we have re-evaluated this approach over the new dataset of CiteSeer papers. The evaluation of in-text citation frequency identification has been done over the randomly selected 200 citing documents. In these citing document, the citation frequency of 400 co-cited documents are manually analyzed in text of citing documents. After the manual analysis of in-text citation frequencies of cited documents, the precision, recall, and F-score were calculated. The results are shown in Figure 6.7. F-score of our second component over the new set of CiteSeer dataset is 0.89 while the previous F-score of our approach was 0.97 as discussed in section 5.6.3.

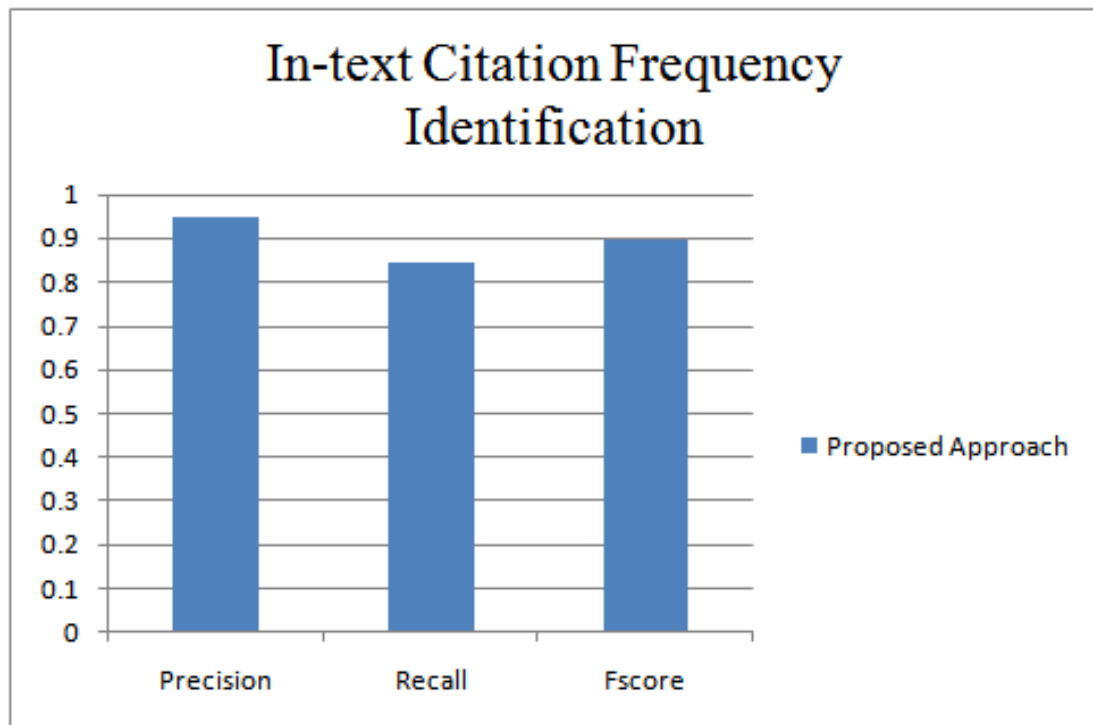


FIGURE 6.7: Precision, Recall, and F-score of In-text citation frequency Identification over CiteSeer dataset

6.3 Evaluation and comparison of SWCA approach with State-of-the-art approaches

In this section, we are evaluating and comparing the results of proposed approach (SWCA) with state-of-the-art approaches including standard co-citation technique [32] and citation-proximity analysis (Boyack et al) technique [21, 22, 37]. The proposed approach and state-of-the-art approaches will produce ranked list of relevant papers for each query paper. Now the problem is to evaluate and compare the proposed and state-of-the-art techniques against some benchmark data and/or method. Beel et al have evaluated the research paper recommendation approaches published in the last 15 years [9], after thorough analysis, they have concluded that there is no standard dataset (Gold Standard dataset) on which a system can be evaluated in this domain. They highlighted that the gold standard dataset was prepared by different researchers for the evaluation of their systems and such a

gold standard was not made available openly by any of the researcher which can be further utilized by others working in this domain. However, the strategies for making gold standard were reviewed by Beel et al [9] and they concluded that there are two types of evaluation methods (1) Online evaluation and (2) Offline evaluation. They also highlighted that the offline evaluation method has been used in evaluation of 53% paper recommendation systems. The offline evaluation is conducted in two ways: (1) user study and (2) offline evaluation metrics. The user study is not possible for the huge dataset. Therefore, the evaluation of the huge dataset is performed by different offline evaluation metrics such as recall, F-measure, mean reciprocal rank (MRR), normalized discounted cumulative gain (nDCG), mean absolute error, root mean square error and considering benchmark ranking made by content of research papers. According to Beel et al [9], 53% approaches were compared with content-based filtering. The content of documents are strong evidences for similarity purpose. Some of the recent studies [21, 90–93] have shown the importance of new version of kullback leibler divergence (KLD) which is called JSD (Jensen Shannon Divergence). They considered JSD as good measure for the difference or divergence between two distribution or ranking. We have selected two content-based measures including JSD and cosine Similarity [94–96] as Baselines or Standards. Therefore, first the JSD measure has been explained in section 6.3.1 and then we have discussed the content based similarity measure in section 6.3.2. Finally the co-citation, Boyack et al and the proposed approach will be compared with JSD and cosine similarity.

6.3.1 Jensen-Shannon Divergence (JSD)

The JSD measure is used to compute the distance between two probability distributions [90]. First the word probability vector for each document is prepared and then the word probability vector is prepared for the cluster that consists of all documents. The JSD value is calculated for each document by using the word probability vector for a document and the word probability vector for the cluster

in which the document resides. The JSD formula is shown in Equation 6.5.

$$JSD(p, q) = \frac{1}{2}D_{KL}(p, m) + \frac{1}{2}D_{KL}(q, m) \quad (6.5)$$

In Equation 6.5, ‘p’ is the probability of a word in a document and ‘q’ is the probability of the same word in the cluster of documents. D_{KL} is the Kullback-Leibler divergence as shown in Equation 6.6. ‘N’ is the number of words in a cluster of documents

$$D_{KL}(p, m) = \sum_{i=1}^N (p_i \log(p_i/m_i)) \quad (6.6)$$

The cluster JSD is calculated as the average of JSD values for all documents in the cluster. JSD is a divergence measure, meaning that if the documents in a cluster are very different from each other, using different sets of words, the JSD value will be very high. Clusters of documents with similar sets of words (a less diverse set of words) will have a lower divergence. The steps for JSD value calculation of a document and a cluster is shown below with proper example.

In first step, we will take the set of documents which consists of different keywords as shown in Figure 6.7.

TABLE 6.7: Cluster of documents

Doc#	Number of words in three different documents in a cluster	Number of Words
Doc1	cross, Validated, answers, computer, good	5
Doc2	simply, validated, answers, computer, nice	5
Doc3	simply, cross, bye, hello, good, cross	6
Total Number of Words in a Cluster:		16

In second step, we will prepare the word probability vectors across each document and a cluster as shown in table 6.8.

TABLE 6.8: Word count and probability vectors for each document and cluster

Words	Word count vectors				Word probability vectors			
	doc1	doc2	doc3	cluster	doc1 (p1)	doc2 (p2)	doc3 (p3)	cluster (q)
answers	1	1	0	2	0.2	0.2	0	0.125
computer	1	1	0	2	0.2	0.2	0	0.125
cross	1	0	2	3	0.2	0	0.33	0.187
good	1	0	1	2	0.2	0	0.16	0.125
nice	0	1	0	1	0	0.2	0	0.062
simply	0	1	1	2	0	0.2	0.16	0.125
validated	1	1	0	2	0.2	0.2	0	0.125
bye	0	0	1	1	0	0	0.16	0.062
hello	0	0	1	1	0	0	0.16	0.062

In third step, we will find the mean distribution using $m = (p + q)/2$. Let us suppose, we want to find the ‘m’ of ‘answers’ word in doc1, then we will get the mean of probability values of ‘answers’ word in doc1 and its cluster such as ‘0.2 + 0.125’. In this way, the ‘m’ values can be calculated for the other words in a cluster as shown in Table 6.9.

TABLE 6.9: Mean of ‘p1’, ‘p2’, and ‘p3’ with ‘q’ distribution

Words	$m1 = (p1 + q1)/2$	$m2 = (p2 + q2)/2$	$m3 = (p3 + q3)/2$
answers	0.1625	0.1625	0.0625
computer	0.1625	0.1625	0.0625
cross	0.1937	0.0937	0.2604
good	0.1625	0.0625	0.1458
nice	0.0312	0.1312	0.0312
simply	0.0625	0.1625	0.1458
validated	0.1625	0.1625	0.0625
bye	0.0312	0.0312	0.1145
hello	0.0312	0.0312	0.1145

Now we are able to find the Kullback Leibler (KL) Divergence by using ‘p’ and ‘m’ for particular word. The ‘ D_{KL} ’ values for each word are calculated by Equation 6.6 as shown in Table 6.10.

TABLE 6.10: Kullback Leibler Divergence for 'p' and 'q'

Words	(p1,m1)	(p2,m2)	(p3,m3)	(q1,m1)	(q2,m1)	(q3,m3)
answers	0.04152	0.04152	0	-0.03279	-0.0327	0.0866
computer	0.04152	0.04152	0	-0.03279	-0.03279	0.0866
cross	0.0063	0	0.0822	-0.0061	0.1299	-0.0615
good	0.0415	0	0.0222	-0.0327	0.0866	-0.0192
nice	0	0.0842	0	0.0433	-0.0463	0.0433
simply	0	0.0415	0.0222	0.0866	-0.0327	-0.0192
validated	0.04152	0.04152	0	-0.0327	-0.0327	0.0866
bye	0	0	0.0624	0.0433	0.0433	-0.0378
hello	0	0	0.0624	0.0433	0.0433	-0.0378
$\sum_{i=1}^N D_{KL}(p/q, m)$	0.1724	0.2503	0.2516	0.0792	0.1256	0.1273

Now the JSD values of each document in cluster can be calculated by using Kullback Leibler divergence as mentioned below.

$$\begin{aligned}
 JSD(doc1) &= \frac{D_{KL}(p1, m1) + D_{KL}(q1, m1)}{2} = \frac{0.1724 + 0.0792}{2} = \mathbf{0.1258} \\
 JSD(doc2) &= \frac{D_{KL}(p1, m1) + D_{KL}(q1, m1)}{2} = \frac{0.2503 + 0.1256}{2} = \mathbf{0.1880} \\
 JSD(doc3) &= \frac{D_{KL}(p1, m1) + D_{KL}(q1, m1)}{2} = \frac{0.2516 + 0.1273}{2} = \mathbf{0.1895} \\
 JSD(Cluster) &= \frac{JSD(doc1) + JSD(doc2) + JSD(doc3)}{3} = \mathbf{0.1678}
 \end{aligned} \tag{6.7}$$

The divergence or difference between a cluster and a document is calculated by using Equation 6.8. The low divergence value of a document shows more relevancy with a cluster. The divergence values of different documents in a cluster are used to make benchmarks ranking.

$$Divergence(document) = JSD(Cluster) - JSD(document)$$

e.g

$$Divergence(doc1) = Abs(0.1678 - 0.1258) = 0.0419 \tag{6.8}$$

$$Divergence(doc2) = Abs(0.1678 - 0.1880) = 0.0202$$

$$Divergence(doc3) = Abs(0.1678 - 0.1895) = 0.0217$$

Now, it is time to make the benchmark using JSD measure for comparison of proposed approach and state-of-the-art-approaches. First we have randomly selected ten clusters of documents. Each cluster consisted of nine documents. Before calculating JSD of each document and JSD of cluster, we removed the stopwords and special symbols from documents. Then we have obtained automatically the document JSD and cluster JSD for different clusters of co-cited papers. Finally, the divergence values have been found for different documents in their respective cluster. These divergence values are used to rank the documents in a particular cluster. The ten ranking are prepared based on JSD values as shown in Table 6.11 which will be used as benchmark in the comparison of section wise co-citation analysis and state-of-the-art approaches. Each rank list is prepared on different set of documents and represented by the unique column in Table 6.11. The values of JSD measure were unique for each rank list hence no tie condition occurred in JSD values among each cluster of documents.

