

CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD



Ontology for Inter-Research Paper Similarity Measures (COReS) and its Applications

by

Qamar Mahmood

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

in the

Faculty of Computing

Department of Computer Science

2019

Ontology for Inter-Research Paper Similarity Measures (COReS) and its Applications

By

Qamar Mahmood

PC103006

Dr. Yllias Chali

University of Lethbridge, Canada

(Foreign Evaluator 1)

Dr. Ozgu Can

Ege Universitesi Muhendislik Fakultesi, Turkey

(Foreign Evaluator 2)

Dr. Muhammad Abdul Qadir

(Thesis Supervisor)

Dr. Nayyer Masood

(Head, Department of Computer Science)

Dr. Muhammad Abdul Qadir

(Dean, Faculty of Computing)

DEPARTMENT OF COMPUTER SCIENCE
CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
ISLAMABAD

2019

Copyright © 2019 by Mr. Qamar Mahmood

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 “**Ontology for Inter-Research Paper Similarity Measures (COREs) and its Applications**” was conducted under the supervision of **Dr. Muhammad Abdul Qadir**. No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the **Department of Computer Science, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Computer Science**. The open defence of the thesis was conducted on **September 12, 2019**.

Student Name : Mr. Qamar Mahmood
(PC103006)

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

Examination Committee :

(a) External Examiner 1: Dr. Majid Iqbal Khan
Associate Professor
COMSATS University, Islamabad

(b) External Examiner 2: Dr. Hammad Majeed
Associate Professor
FAST-NUCES, Islamabad

(c) Internal Examiner : Dr. Muhammad Tanvir Afzal
Professor
CUST, Islamabad

Supervisor Name : Dr. Muhammad Abdul Qadir
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. Qamar Mahmood (Registration No. PC103006)**, hereby state that my PhD thesis titled, '**Ontology for Inter-Research Paper Similarity Measures (CORES) and its Applications**' is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

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



(Mr. Qamar Mahmood)

Dated: 12th. September, 2019

Registration No : PC103006

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled “**Ontology for Inter-Research Paper Similarity Measures (CORES) and its Applications**” is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me.

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

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



(Mr. Qamar Mahmood)

Dated: 12th September, 2019

Registration No : PC103006

List of Publications

It is certified that following publications have been made out of the research work that has been carried out for this thesis.

Journal Papers

1. Mahmood, Qamar, Muhammad Abdul Qadir, and Muhammad Tanvir Afzal.
“Application of CORES to Compute Research Papers Similarity.” *IEEE Access* 5 (2017): 26124-26134.

Qamar Mahmood

Registration No. PC103006

Acknowledgements

First of all praise be to Allah, The most gracious, Who blessed me with the opportunity, capability and resources to pursue the doctoral programme, and it's all due to His grace that I saw it through.

One of the most notable of Allah's blessings upon me was in the form of my supervisor. PhD supervisors need a special skill; they must have night-vision eyes so that when the student gets totally lost in the pitch dark ravines of the research landscape, they can find a way and guide the lost soul to the right track. My supervisor has not only got such vision, he's got additional telescopic lens built in. I don't have words to thank you, Dr. Muhammad Abdul Qadir, for the motivation, guidance, support and encouragement that you provided to me on constant basis.

I am especially grateful to Dr. Muhammad Tanvir Afzal whose advice led me to some quick decisions that saved me a lot of precious time. He boosted my morale at every point I was feeling shaky, which was just about every other day. I am also very thankful to Mr. Fahad a student of Dr. Tanvir Afzal for providing support during my research experiment. Very special thanks to Dr. Munir Ahmad, my senior research fellow at CDSC, for his support and help. I am also thankful to other members of CDSC whose discussion and constructive criticism maintained an environment that was conducive for research. There are so many other well-wishers including friends, colleagues and relations who remembered me in their prayers. Allah bless you all

In the end I must mention the being who cares more about me than I do for myself, my mother. I always felt her prayers by my side in the time of despair and frustration, and this feeling gave me energy to put myself together. And last but not the least; I have profound gratitude for my wife who has supported me a lot during my research work for PhD.