TABLE 6.11: Ten rankings prepared for ten clusters of documents based on Divergence measure

Paper#	Rank1	Rank2	Rank3	Rank4	Rank5	Rank6	Rank7	Rank8	Rank9	Rank10
1	1	2	7	1	2	3	2	3	2	1
2	3	5	2	2	5	6	5	2	1	9
3	4	1	3	5	6	1	7	5	4	4
4	2	9	4	6	1	8	4	1	6	7
5	5	6	9	7	7	9	9	6	7	2
6	7	3	8	4	3	7	6	7	8	6
7	6	4	6	8	4	4	1	4	5	3
8	9	8	1	9	9	5	8	8	9	5
9	8	7	5	3	8	2	3	9	3	8

6.3.2 Content based Similarity

Another important state-of-the-art approach with which we will be comparing our results, is content based similarity. To implement the content based similarity, different measures are used that include Cosine Similarity [97], Jaccard [98], Euclidean [99] etc. Cosine similarity is used to measure the distance between two

vectors. Subhashini and Kumar [100] conducted the experimental study of similarity measures for both information retrieval and document clustering. They indicated that the cosine similarity measure is superior than the other measures such as Jaccard measure, Euclidean, and Pearson correlation distance. It is used to find the ranking of documents [97]. The cosine similarity measure formula is given in Equation 6.9. If the value of cosine similarity function is zero between that two documents then it means the two documents are not related with each other. If the value of cosine similarity function is one then it means the two documents are identical.

$$\text{Similarity} = \text{Cos}(\theta) = \frac{q \cdot d}{\|q\| \cdot \|d\|} = \frac{\sum_{i=1}^m q_i d_i}{\sqrt{\sum_{i=1}^m q_i^2} \sqrt{\sum_{i=1}^m d_i^2}} \quad (6.9)$$

Let us see the example of similarity between text documents by using the cosine similarity measure. First we selected the dataset of three document as mentioned in Table 6.12.

TABLE 6.12: Collection of text documents

Doc#	Number of words in three different documents in a cluster
Doc1	cross, Validated, answers, computer, good
Doc2	simply, validated, answers, computer, nice
Doc3	simply, cross, validated, good, cross

Before to performing any task in information retrieval over text document, the Term Frequency Vector (TFV) of content is prepared as shown in Table 6.13.

TABLE 6.13: Document TFV with tf-idf score

Terms	tf_{t,d_1}	tf_{t,d_2}	tf_{t,d_3}	idf	$d_1(\text{tf-idf})$	$d_2(\text{tf-idf})$	$d_3(\text{tf-idf})$
answers	1	1	0	0.176	0.176	0.176	0
computer	1	1	0	0.176	0.176	0.176	0
cross	1	0	2	0.176	0.176	0	0.299
good	1	0	1	0.176	0.176	0	0.176
nice	0	1	0	0.477	0	0.477	0
simply	0	1	1	0.176	0	0.176	0.176
validated	1	1	1	0	0	0	0

The weight of each term in a document of corpus is denoted by the ‘tf-idf’ measure. The ‘tf-idf’ is shown in Equation 6.10. The ‘ $tf_{t,d}$ ’ shows the term frequency in a particular document. The ‘idf’ represents the inverse document frequency calculated by the ‘ $\log_{10}(N/df_t)$ ’. ‘N’ represents the total number of documents in a corpus and ‘ df_t ’ shows the number of documents that consist of the term ‘t’.

$$W_{t,d} \text{ or } tf - idf = (1 + \log_{10}tf_{t,d}) \times \log_{10}\left(\frac{N}{df_t}\right) \quad (6.10)$$

Now, the cosine similarity between any two documents can be calculated by using the Equation 6.9. The ‘tf-idf’ of each term in ‘ d_1 ’, ‘ d_2 ’, and ‘ d_3 ’ in Table 6.8 will be used to find the cosine similarity between documents.

Let us find the cosine similarity score between the pairs of documents such as (d_1, d_2) , (d_1, d_3) , and (d_2, d_3) by putting the values of ‘ d_1 ’, ‘ d_2 ’, and ‘ d_3 ’ from Table 6.14 in Equation 6.9.

TABLE 6.14: Terms with 'tf-idf' scores in d_1 , d_2 , and d_3

Terms	d_1	d_2	d_3	d_1^2	d_2^2	d_3^2
answers	0.176	0.176	0	0.031	0.031	0
computer	0.176	0.176	0	0.031	0.031	0
cross	0.176	0	0.299	0.031	0	0.052
good	0.176	0	0.176	0.031	0	0.031
nice	0	0.477	0	0	0.228	0
simply	0	0.176	0.176	0	0.031	0.031
validated	0	0	0	0	0	0

The cosine similarity score is high between ' d_1 ' and ' d_3 ' as shown below which means that ' d_1 ' and ' d_3 ' are more relevant documents.

$$\begin{aligned} \text{CosineSim}(d_1, d_2) &= \frac{0.176 \times 0.176 + 0.176 \times 0.176 + 0.176 \times 0}{\sqrt{0.031 + 0.031 + 0.031 + 0.031 + 0 + 0 + 0}} \\ &\quad + \frac{0.176 \times 0 + 0 \times 0.477 + 0 \times 0.176 + 0 \times 0}{\sqrt{0.031 + 0.031 + 0 + 0 + 0.228 + 0.031 + 0}} = \mathbf{0.311} \end{aligned}$$

$$\begin{aligned} \text{CosineSim}(d_1, d_3) &= \frac{0.176 \times 0 + 0.176 \times 0 + 0.176 \times 0.299}{\sqrt{0.031 + 0.031 + 0.031 + 0.031 + 0 + 0 + 0}} \\ &\quad + \frac{0.176 \times 0.176 + 0 \times 0 + 0 \times 0.176 + 0 \times 0}{\sqrt{0 + 0 + 0.052 + 0.031 + 0 + 0.031 + 0}} = \mathbf{0.597} \end{aligned}$$

$$\begin{aligned} \text{CosineSim}(d_2, d_3) &= \frac{0.176 \times 0 + 0.176 \times 0 + 0 \times 0.299}{\sqrt{0.031 + 0.031 + 0 + 0 + 0.228 + 0.031 + 0}} \\ &\quad + \frac{0 \times 0.176 + 0.477 \times 0 + 0.176 \times 0.176 + 0 \times 0}{\sqrt{0 + 0 + 0.052 + 0.031 + 0 + 0.031 + 0}} = \mathbf{0.162} \end{aligned}$$

Similarly to JSD, the cosine similarity scores has been found among the documents of same ten clusters. In this way, the ten rankings are prepared based on Cosine

similarity values as shown in Table 6.15 which will be compared with the proposed approach.

TABLE 6.15: Ten rankings prepared for ten clusters based on cosine similarity score

Paper#	Rank1	Rank2	Rank3	Rank4	Rank5	Rank6	Rank7	Rank8	Rank9	Rank10
1	1	2	7	2	2	5	2	2	7	2
2	2	8	3	1	4	4	5	1	2	9
3	3	1	2	4	6	1	8	5	5	5
4	4	9	9	9	3	7	6	3	4	6
5	8	6	8	8	7	3	9	4	1	3
6	7	3	6	3	1	9	7	9	3	7
7	5	5	5	6	5	8	1	6	9	1
8	9	7	1	7	8	6	4	7	8	4
9	6	4	4	5	9	2	3	8	6	8

6.3.3 Co-citation Technique

We are going to compare different state-of-the-art approaches with the proposed approach on the same dataset. In this context, co-citation approach proposed by small [32] becomes one of the right choice for comparison as it is considered a benchmark by scientific community to compare their own approaches [21, 90]. The co-citation is a relationship which is established between cited documents by the authors of citing documents. The degree of relationship between co-cited documents is measured by the number of times they appear together in citing documents. This co-citation measure is also used to rank the co-cited documents with the query paper. Those co-cited documents which have the highest co-citation with the query paper will come at the top of ranking list. In our research work, we have also selected the co-citation approach for the comparison of our approach. For the comparison, the ten rank lists of co-cited documents are prepared based on the co-citation measure over the same document clusters which is utilized for the JSD calculation. The ten rank list are shown in Table 6.16.

TABLE 6.16: Ten ranking prepared for ten cluster of documents based on Co-citation measure

Paper#	Rank1	Rank2	Rank3	Rank4	Rank5	Rank6	Rank7	Rank8	Rank9	Rank10
1	1	1	2	2	1	1	1	1	1	1
2	2	9	1	1	6	6	4	2	9	5
3	5	7	3	9	2	3	3	7	7	3
4	6	6	9	4	9	7	2	6	4	9
5	3	8	8	7	4	9	5	9	3	2
6	7	5	7	5	7	8	6	8	5	6
7	8	4	4	6	3	4	8	4	6	4
8	9	3	6	8	8	5	7	5	8	8
9	4	2	5	3	5	2	9	3	2	7

6.3.4 Citation Proximity Analysis (Boyack et al)

In the citation proximity analysis, Boyack et al [21] performed ranking on co-cited documents based on the proximity measure in text of citing document. They considered the whole document as a set of bytes. The citations that are within the same bracket such as ‘[22,3 ,4,55]’, a weight of 4 is assigned, while co-cited pairs within 375, 1500, and 6000 bytes are given weights of, 3, 2, and 1, respectively. The co-cited pairs that are more than 6000 bytes apart are given a weight of zero. In this approach, there is no need to get the sections of citing document. Based on this proximity analysis the following ranking lists are prepared for the evaluation and comparison with proposed approach (SWCA) in results section. The ranking lists are given in Table 6.17. Each rank in column wise order represents the ranking for each cluster of co-cited documents.

TABLE 6.17: Ten rankings prepared for ten clusters of documents based on Proximity measure

Paper#	Rank1	Rank2	Rank3	Rank4	Rank5	Rank6	Rank7	Rank8	Rank9	Rank10
1	1	1	1	1	1	1	1	1	1	1
2	2	7	2	2	5	9	9	2	9	6
3	4	3	3	9	7	5	7	7	6	5
4	9	9	9	4	2	6	6	4	3	9
5	3	6	8	6	6	4	8	9	4	2
6	7	4	7	5	8	7	4	5	7	4
7	6	8	6	3	4	8	3	6	5	3
8	5	5	4	8	9	2	5	8	8	8
9	8	2	5	7	3	3	2	3	2	7

6.3.5 Section Wise Co-citation Analysis(SWCA)

The details of section wise co-citation analysis have been given in section 6.1. In this section, the ten ranking lists are prepared after executing SWCA approach. These ranking will be compared in the results section with the state-of-the-art approaches including JSD ranking and cosine similarity. The ranking lists are shown in Table 6.18.

TABLE 6.18: Ten rankings prepared for ten clusters based on Relevancy Score in SWCA approach

Paper#	Rank1	Rank2	Rank3	Rank4	Rank5	Rank6	Rank7	Rank8	Rank9	Rank10
1	1	1	1	2	1	1	1	1	1	1
2	2	9	2	1	5	5	5	2	9	5
3	3	2	3	5	6	3	8	6	6	3
4	4	8	7	8	3	8	4	5	3	9
5	8	6	9	7	7	6	7	9	5	2
6	7	5	8	4	2	7	6	7	7	4
7	6	4	6	6	4	9	3	4	4	7
8	9	7	4	9	8	4	9	8	8	8
9	5	3	5	3	9	2	2	3	2	6

6.3.6 Results

In this section, we have compared the ranking of SWCA approach and state-of-the-art approaches such as co-citation, and citation proximity analysis(Boyack et al) with JSD and cosine similarity based ranking to find the correlation. The JSD and Cosine similarity rankings are used as Baseline [94, 101]. The correlation is the distribution analysis that is used to measures the strengths of association between two distribution and the direction of the relationship. Usually the score of co-relation occurred between +1 and -1. The value of '+1' means perfect positive correlation and the value of '-1' will be perfect negative correlation. Similarly the value of '0' means "no correlation". The two rank correlation measures: Spearman's (p) and Kendall's (T) are selected to evaluate the correlation between JSD or cosine similarity and proposed approach as well as state-of-the-art approaches. These two measures are widely used for evaluating the rankings [102, 103]

(1) Spearman's Rank Correlation Coefficient

Let us see the comparison between the ranking of proposed and state-of-the-art approaches against the benchmark ranking of JSD using Spearman rank correlation measure. The formula of Spearman rank correlation is given in Equation 6.11. ' p ' is the spearman rank correlation. ' d_i ' represents the difference between the ranks of corresponding values ' X_i ' and ' Y_i '. The ' n ' denotes the number of values in each dataset.

$$p = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (6.11)$$

For single randomly selected cluster of co-cited documents, the rank lists of proposed approach ,state-of-the-art approaches, and JSD are shown in Table 6.19.

TABLE 6.19: The ranking dataset of single cluster for proposed approach, state-of-the-art approaches, and JSD approach

Paper#	Rank(JSD)	Rank (Co-citation)	Rank (Boyack et al)	Rank(SWCA)
1	1	1	1	1
2	3	2	2	2
3	4	5	4	3
4	2	6	9	4
5	5	3	3	8
6	7	7	7	7
7	6	8	6	6
8	9	9	5	9
9	8	4	8	5

First, the comparison of co-citation approach with JSD measure has been shown in Table 6.20 using Spearman rank correlation. The ‘ p ’ value between JSD and co-citation approaches has been calculated by using Spearman Equation 6.11. The Equation exploits the statistic in Table 6.20.

$$p(JSD \text{ vs } Cocitation) = 1 - \frac{6 \times (0+1+1+16+4+0+4+0+16)}{9 \times (9 \times 9 - 1)} = 1 - 0.35 = \mathbf{0.65}$$

TABLE 6.20: Spearman rank correlation between JSD Vs Co-citation ranks

Rank(JSD) X_i	Rank(Co-citation) Y_i	Difference d_i	d_i^2
1	1	0	0
3	2	1	1
4	5	-1	1
2	6	-4	16
5	3	2	4
7	7	0	0
6	8	-2	4
9	9	0	0
8	4	4	16

Secondly, we have found the correlation between JSD rank and Boyack et al rank. The process of finding the spearman rank correlation between these two distribution is given in Table 6.21.

For one paper, the ‘ p ’ value between JSD and Boyack et al approaches has been calculated by Equation 6.11 using the statistic in Table 6.21.

$$p(\text{JSD vs Boyacketal}) = 1 - \frac{6 \times (0+1+0+49+4+0+0+16+0)}{9 \times (9 \times 9 - 1)} = 1 - 0.58 = \mathbf{0.42}$$

TABLE 6.21: Spearman rank correlation between JSD Vs Boyack et al ranks

Rank(JSD) X_i	Rank(Boyack et al) Y_i	Difference d_i	d_i^2
1	1	0	0
3	2	1	1
4	4	0	0
2	9	-7	49
5	3	2	4
7	7	0	0
6	6	0	0
9	5	4	16
8	8	0	0

Finally, we have found the correlation between JSD rank and SWCA rank. The Spearman rank correlation between these two distribution is given in Table 6.22.

TABLE 6.22: Spearman rank correlation between JSD Vs SWCA ranks

Rank(JSD) X_i	Rank(SWCA) Y_i	Difference d_i	d_i^2
1	1	0	0
3	2	1	1
4	3	1	1
2	4	-2	4
5	8	-3	9
7	7	0	0
6	6	0	0
9	9	0	0
8	5	3	9

For one paper, the ‘ p ’ value between JSD and SWCA approaches has been calculated by Equation 6.11 based on statistic in Table 6.22.

$$p(\text{JSD vs SWCA}) = 1 - \frac{6 \times (0+1+1+4+9+0+0+0+0)}{9 \times (9 \times 9 - 1)} = 1 - 0.2 = \mathbf{0.8}$$

(2) Kendall Rank Correlation Coefficient

Kendall Tau is a measure of the correlation between two ranked lists. It compares the number of concordant pairs with the number of discordant pairs between each list. The concordant pair is defined over two observations (x_i, y_i) and (x_j, y_j) [101]. if $x_i > x_j$ and $y_i > y_j$, then the pair at indices i, j is concordant. It means that the ranking at i, j in both ranking sets X and Y agree with each other. Similarly, the pair i, j is discordant if $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$. Kendall’s Tau coefficient is calculated using Equation 6.12.

$$T = \frac{C - D}{n(n - 1)/2} \quad (6.12)$$

where C is the number of concordant pairs, D is the number of discordant pairs, and the denominator represents the total number of possible pairs. The ‘ n ’ symbol shows the total number of elements in the each rank list. Thus, Kendall’s coefficient

falls in the range of $[-1, 1]$, where -1 means that the ranked lists are perfectly negatively correlated, 0 means that they are not significantly correlated, and 1 means that the ranked lists are perfectly correlated.

Like the Spearman rank correlation coefficient, we have found the correlation between the ranking lists in Table 6.19 by using Kendall tau Equation 6.12. The Kendall's tau coefficients of co-citation, Boyack et al, and SWCA ranks with JSD rank list has been calculated by Equation 6.12 respectively as shown below.

$$T(\text{JSD vs Co-citation}) = \frac{27-9}{9 \times (9-1)/2} = \mathbf{0.5}$$

$$T(\text{JSD vs Boyacketal}) = \frac{25-11}{9 \times (9-1)/2} = \mathbf{0.38}$$

$$T(\text{JSD vs SWCA}) = \frac{29-7}{9 \times (9-1)/2} = \mathbf{0.61}$$

(A). The Analysis of SWCA approach with state-of-the-approaches using JSD as Baseline

In this section, the rankings by the proposed approach have been compared with the state-of-the-art approaches: co-citation and Boyack et al against the benchmark ranking by JSD. There are total of 10 clusters and each cluster has nine ranked documents by each approach and the benchmark. The evaluation methodology compares the results in different aspects for example:

- It would be important to identify the average correlations (both Spearman and Kendall Tau) between the JSD and all other approaches.
- It would also be important to study the correlation between JSD and other approaches in different chunks of the ranking. For this purpose, the following chunks have been identified in the ranked results like: top@3, top@5, top@7, and top@9. It would be interesting to know that which approach (proposed

or state-of-the-art approaches) achieve better ranking at top of the rankings or in different defined chunks.

In Figure 6.8, the proposed approach SWCA has been compared with the state-of-the-art techniques: Co-citation and Boyack et al against the JSD ranking. The comparisons were done in all defined ranking chunks. The Figure 6.8 has total of four sub figures. The Figure 6.8(a) presents the comparisons between the proposed and state-of-the-art approaches in top 3 ranked papers only. Similarly, the comparisons between the proposed and state-of-the-art approaches in sets of top 5, top 7, and top 9 ranked papers have been shown in Figure 6.8(b), Figure 6.8(c), and Figure 6.8(d) respectively. In Figure 6.9, the overall comparison between proposed and state-of-the-art approaches has been shown in different sets of queries.

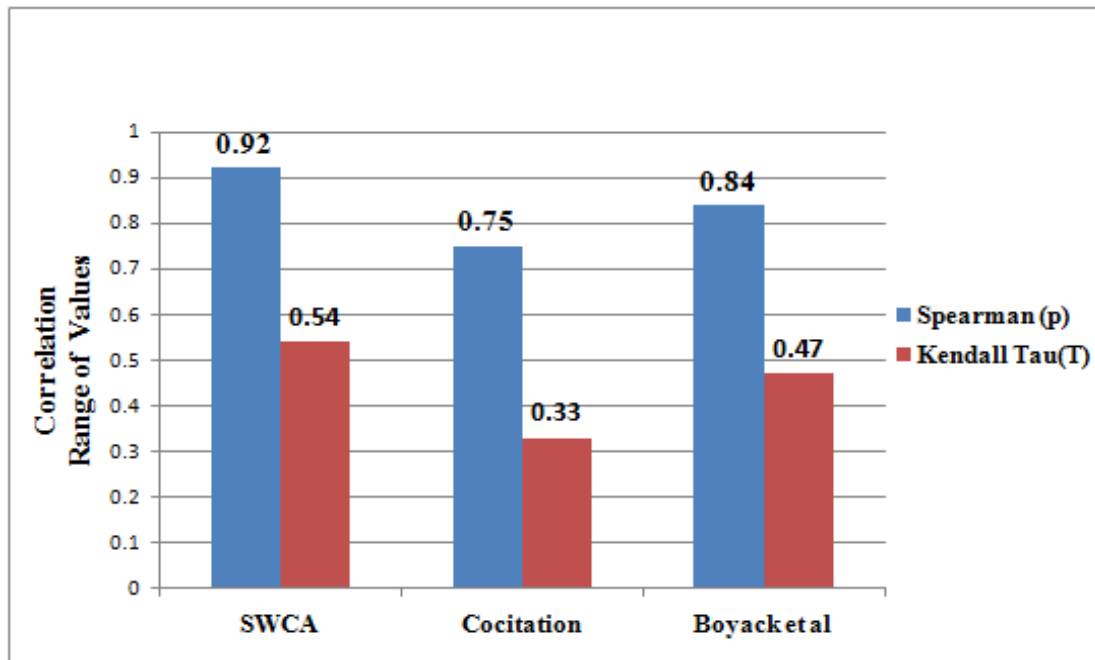
In first four subgraphs in Figure 6.8, on the X-axis, ranking approaches have been displayed like the proposed approach (SWCA) and the state-of-the-art approaches (co-citation and Boyack et al). The blue line represents the Spearman's correlation between JSD and all other approaches whereas the red line represents the Kendall tau's correlation between JSD and all other approaches. The Y-axis represents the correlation values. After critical study of results in these subgraphs, the following observations have been made.