Abstract

A large number of research papers are being published and indexed regularly by systems such as search engines, citation indexers, and digital libraries, enabling researchers to explore through these papers. Most of the users feel frustrated due to the large number of results for similar research papers with many of these results are not similar at all. A careful analysis of these systems to find similar research papers reveals a major problem that research paper based similarity measuring techniques have not been conceptually modelled to find the similarity measures with a reasonable accuracy. In order to solve this problem an ontology to model the domain of research papers similarity measures is required. While surveying content based similarity measuring techniques, it was found that these techniques were not integrated with each other, to formulate a hybrid technique without overlappings and redundancies in methods and features. We have surveyed different ontologies relevant to research paper similarity measures domain, finding that none of these were modeling this domain. In this thesis, content based similarity measuring techniques were modelled in the form of ontology named as CORES (Content based Ontology for Research paper Similarity) which has been evaluated using automated evaluation tools and user study based evaluation techniques. An important application of CORES is finding research paper similarity measures in a comprehensive by using knowledge about relationships between different similarity measuring techniques demonstrated using four use cases. An experiment was also performed on a gold standard data set of research papers to compute comprehensive similarity measures using CORES. The results of Fractional Regression Coefficient (Percentage Difference) between user study based similarity measure (as a benchmark) and comprehensive similarity measure were computed. It was found that comprehensive similarity measure was more correlated to user study based similarity measure with a value of 47% for Fractional Regression Coefficient as compared to vector space based and InText citation based similarity measuring techniques and their combinations. CORES models only the content based similarity measuring techniques, the model can be extended for other similarity measuring techniques for example Collaborative Filtering, Item Centric etc.

CORes can also be aligned with other relevant ontologies (SPAR) to enhance its adaptation by community.

Contents

Author’s Declaration	v
Plagiarism Undertaking	vi
List of Publications	vii
Acknowledgements	viii
Abstract	ix
List of Figures	xiv
List of Tables	xvi
Abbreviations	xix
1 Introduction and Research Methodology	1
1.1 Introduction	1
1.2 Problem Statement	6
1.3 Research Questions	6
1.4 Research Methodology	7
1.5 Dissertation Organization	8
1.6 Research Contributions	8
2 Literature Review for CORES	9
2.1 Survey of the Existing Ontologies	9
2.1.1 SwetoDblp Ontology	10
2.1.2 SPAR Ontologies	10
2.2 Survey of Ontology Based Semantic Similarity Measuring Techniques	11
2.3 Survey of Approaches Classifying Different Similarity Measuring Techniques	16
2.4 Conclusions	19
3 Knowledge Acquisition for Development of CORES	20
3.1 Introduction to the Acquisition Process	20

3.2	Basic Terminology about Content Based Similarity Measuring Techniques:	22
3.3	Survey of Content Based Similarity Measuring Techniques	35
3.4	Vector Space Based Similarity Measuring Techniques	36
3.4.1	Finding Concepts With Disjoint and Overlapping Relationships	36
3.4.2	Conclusions from Analysis of Vector Space Based Similarity Measuring Techniques	51
3.4.3	Survey of Vector Space Based Similarity Measuring Techniques	52
3.5	Probabilistic Similarity Measuring Techniques	57
3.5.1	Finding Concepts with Disjoint and Overlapping Relationships	58
3.5.2	Conclusions from Analysis of Probabilistic Similarity Measuring Techniques	73
3.5.3	Survey of Probabilistic Similarity Measuring Techniques	76
3.6	Citation Based Similarity Measuring Techniques	78
3.6.1	Finding Concepts with Disjoint and Overlapping Relationships	78
3.6.2	Conclusions from Analysis of Citation Based Similarity Measuring Techniques	88
3.6.3	Survey of Citation Based Similarity Measuring Techniques	89
3.7	Survey of Miscellaneous Similarity Measuring Techniques	91
3.8	Finding Relationships Between Content Based Similarity Measures	95
3.9	Conclusions	96
4	Conceptualization and Implementation of CORES	107
4.1	Introduction to Development of CORES	107
4.2	Abstract Definition of CORES	108
4.2.1	Abstract Layers of CORES	108
4.2.2	Semantic Model for Operations of Content Based Similarity Measuring Techniques	109
4.2.3	Semantic Model for Contents of Research Paper	111
4.3	Definition of CORES Using Protégé	113
4.4	SPARQL Queries for CORES	119
4.5	Ontology Metrics for CORES	120
4.6	Rules for CORES	122
4.7	Comparison of CORES With the Surveyed Ontologies	123
4.8	Conclusions	125
5	Evaluation of CORES	127
5.1	Ontology Evaluation Criteria	127
5.2	Ontology Evaluation Approaches	129
5.3	Ontology Evaluation Tools	130
5.4	User Study Based Evaluation	131
5.5	Evaluation of CORES	132
5.5.1	Evaluation metrics and methods used for CORES	132
5.5.2	Evaluation of CORES Using Ontology Evaluation Tools	135