1. The proposed approach has outperformed the state-of-the-art approaches based on JSD benchmark ranking using Spearman's measure.
2. The Boyack et al remained runners up approach which performed well than the co-citation technique based on both Spearman's and Kendall's tau.
3. One of interesting findings is that the Spearman's correlation of proposed and state-of-the-art approaches with JSD ranking is decreasing as long as we move downward in the ranking. It means that all compared approaches and the proposed approach have a potential to bring the important papers in the top of the ranking.
4. The SWCA approach has also performed well than other approaches in all subgraphs based on Kendall's tau measure except the Figure 6.8(c). In this

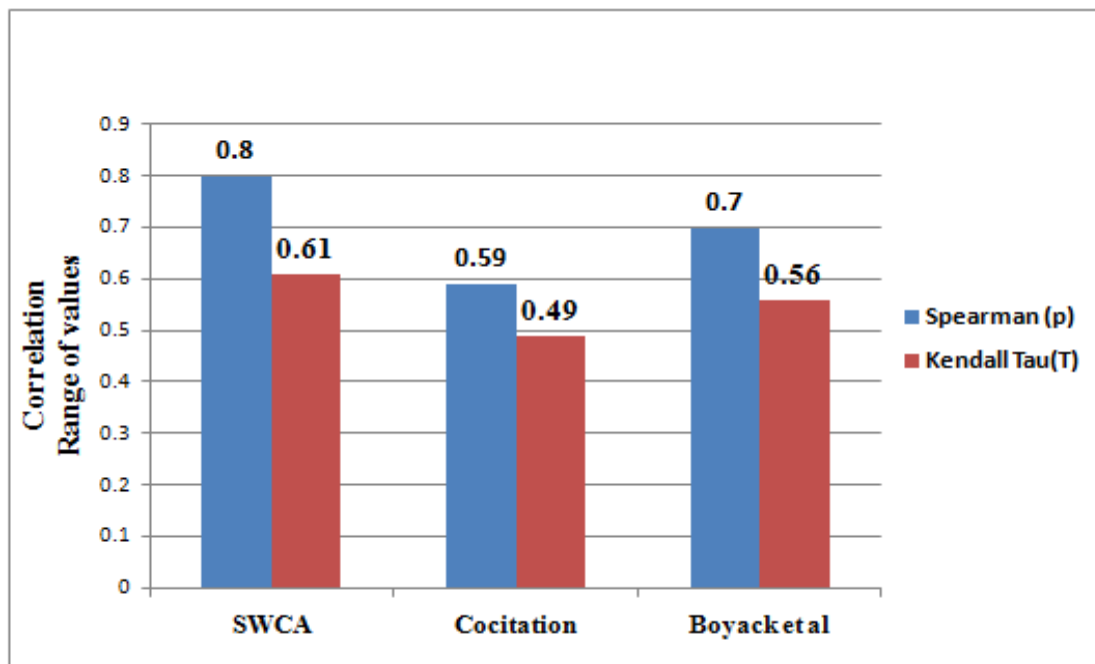
Figure, the correlation of proposed approach has decreased than state-of-the-art approaches. The reason is that Kendall's tau work on the number of concordant pairs and discordant pairs. In case of top@7, the number of concordant pairs were always noticed greater than the discordant pairs for the proposed approach with JSD ranking whereas, the Spearman's correlation works on the overall ranking distributions instead of noticing concordant and discordant pairs.

5. In Figure 6.8, the values of Spearman's correlation in all subgraphs are greater than the values of Kendall's tau correlation³, such behavior was also recorded by other researchers as well [102].

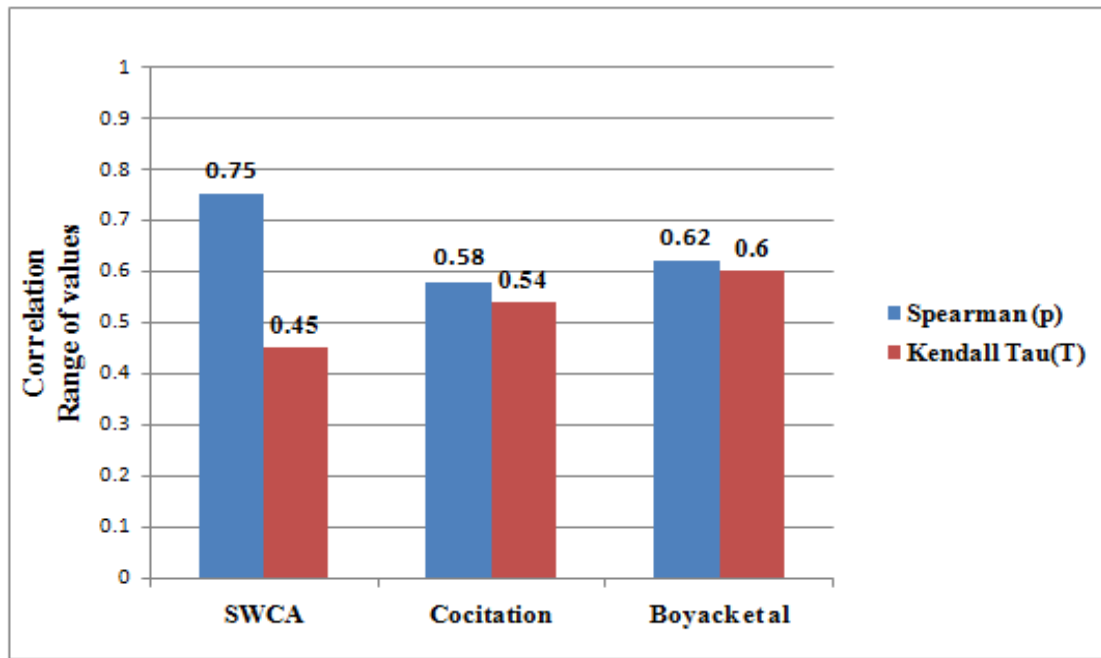
³<http://www.statisticssolutions.com/kendalls-tau-and-spearman-rank-correlation-coefficient/>



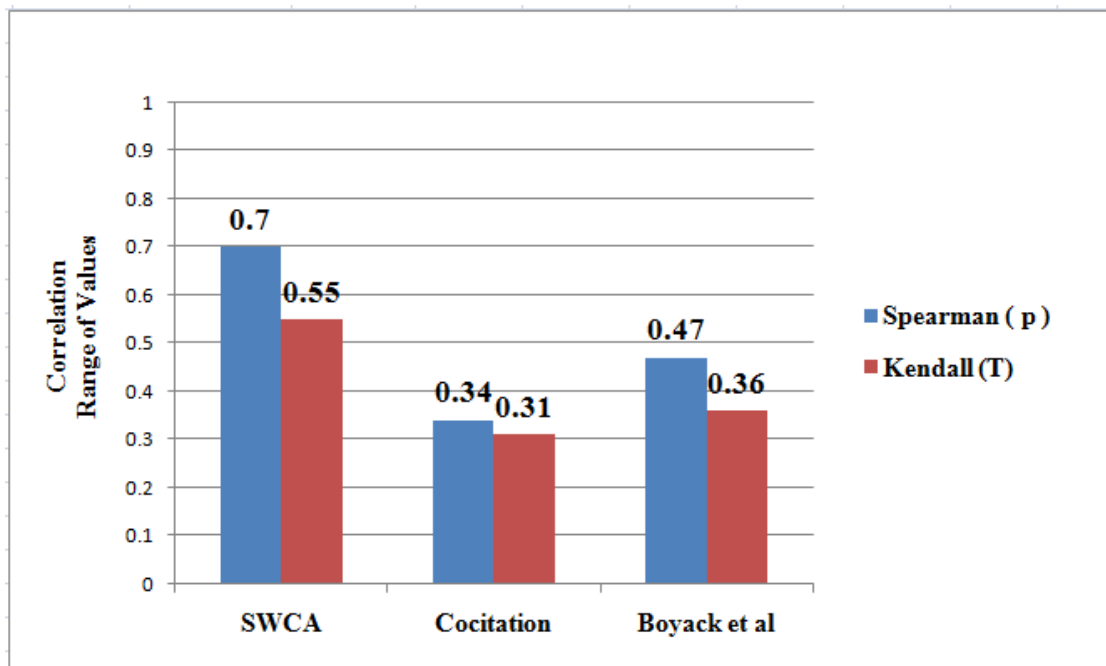
(a)



(b)



(c)



(d)

FIGURE 6.8: Proposed approach comparison with State-of-the-art approaches based on JSD ranking a) Average Correlation with JSD @ 3 b) Average Correlation with JSD @ 5 c) Average Correlation with JSD @ 7 d) Average Correlation with JSD @ 9

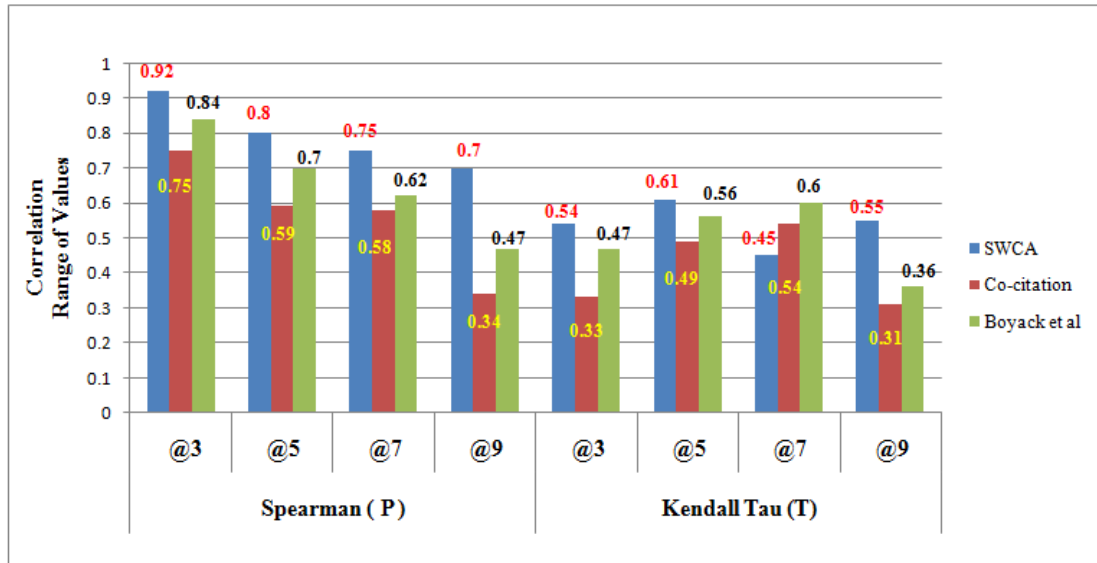


FIGURE 6.9: Comparison of Proposed technique with State-of-the-art techniques for different set of queries