5.5.3	User Study Based Evaluation of CORES	136
5.6	Conclusions	142
6	Comprehensive Research Paper Similarity Measure-Application of CORES	143
6.1	Different Possible Applications of CORES	143
6.2	Algorithm for Computation of Comprehensive Similarity Measure .	145
6.3	Use Cases for CORES	146
6.3.1	Use Case-1: Vector Space Based Similarity Measures Computation	146
6.3.2	Use Case-2: Probabilistic Similarity Measures Computation	149
6.3.3	Use Case-3: Citation-Based Similarity Measures Computation	151
6.3.4	Use Case-4: Comprehensive Similarity Measures Computation	153
6.4	Conclusions	154
7	Analysis of Comprehensive Research Paper Similarity Measure	156
7.1	Case Study for Experiment	157
7.2	Experimental Setup	159
7.3	Data Set Description	160
7.4	Application Built for Experiment	160
7.5	Results and Discussions	161
7.5.1	Comparison of Different Similarity Measures	161
7.5.2	Performance Analysis Using Percentage Difference Coefficient	164
7.6	Conclusions	167
8	Conclusions and Future Tasks	169
	Bibliography	172
	Appendix A Questionnaire for User-based evaluation of CORES	185
	Appendix B Algorithm for Computation of Comprehensive Similarity Measure	189
	Appendix C Different Similarity Measure values computed for a set of research papers from Gold Standard Dataset	191
	Appendix D Documentation of CORES	193

List of Figures

2.1	Classification of similarity techniques by Metzler et al [3]	17
2.2	Classification of similarity techniques by Yih et al [4]	18
2.3	Classification of similarity techniques by Bar et al [9]	18
4.1	Abstract layers of CORES Ontology	108
4.2	Semantic model for classification operations of content-based similarity measuring techniques	110
4.3	Semantic model for representation of contents and weighting schemes of a research paper	112
4.4	Top level classes in CORES as viewed in Protégé	113
4.5	Content based similarity modelling classes in CORES as viewed in Protégé	114
4.6	Research Paper content based modelling classes in CORES as viewed in Protégé	115
4.7	“Used_in” object property in CORES as viewed in Protégé	116
4.8	“Generates” object property in CORES as viewed in Protégé	117
4.9	Data properties in CORES as viewed in Protégé	117
4.10	Individuals for concepts in CORES as viewed in Protégé	117
4.11	Visualization of content based similarity methods in CORES as viewed in Protégé	118
4.12	SPARQL Query for Vector Space based similarity methods in CORES as viewed in Protégé	119
4.13	SPARQL Query for research paper weighting schemes in CORES as viewed in Protégé	120
4.14	Ontology Metrics for CORES (First View) as viewed in Protégé	121
4.15	Ontology Metrics for CORES (Second View) as viewed in Protégé	121
5.1	Hermit reasoner running on CORES without generating any errors	137
5.2	Fact++ reasoner running on CORES without generating any errors	138
5.3	Plot of User Study based evaluation of CORES for completeness	140
5.4	Plot of User Study based evaluation of CORES for accuracy and clarity	140
5.5	“Section Wise” similarity concept addition in the CORES	141
6.1	Algorithm for comprehensive similarity computation, an abstract view	145

6.2	Use Case-1 representing usage of a conceptual model of Vector Space-based Similarities from CORES	148
6.3	Use Case-2 representing usage of a conceptual model of Probabilistic Similarities from CORES.	150
6.4	Use Case-3 representing usage of a conceptual model of Citation-Based Similarities from CORES	152
7.1	Concepts and Knowledge based on content based similarities from CORES to be used for Case Study	158
7.2	Different similarity measures compared with User Study based similarity	162
7.3	Comparison of user study based similarity with different combinations involving InText Citation based similarity	163
7.4	Comparison of user study based similarity with a combination of vector space based similarity measures	164
7.5	Comparison of percentage difference values for different similarity measures	166
7.6	Comparison of percentage difference values with different combinations of similarity measures	167