(B). The Analysis of SWCA approach with state-of-the-approaches using Cosine Similarity as Baseline

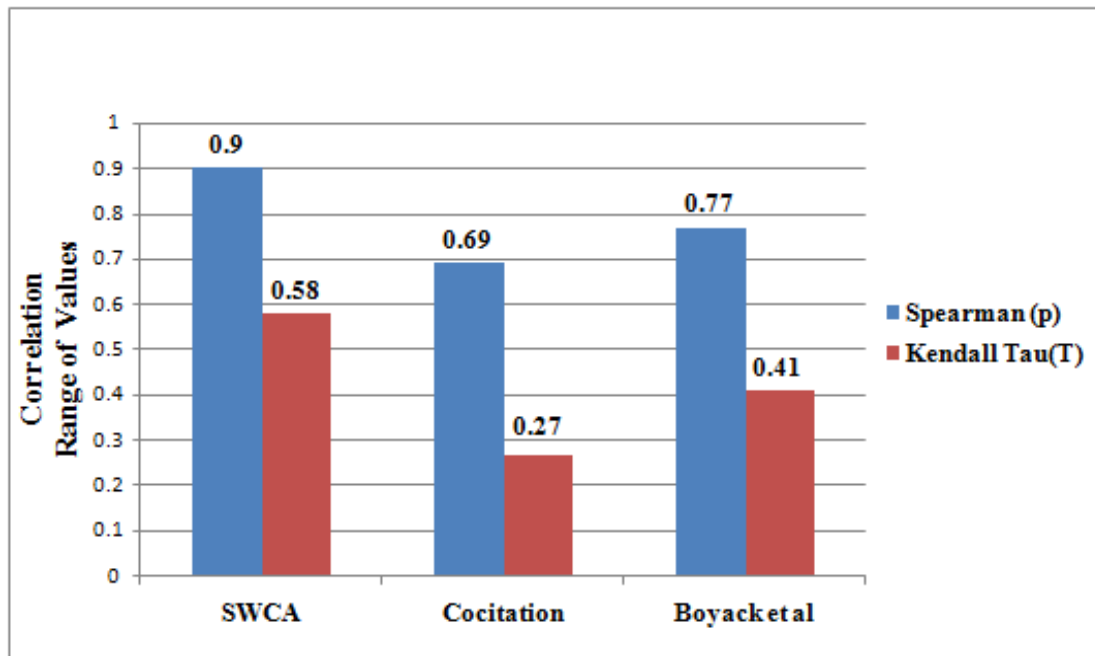
In this section, we have compared the ranking lists of proposed and state-of-the-approaches with the ranking of content based measure which is called cosine similarity [94, 101] as discussed in section 6.3.2.

In Figure 6.9, the proposed approach (SWCA) has been compared with the state-of-the-art techniques: Co-citation and Boyack et al against the Cosine ranking. The comparisons were done in all defined ranking chunks like were done in JSD based comparisons. The Figure 6.9 also has total of four sub figures. Figure 6.10(a) presents the comparisons between the proposed and state-of-the-art approaches in top 3 ranked papers only. Similarly, the comparisons between the proposed and state-of-the-art approaches in sets of top 5, top 7, and top 9 ranked papers have been shown in Figure 6.10(b), Figure 6.10(c), and Figure 6.10(d) respectively. In Figure 6.10, the overall comparison between proposed and state-of-the-approaches has been shown in different sets of queries.

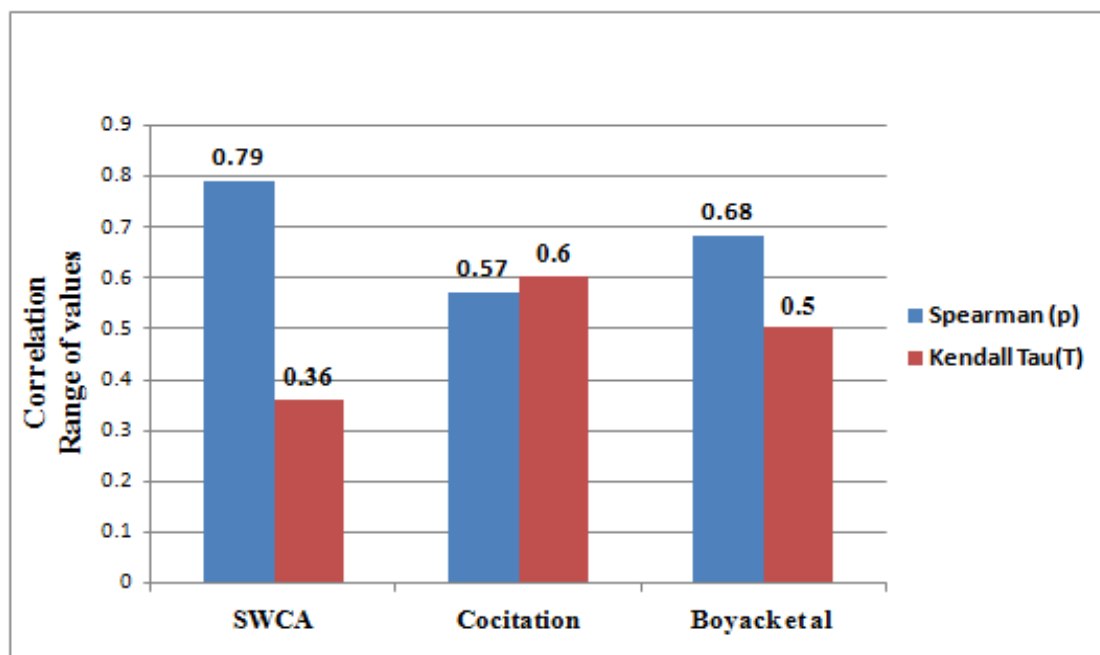
After critical study of results in these subgraphs, the following findings have been achieved.

1. The proposed approach has also outperformed the state-of-the-art approaches based on Cosine benchmark ranking using Spearman's measure.
2. The Boyack et al remained runners up approach which performed well than the co-citation technique based on both Spearman's and Kendall's tau.
3. Like in Figure 6.8, the Spearman's correlation of proposed and state-of-the-art approaches with Cosine ranking is also decreasing as long as we move downward in the ranking. It means that all compared approaches and the proposed approach have a potential to bring the important papers in the top of the ranking.
4. The SWCA approach has also performed well than other approaches in two subgraphs as shown in Figure 6.10(a) and Figure 6.10(d) based on Kendall's tau measure. While, in the remaining two subgraphs in Figure 6.10(b) and Figure 6.10(c), the score of proposed approach remained low due to the same reason as discussed in finding number 4 with Figure 6.8(c).
5. Once again, the values of Spearman's correlation in all subgraphs in Figure 6.9 are greater than the values of Kendall's tau correlation³ as explained above and was also pointed out by other researchers too [102].

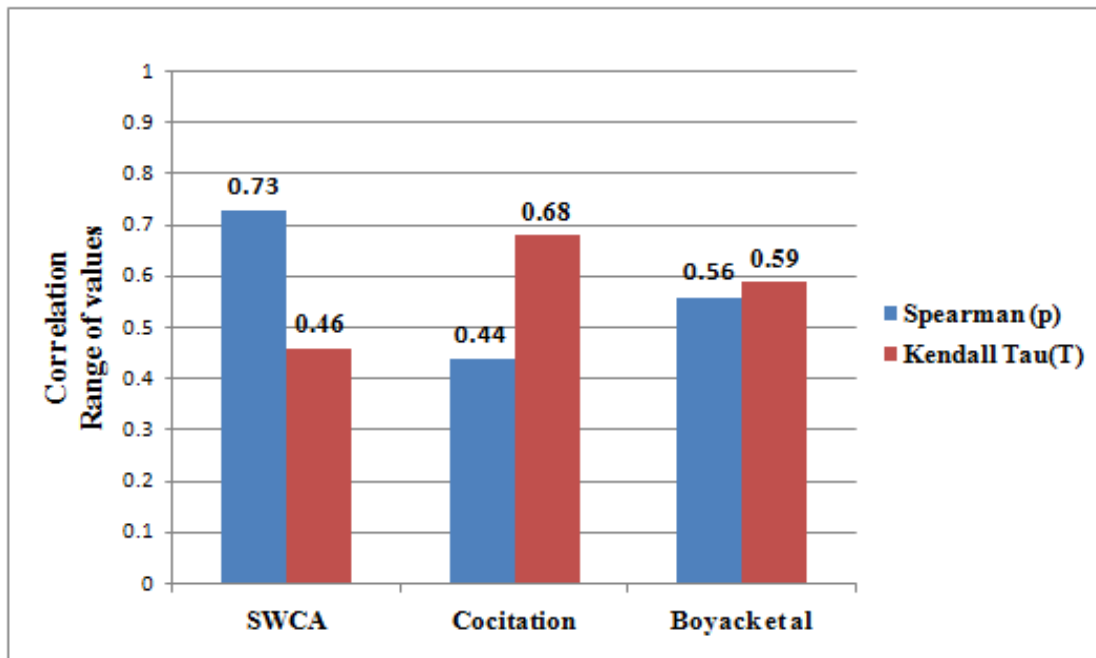
³<http://www.statisticssolutions.com/kendalls-tau-and-spearman-rank-correlation-coefficient/>



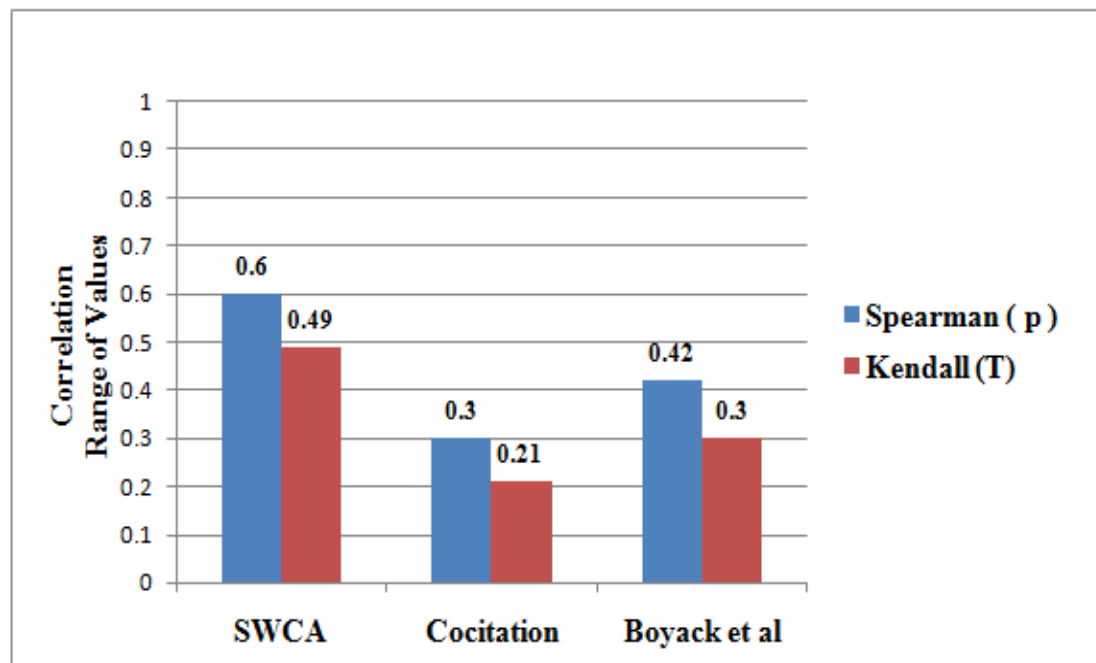
(a)



(b)



(c)



(d)

FIGURE 6.9: Proposed approach comparison with State-of-the-art approaches based on Cosine ranking a) Average Correlation with Cosine @ 3 b) Average Correlation with Cosine @ 5 c) Average Correlation with Cosine @ 7 d) Average Correlation with Cosine @ 9

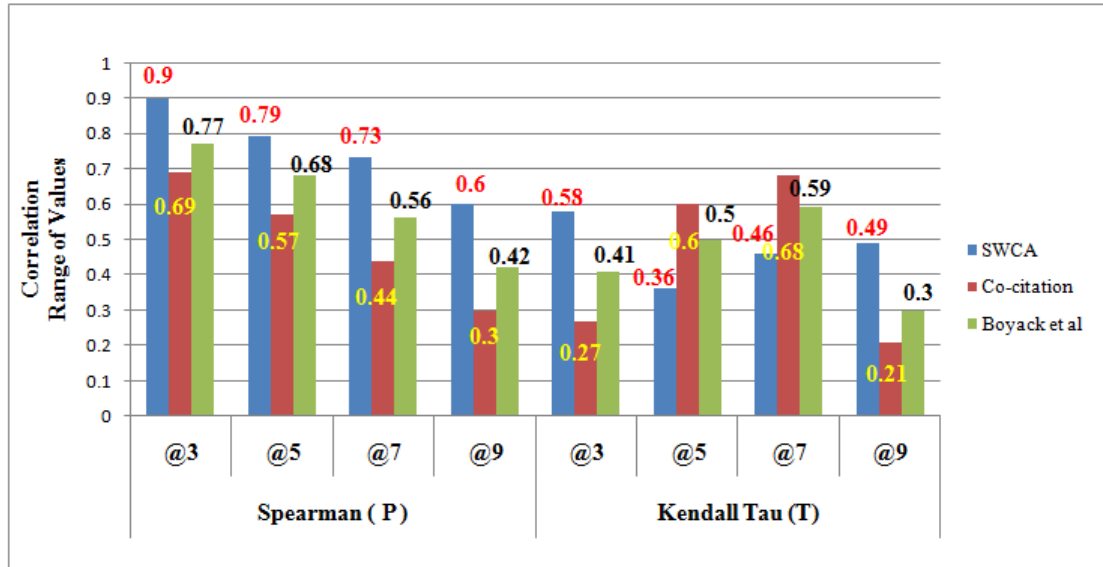


FIGURE 6.10: Comparison of Proposed technique with State-of-the-art techniques for different set of queries

6.4 Summary

In this chapter, first the final dataset was prepared for the empirical analysis of proposed approach. Second, the SWCA algorithm was elaborated along with relevancy score computation. In third step, we evaluated the two main components (1) generic section identification and (2) In-text citation patterns and frequencies identification, over the dataset which was prepared for the final task. In the fourth step, two benchmarks rank lists were prepared by JSD and cosine similarity approaches for the comparison of proposed approach, and state-of-the-art rankings. After the critical analysis of results, it is observed that the proposed approach (SWCA) has strong correlation with the two benchmarks: JSD and Cosine similarity than the state-of-the-art approaches. It means that the proposed approach has outperformed state-of-the-art approaches. The correlation comparisons of proposed and state-of-the-art approaches against benchmarks were made at two levels: (1) overall ranking level correlation comparisons, and (2) comparisons in the top 3, top 5, top 7, and top 9 ranked papers. In all of the cases, the proposed approach has outperformed the state-of-the-art approaches. Furthermore, all approaches were able to rank better papers in the top of their rankings, however, when they are compared with the proposed approach, the proposed approach was able to

consistently win in all cases. For correlation comparisons, two ranking evaluation metrics were used: Spearman, and Kendal Tau. In both of the evaluation parameters, the proposed approach was able to win, however, the correlation values of Spearman always remained higher than the Kandle Tau, which is also consistent with the findings of other researchers [102]

Chapter 7

Conclusion and Future Work

7.1 Conclusions

This thesis has critically evaluated the classical as well as state-of-the-art approaches in the domain of research paper recommendations. These approaches were classified into different categories like approaches based on: metadata of research documents, content of scientific papers, citation network, and user interactions in collaborative environment. The strengths and weaknesses of each type of approaches were highlighted in chapter 2.

Citation based approaches remained important as the relevant papers are picked and citing the authors of the papers. One recent approach, under the umbrella of citation based approaches is co-citation. The original co-citation model [32] considers two documents as most relevant, if both papers are occurring together in the reference sections of many other papers. Recently, citation models have started to consider the content of the citing papers where two or more papers were co-cited. In such approaches, the proximity of co-cited papers is identified in different ways for example at sentence level, paragraph level, and byte level etc.

However, the following two observations were made which serves as a research gap in this domain and is being focused in this research . The first problem in our focus is that the Citation proximity Analysis (CPA) and Citation Order

Analysis (COA) have different meanings in different structural components of research papers. For example two papers cited together in the “Discussion” section generally are not as important as two papers cited together in the “Results” section or “Methodology” section even if they are not cited in the same sentence. The second problem is that clustering content in equal bytes will again have logically many limitations, when two relevant references are placed in different clusters or when two irrelevant references are placed within the same cluster. Based on the above issues, this thesis comprehensively evaluated and experimented co-cited papers in the generic structure/sections (Introduction, Literature, Methodology, Results and Discussion) of scientific documents and evaluated whether considering the proximity of research papers in section level may increase the accuracy of identifying relevant research papers or not.

For the section wise co-citation analysis (SWCA), we developed the comprehensive methodology as discussed in details in chapter 3. This methodology consisted of three main phases (1) Data preparation phase (2) Section wise co-citation analysis phase, and (3) Result evaluation phase. Furthermore, the section wise co-citation analysis phase consisted of three main components: Generic sections identification, In-text co-citation patterns and frequencies identification, and section wise co-citation analysis. The first component has been discussed in details in chapter 4. The second component has been explained in chapter 5. The third and last component has been discussed in chapter 6.

7.2 Contributions

In chapter 4, we discussed our first contribution. In this contribution, we proposed, implemented, and evaluated a novel approach for structural components identification and mapping on generic sections. Furthermore, in the evaluation process, two annotated testing dataset were selected with 150 and 300 citing documents respectively. The technique was evaluated based on well-known measure of precision, recall and F-score. The precision and recall values were computed

for each standard section “Introduction”, “Related Work”, “Methodology”, “Results”, “Discussion” and “Conclusion”. For the comparison of proposed approach, the state-of-the-art [28] technique was also applied on the same dataset. The aggregated F-score of proposed approach was 0.92 over the both datasets while the F-score of state-of-the-art technique was 0.81. The state-of-the-art approach only considered the key-terms in section labels and the position of sections in the research papers. In our approach, the patterns such as the key-terms, section number, number of citations, number of figures, number of tables, first person plural pronoun, number of pages, and number of structural components were used for the accurate identification of section mapping.

In chapter 5, we discussed the second contribution. In the second contribution, we proposed citation-anchors taxonomy after the critical analysis of citation-anchors in the citing documents, literature approaches, and well known citation representation formats such as APA, MLA, AMA, and CBE. Secondly, we proposed, implemented, and evaluated a novel approach for the identification of in-text citation patterns and frequencies in the citing documents. For the evaluation of proposed approach, two datasets were prepared, one from Computer Science Journal J.UCS and the other from digital library CiteSeerX having indexed many conferences and journals. The testing set of J.UCS dataset consisted of 3,000 citations. The testing set of CiteSeerX dataset consisted of 5,000 citations. For the comparison, the state-of-the-art technique was also implemented over the same datasets. Both approaches were evaluated based on well-known measure of precision, recall and F-score. The results were compared with the state-of-the-art approach proposed by Shahid et al [18]. The proposed model has comprehensively outperformed the state-of-the-art approach by scoring average F-score of 0.97 as compared to baseline of 0.58. The state-of-the-art technique used the exact matching of citation-tag with citation-anchor. But the highlighted issues in section 5.2 of in-text citation anchor were not detected with exact matching. However, the proposed approach used multiple evidences such as: innovative heuristics, lessons learned from previous approaches in the literature and learning from initial experimentations.

After the above mentioned two important innovative approaches, we used the above approaches into the overall proposed approach known as Section Wise Co-citation Analysis (SWCA). The last and third contribution: section wise co-citation analysis (SWCA) is discussed in chapter 6. In that chapter, first the dataset was prepared for the empirical analysis of proposed approach. The dataset consisted of ‘50’ query papers, ‘450’ co-cited papers pairs, and ‘11,875’ citing documents. Second, the SWCA algorithm has been elaborated along with relevancy score computation. In third step, we evaluated the two core components (1) generic section identification and (2) In-text citation patterns and frequencies identification, over the dataset which was prepared for the final task. In the fourth step, two benchmarks rank lists were prepared by JSD and cosine similarity approaches for the comparison of proposed approach with the state-of-the-art approaches rankings.

The state-of-the-art approaches used for comparisons were: (1) Standard co-citation approach [32] and (2) Citation Proximity Analysis based on bytes [21]. There were two benchmark rankings, one ranking by the proposed approach and two rankings by the state-of-the-art approaches. To compare rankings, two well known measures are used in the literature known as: Kandle Tau and Spearsman’s correlations. Both of these measures were used to compare the rankings obtained by the proposed approach, and the state-of-the-art approaches with both benchmark rankings. The interesting findings were: (1) The average correlation for the proposed approach remained 0.65 as compared to 0.5 by CPA and 0.48 by standard co-citation. (3) The rankings were also compared into different ranking chunks such as comparing in the top@3, top@5, top@7, and top@9 ranked lists. The result shows that all approaches were able to bring most relevant papers to the top of the rankings; however, the proposed approach was able to bring most of the relevant ones in the top of the ranking

7.3 Limitations of Proposed Approach

The proposed approach SWCA is developed based on two main components (1) ILMRaD Structure Identification and (2) Intext-citation patterns and their frequencies identification. The experimental study of our research work shows that SWCA approach has better results than state-of-the-art approaches. However, the proposed approach has also some deficiencies due to the limitations that exist in two basic components. The limitations of these components are discussed below.

1. In the proposed approach SWCA, the metadata of research papers have been prepared from two openly available digital libraries, such as J.UCS and CiteSeerX. In the absence of these libraries, our approach is not able to construct the set of citing documents, and the set of co-cited pairs. These two sets of research papers are used as input in the section wise co-citation analysis (SWCA) technique.
2. The input of the proposed approach SWCA is a research document. This approach has been built based on two formats, PDF and plain-text of research document. The SWCA technique will not work with other formats like Postscript (.ps).
3. We have used the online PDFx conversion tool for the PDF-to-XML conversion. Hence, the proposed approach also depends on the PDFx tool. There is, therefore, a possibility that the proposed approach may not properly work with the XML document converted by other PDF-to-XML tool.
4. The rules in the proposed approach are prepared from J.UCS and CiteSeerX research documents. The research papers of these two digital libraries are related to computer science domain. These rules are constructed for the identification of ILMRaD structure and in-text citation frequencies and their pattern. There is a possibility that these rules may not properly work for the research papers of PubMed Journal because these research papers follow the IMRaD structure while our approach follows the ILMRaD structure.

7.4 Future Work

This research has opened many research avenues for future, some of them have been highlighted below:

1. Although the proposed approach for section identification and mapping has acquired good accuracy on different datasets such as J.UCS, CiteSeer and the World challenge at ESWC conference. The approach should further be tested on diversified domains such as: Medical, Chemistry, Physics, Neuroscience etc, There might be a need for some further fine tuning of the approach in different domains. Furthermore, the technique may be extended and evaluated by: (a) incorporating authors styles of writing, (b) by analyzing the content represented in the sections, (c) language constructs for example the phrases are in passive voice or active voice etc.
2. The proposed approach of in-text citation frequency identification may be experimented and extended in future by analyzing all citation styles available in the scientific community. Furthermore, when authors make mistakes in writing references or providing citations, there could be an automated approach which will be able to identify such anomalies. This will of course, increase the overall accuracy.
3. The conversion of PDF to XML and text remained a challenge. Around 5% PDFs were not in a format to be recognized by the state-of-the-art tools for conversion into text. Although a researcher earned his PhD for proposing propose an innovative approach to convert PDF to XML, but still around 5% documents were not converted accurately. Therefore, there is a need for more experimentations and innovative approaches which could lead to convert all PDFs to XML and text formats.
4. In the proposed approach of Section Wise Co-citation Analysis (SWCA), the weights of different sections were assigned based on the state-of-the-art approaches, Some more experiments can be done to evaluate different weights that might lead to more accurate results.

5. In SWCA, the sections weights were considered for papers co-cited in the same sections; there might be some meanings of co-citation into different sections. Such phenomena needs to be experimented and evaluated in different sections.

References

- [1] P. O. Larsen and M. Von Ins, “The rate of growth in scientific publication and the decline in coverage provided by science citation index,” *Scientometrics*, vol. 84, no. 3, pp. 575–603, 2010.
- [2] L. Bornmann and R. Mutz, “Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references,” *Journal of the Association for Information Science and Technology*, vol. 66, no. 11, pp. 2215–2222, 2015.
- [3] A. Hodgson and L. Schlager, “Closing the pdf gap: Readcubes experiments in reader-focused design,” *Learned Publishing*, vol. 30, no. 1, pp. 65–69, 2017.
- [4] M. Ware and M. Mabe, “The stm report: An overview of scientific and scholarly journal publishing. http://www.stmassoc.org/2009_10_13_mwc_stm_report.pdf,” 2015.
- [5] A. E. Jinha, “Article 50 million: an estimate of the number of scholarly articles in existence,” *Learned Publishing*, vol. 23, no. 3, pp. 258–263, 2010.
- [6] 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. ACM.
- [7] S. Li, P. Brusilovsky, S. Su, and X. Cheng, “Conference paper recommendation for academic conferences,” *IEEE Access*, vol. 6, pp. 17 153–17 164, 2018.

-
- [8] J. Beel, S. Langer, M. Genzmehr, B. Gipp, C. Breitingner, and A. Nürnberger, “Research paper recommender system evaluation: a quantitative literature survey,” in *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, pp. 15–22. ACM, 2013.
- [9] J. Beel, B. Gipp, S. Langer, and C. Breitingner, “paper recommender systems: a literature survey,” *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, 2016.
- [10] X. Su and T. M. Khoshgoftaar, “A survey of collaborative filtering techniques,” *Advances in artificial intelligence*, vol. 2009, p. 4, 2009.
- [11] Y. Cai, H.-f. Leung, Q. Li, H. Min, J. Tang, and J. Li, “Typicality-based collaborative filtering recommendation,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 3, pp. 766–779, 2014.
- [12] C.-H. Chen, S. D. Mayanglambam, F.-Y. Hsu, C.-Y. Lu, H.-M. Lee, and J.-M. Ho, “Novelty paper recommendation using citation authority diffusion,” in *Technologies and Applications of Artificial Intelligence (TAAI), 2011 International Conference on*. IEEE, 2011, pp. 126–131.
- [13] A. Livne, E. Adar, J. Teevan, and S. Dumais, “Predicting citation counts using text and graph mining,” in *Proc. the iConference 2013 Workshop on Computational Scientometrics: Theory and Applications*, 2013.
- [14] S. Pruitikaneer, L. Di Jorio, A. Laurent, and M. Sala, “Paper recommendation system: A global and soft approach,” in *FUTURE COMPUTING’2012: Fourth International Conference on Future Computational Technologies and Applications*, 7, 2012.
- [15] D. Kaplan, R. Iida, and T. Tokunaga, “Automatic extraction of citation contexts for research paper summarization: A coreference-chain based approach,” in *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*. Association for Computational Linguistics, 2009, pp. 88–95.

-
- [16] S. Tuarob, P. Mitra, and C. L. Giles, “A classification scheme for algorithm citation function in scholarly works,” in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital Libraries*, pp. 367–368. ACM, 2013.
- [17] W.-R. Hou, M. Li, and D.-K. Niu, “Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution,” *BioEssays*, vol. 33, no. 10, pp. 724–727, 2011.
- [18] 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.
- [19] Y. Liang, Q. Li, and T. Qian, “Finding relevant papers based on citation relations,” in *Web-age information management*, pp. 403–414, Springer, 2011.
- [20] M. T. Afzal, *Context aware information discovery for scholarly e-community*. PhD thesis, Graz University of Technology, Austria, 2010.
- [21] 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.
- [22] B. Gipp and J. Beel, “Citation proximity analysis (cpa)-a new approach for identifying related work based on co-citation analysis,” in *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI09)*. 2:571–575, Rio de Janeiro (Brazil): International Society for Scientometrics and Informetrics, 2009.
- [23] P. R. Nair and V. D. Nair, *Scientific writing and communication in agriculture and natural resources*. Springer, 2014.
- [24] M. Bertin, I. Atanassova, Y. Gingras, and V. Larivière, “The invariant distribution of references in scientific articles,” *Journal of the Association for Information Science and Technology*, vol. 67, no. 1, pp. 164–177, 2016.
- [25] M. Bertin and I. Atanassova, “Weak links and strong meaning: The complex phenomenon of negational citations.” in *BIR@ ECIR*, pp. 14–25, 2016.

- [26] Teufel, “Citations and sentiment,” *Workshop on Text mining for Scholarly Communications and Repositories, Manchester Interdisciplinary Biocentre, University of Manchester Researcher Thesis*, 2009.
- [27] M.-T. Luong, T. D. Nguyen, and M.-Y. Kan, “Logical structure recovery in scholarly articles with rich document features,” *Multimedia Storage and Retrieval Innovations for Digital Library Systems*, vol. 270, p. 2, 2012.
- [28] A. Shahid and M. T. Afzal, “Section-wise indexing and retrieval of research articles,” *Cluster Computing*. pp. 1–12, 2017.
- [29] R. Babeley, “A study on automated citation analysis in the field of library and information science literature,” *International Journal of Contemporary Research and Review*, vol. 7, no. 12, 2016.
- [30] E. Garfield, “The history and meaning of the journal impact factor,” *Jama*, vol. 295, no. 1, pp. 90–93, 2006.
- [31] —, “Science citation index—a new dimension in indexing,” *Science*, vol. 144, no. 3619, pp. 649–654, 1964.
- [32] 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.
- [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)*, pp. 309–315, 2009.
- [34] A. Shahid, M. Afzal, and M. Qadir, “Discovering semantic relatedness between scientific articles through citation frequency,” in *Workshop on Text mining for Scholarly Communications and Repositories, Australian Journal of Basic Applied Sciences*. 5:1599–1604, 2011.
- [35] M. M. Kessler, “Bibliographic coupling between scientific papers,” *American documentation*, vol. 14, no. 1, pp. 10–25, 1963.

- [36] S. Liu, C. Chen, K. Ding, B. Wang, K. Xu, and Y. Lin, "Literature retrieval based on citation context," *Scientometrics*, vol. 101, no. 2, pp. 1293–1307, 2014.
- [37] B. Gipp and J. Beel, "Identifying related documents for research paper recommender by cpa and coa," in *International Conference on Education and Information Technology (ICEIT09), Lecture Notes in Engineering and Computer Science*. 1:636–639, 2009.
- [38] Z. Hu, C. Chen, and Z. Liu, "The recurrence of citations within a scientific article." in *Salah, A., Tonta, A., Akdag Salah, C., Sugimoto, U.A. (eds.) 15th International Society of Scientometrics and Informetrics Conference. ISSI, Bogazii University Printhouse, Istanbul, Turkey (June 29 to July 3 2015)*, 2015.
- [39] H. D. White, J. Buzydlowski, and X. Lin, "Co-cited author maps as interfaces to digital libraries: designing pathfinder networks in the humanities," in *International Conference on Information Visualization, 2000. Proceedings. IEEE*, pp. 25–30. IEEE, 2000.
- [40] X. Lin, H. D. White, and J. Buzydlowski, "Real-time author co-citation mapping for online searching," *Information processing & management*, vol. 39, no. 5, pp. 689–706, 2003.
- [41] Z. Yang, L. Hong, and B. D. Davison, "Topic-driven multi-type citation network analysis," in *Adaptivity, Personalization and Fusion of Heterogeneous Information*, pp. 24–31, 2010.
- [42] E. Garfield *et al.*, "Citation analysis as a tool in journal evaluation." American Association for the Advancement of Science, 178(4060):471–479, 1972.
- [43] H. F. Moed, "Measuring contextual citation impact of scientific journals," *Journal of Informetrics*, vol. 4, no. 3, pp. 265–277, 2010.
- [44] R. Kumar, *Research Methodology-A step by step guide for beginners*, 3rd ed. London: Sage, 2011.

- [45] K. McCain and K. Turner, "Citation context analysis and aging patterns of journal articles in molecular genetics," *Scientometrics*, vol. 17, no. 1-2, pp. 127–163, 1989.
- [46] S. Maričić, J. Spaventi, L. Pavičić, and G. Pifat-Mrzljak, "Citation context versus the frequency counts of citation histories," *Journal of the Association for Information Science and Technology*, vol. 49, no. 6, pp. 530–540, 1998.
- [47] M. Bertin and I. Atanassova, "A study of lexical distribution in citation contexts through the imrad standard," *PloS Negl. Trop. Dis.*, vol. 1, pp. 83–402, 2014.
- [48] Z. Hu, C. Chen, and Z. Liu, "Where are citations located in the body of scientific articles? a study of the distributions of citation locations," *Journal of Informetrics*, vol. 7, no. 4, pp. 887–896, 2013.
- [49] Y. Ding, X. Liu, C. Guo, and B. Cronin, "The distribution of references across texts: Some implications for citation analysis," *Journal of Informetrics*, vol. 7, no. 3, pp. 583–592, 2013.
- [50] A. G. Hu and A. B. Jaffe, "Patent citations and international knowledge flow: the cases of korea and taiwan," *International journal of industrial organization*, vol. 21, no. 6, pp. 849–880, 2003.
- [51] P. M. Editors *et al.*, "The impact factor game: it is time to find a better way to assess the scientific literature: conversations," *PLos Med*, vol. 3, no. 6, p. e291, 2006.
- [52] S. Liu and C. Chen, "The proximity of co-citation," *Scientometrics*, vol. 91, no. 2, pp. 495–511, 2011.
- [53] 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*, pp. 89–98. ACM, 1998.

- [54] D. Bergmark, “Automatic extraction of reference linking information from online documents,” Cornell Digital Library Research Group, 2000–1821, Tech. Rep., 2000.
- [55] F. Nadirman, A. Ridha, and A. Annisa, “Searching and visualization of references in research documents,” *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 12, no. 2, pp. 447–454, 2014.
- [56] D. Tkaczyk, A. Collins, P. Sheridan, and J. Beel, “Evaluation and comparison of open source bibliographic reference parsers: A business use case,” *arXiv preprint arXiv:1802.01168*, 2018.
- [57] D. Tkaczyk and Ł. Bolikowski, “Extracting contextual information from scientific literature using cermine system,” in *Semantic Web Evaluation Challenge*. Springer, 2015, pp. 93–104.
- [58] P. Lopez, “Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications,” in *International Conference on Theory and Practice of Digital Libraries*, pp. 473–474, Springer, 2009.
- [59] L. R. Zientek, J. M. Werner, M. V. Campuzano, and K. Nimon, “The use of google scholar for research and research dissemination,” *New Horizons in Adult Education and Human Resource Development*, vol. 30, no. 1, pp. 39–46, 2018.
- [60] S. Bhatia, C. Caragea, H.-H. Chen, J. Wu, P. Treeratpituk, Z. Wu, M. Khabsa, P. Mitra, and C. L. Giles, “Specialized research datasets in the citeseerx digital library,” *D-Lib Magazine*, vol. 18, no. 7/8, 2012.
- [61] J. Beel and B. Gipp, “Google scholars ranking algorithm: an introductory overview,” in *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI09)*, vol. 1. Rio de Janeiro (Brazil), 2009, pp. 230–241.

- [62] 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.
- [63] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: an open architecture for collaborative filtering of netnews," in *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pp. 175–186, ACM, 1994.
- [64] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, "Eigentaste: A constant time collaborative filtering algorithm," *Information Retrieval*, vol. 4, no. 2, pp. 133–151, 2001.
- [65] G. Koutrika, B. Bercovitz, F. Kaliszan, H. Liou, and H. Garcia-Molina, "Courserank: A closed-community social system through the magnifying glass." in *Proc. Of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM09), San Jose, California, May 17 20, 2009.ICWSM, 2009*.
- [66] M. Zhang, W. Wang, and X. Li, "A paper recommender for scientific literatures based on semantic concept similarity," in *International Conference on Asian Digital Libraries*, pp. 359–362. Springer, 2008.
- [67] K. Hong, H. Jeon, and C. Jeon, "Researchers' interesting areas recognition system using implicit feedback," *International Journal of Advancements in Computing Technology*, vol. 5, no. 15, p. 127, 2013.
- [68] K. Hoxha, A. Kika, E. Gani, and S. Greca, "Towards a modular recommender system for research papers written in albanian," *arXiv preprint arXiv:1405.0190*, 2014.
- [69] 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*, pp. 421–430. ACM, 2010.

- [70] E. Garfield, "Use of journal citation reports and journal performance indicators in measuring short and long term journal impact," *Croatian Medical Journal*, vol. 41, no. 4, pp. 368–374, 2000.
- [71] G. Colavizza, K. W. Boyack, N. J. van Eck, and L. Waltman, "The closer the better: Similarity of publication pairs at different cocitation levels," *Journal of the Association for Information Science and Technology*, vol. 69, no. 4, pp. 600–609, 2018.
- [72] 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*, pp. 705–706. ACM, 2007.
- [73] M. T. Afzal, W. Balke, H. Maurer, and N. Kulathuramaiyer, "Improving citation mining," in *First International Conference on Networked Digital Technologies, 2009*, pp. 116–121. IEEE, 2009.
- [74] Y. K. Jeong, M. Song, and Y. Ding, "Content-based author co-citation analysis," *Journal of Informetrics*, vol. 8, no. 1, pp. 197–211, 2014.
- [75] 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.
- [76] R. Ahmad, M. T. Afzal, and M. A. Qadir, "Pattern analysis of citation-anchors in citing documents for accurate identification of in-text citations," *IEEE Access*, vol. 5, pp. 5819–5828, 2017.
- [77] M. T. Afzal, "Applying ontological framework for finding links into the future from web," in *5th International Conference on Semantic Systems (I-Semantics 2009)*, pp. 656–662, 2009.
- [78] L. Sollaci and M. Pereira, "The introduction, methods, results, and discussion (imrad) structure," *a fifty-year survey. Journal of the Medical Library Association*, 92(3): 364., 2004.

- [79] S. A. Socolofsky, “How to write a research journal article in engineering and science. texas a&m university,” 2004.
- [80] T. D. Nguyen and M.-Y. Kan, “Keyphrase extraction in scientific publications,” pp. 317–326, Springer, 2007.
- [81] M. Bertin, I. Atanassova, V. Lariviere, and Y. Gingras, “The distribution of references in scientific papers: An analysis of the inrad structure,” in *Proceedings of the 14th ISSI Conference*. 591:603, 2013.
- [82] G. M. Hall and Z. Sestak, *How to write a paper, 2nd edn.* London:BMJ, 2003.
- [83] M. Palumbo, “How to write a technical paper,” 2013.
- [84] Guigo and Roderic, “An introduction to position specific scoring matrix,” <http://bioinformatica.upf.edu/T12/MakeProfile.html>, Retrieved 26 July 2017, 2013.
- [85] R. Ahmad, M. T. Afzal, and M. A. Qadir, “Information extraction from pdf sources based on rule-based system using integrated formats,” in *the semantic web: ESWC 2016 Challenges, Anissaras, Crete, Greece*. 641:293–308, Communications in computer and information science. Springer., 2016.
- [86] S. Bhatia, S. Tuarob, P. Mitra, and C. L. Giles, “An algorithm search engine for software developers,” in *Proceedings of the 3rd International Workshop on Search-Driven Development: Users, Infrastructure, Tools, and Evaluation*, pp. 13–16. ACM, 2011.
- [87] S. Tuarob, P. Mitra, and C. L. Giles, “Building a search engine for algorithms by suppowong tuarob, prasenjit mitra, and c. lee giles with martin vesely as coordinator,” *ACM SIGWEB Newsletter*, no. Winter, p. 5, 2014.
- [88] U. J. Pape, S. Rahmann, F. Sun, and M. Vingron, “Compound poisson approximation of the number of occurrences of a position frequency matrix (pfm) on both strands,” *Journal of Computational Biology*, vol. 15, no. 6, pp. 547–564, 2008.

- [89] P. Ciancarini, A. Di Iorio, A. G. Nuzzolese, S. Peroni, and F. Vitali, “Semantic annotation of scholarly documents and citations,” in *Congress of the Italian Association for Artificial Intelligence*. Springer, 2013, pp. 336–347.
- [90] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, and K. Börner, “Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches,” *PloS one*, vol. 6, no. 3, p. e18029, 2011.
- [91] S. Mackie, R. McCreadie, C. Macdonald, and I. Ounis, “On choosing an effective automatic evaluation metric for microblog summarisation,” in *Proceedings of the 5th Information Interaction in Context Symposium*, pp. 115–124. ACM, 2014.
- [92] A. Rutherford and N. Xue, “Improving the inference of implicit discourse relations via classifying explicit discourse connectives.” in *HLT-NAACL*, pp. 799–808, 2015.
- [93] L. Wang, H. Raghavan, C. Cardie, and V. Castelli, “Query-focused opinion summarization for user-generated content,” *arXiv preprint arXiv:1606.05702*, 2016.
- [94] A. Louis and A. Nenkova, “Automatically assessing machine summary content without a gold standard,” *Computational Linguistics*, vol. 39, no. 2, pp. 267–300, 2013.
- [95] K. Krstovski, D. A. Smith, and M. J. Kurtz, “Automatic construction of evaluation sets and evaluation of document similarity models in large scholarly retrieval systems.” in *AAAI Workshop: Scholarly Big Data*, 2016.
- [96] Z. Zhang and M. Lan, “Ecnu at semeval 2016 task 6: Relevant or not? supportive or not? a two-step learning system for automatic detecting stance in tweets.” in *SemEval@ NAACL-HLT*, pp. 451–457, 2016.

- [97] M. A. H. Al-Hagery, "Google search filter using cosine similarity measure to find all relevant documents of a specific research topic," *International Journal of Education and Information Technologies*, vol. 10, pp. 229–242, 2016.
- [98] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of jaccard coefficient for keywords similarity," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 1(6), 2013.
- [99] M. Erritali, A. Beni-Hssane, M. Birjali, and Y. Madani, "An approach of semantic similarity measure between documents based on big data," *International Journal of Electrical and Computer Engineering*, vol. 6, no. 5, p. 2454, 2016.
- [100] R. Subhashini and V. J. S. Kumar, "Evaluating the performance of similarity measures used in document clustering and information retrieval," in *First International Conference on Integrated Intelligent Computing (ICIIC), 2010*. 27–31. IEEE, 2010.
- [101] J. M. Lingeman and H. Yu, "Learning to rank scientific documents from the crowd," *arXiv preprint arXiv:1611.01400*, 2016.
- [102] G. A. Fredricks and R. B. Nelsen, "On the relationship between spearman's rho and kendall's tau for pairs of continuous random variables," *Journal of Statistical Planning and Inference*, vol. 137, no. 7, pp. 2143–2150, 2007.
- [103] M. Faisal, A. Daud, and A. Akram, "Expert ranking using reputation and answer quality of co-existing users." *International Arab Journal of Information Technology (IAJIT)*, vol. 14, no. 1, 2017.