List of Tables

3.1	Document Space	24
3.2	Word Space	25
3.3	Binary Vector	25
3.4	Analysis on basis of Research Paper Sections and Text Representation Schemes	38
3.5	Analysis on basis of Research Paper Sections and Text Extraction Schemes	39
3.6	Analysis on basis of Research Paper Sections and Research Paper Weighting Schemes	40
3.7	Analysis on basis of Research Paper Sections and Vector Space based Similarity Methods	42
3.8	Analysis on basis of Text Representation Schemes and Text Extraction Schemes	43
3.9	Analysis on basis of Text Representation Schemes and Research Paper Weighting Schemes	44
3.10	Analysis on basis of Text Representation Schemes and Vector Space based Similarity Methods	46
3.11	Analysis on basis of Text Extraction Schemes and Research Paper Weighting Schemes	47
3.12	Analysis on basis of Text Extraction Schemes and Vector Space based and String based Similarity Methods	49
3.13	Analysis on basis of Research Paper Weighting Schemes and Vector Space based Similarity Methods	50
3.14	Analysis using Research Paper Sections and Text Representation Schemes	59
3.15	Analysis using Research Paper Sections and Research Paper Weighting Schemes	60
3.16	Analysis using Research Paper Sections and Text Extraction Schemes	61
3.17	Analysis using Research Paper Sections and Entities Related to Research Paper	62
3.18	Analysis using Research Paper Sections and Probabilistic Similarity Methods	64
3.19	Analysis using Text Representation Schemes and Text Extraction Schemes	65
3.20	Analysis using Text Representation Schemes and Research Paper Weighting Schemes	65

3.21	Analysis using Text Representation Schemes and Research Entities related to Research Paper	66
3.22	Analysis using Text Representation Schemes and Probabilistic Similarity Methods	67
3.23	Analysis using Term Extraction Schemes and Research Paper Weighting Schemes	68
3.24	Analysis using Term Extraction Schemes and Entities related to Research Paper	69
3.25	Analysis using Term Extraction Schemes and Probabilistic Similarity Methods	70
3.26	Analysis using Research Paper Weighting Schemes and Entities related to Research Paper	71
3.27	Analysis using Research Paper Weighting Schemes and Probabilistic Similarity Methods	72
3.28	Analysis using Entities Related to Research Paper and Probabilistic Similarity Methods	74
3.29	Analysis using Research Paper Sections and Text Representation Schemes	80
3.30	Analysis using Research Paper Sections and Entities related to Research Paper	81
3.31	Analysis using Research Paper Sections and Research Paper Weighting Schemes	82
3.32	Analysis using Research Paper Sections and Citation based Similarity Methods	84
3.33	Analysis using Text Representation Schemes and Entities Related to Research Paper	85
3.34	Analysis using Text Representation Schemes and Research Paper Weighting Schemes	85
3.35	Analysis using Text Representation Schemes and Citation Based Similarity Methods	85
3.36	Analysis using Entities Related to Research Paper and Research Paper Weighting Schemes	86
3.37	Analysis using Entities Related to Research Paper and Citation based Similarity Methods	87
3.38	Analysis using Research Paper Weighting Schemes and Citation based Similarity Methods	87
3.39	Analysis of Content-Based Similarity Measuring Techniques	97
4.1	Comparison of CORES the ontology with other surveyed ontologies	124
5.1	Evaluation metrics and methods considered for CORES evaluation	133
5.2	Evaluation of CORES using Ontology evaluation tools	136
5.3	Mapping of ontology evaluation metrics with questions from user study based questionnaire	139
5.4	Findings after user study based evaluation of CORES	141

6.1	A table representing Vector Space-based document similarity techniques with gap areas	147
6.2	Pairwise Probabilistic Similarity Measures	149
6.3	Pairwise Citation based Similarity Measures	151
7.1	Performance comparison of different similarity measures with user study based similarity	165
7.2	Performance comparison of combinations of different similarity measures with user study based similarity	166
C.1	Comparison of different similarity measures with Comprehensive Similarity measure	191
C.2	Combinations of InText Citation based similarity measure with Vector Space based similarity measures.	192
C.3	Combinations of different Vector Space based Similarity Measures	192

Abbreviations

CC/IDF	Citation Count with Inverse Document Frequency
COReS	Content based Research Paper Similarity Ontology
CSCW	Computer-Supported Cooperative Work and Social Computing
DC	Dublin Core
DL	Description Logic
DoCO	Document Component Ontology
DTD	Document Type Definition
FaBiO	FRBR aligned Bibliographic Ontology
FRBR	Functional Requirement of Bibliographic Records
IDF	Inverse Document Frequency
IR	Information Radius
JCST	Journal of Computer Science and Technology
KL Divergence	Kullback–Leibler Divergence
KPEM	Key Phrase Extraction Module
LDA	Latent Dirichlet Allocation
LOD	Linked Open Data
OWL	Web Ontology Language
PMI	Pointwise Mutual Information
PRISM	Publishing Requirements for Industry Standard Metadata
RDF	Resource Description Framework
SAO	Subject Action Object
SKOS	Simple Knowledge Organization System
SPAR	Semantic Publishing and Referencing
SPARQL	Simple Protocol and RDF Query Language

TF	Term Frequency
TF/IDF	Term Frequency with Inverse Document Frequency
VSM	Vector Space Model
XML	eXtensible Markup Language

Chapter 1

Introduction and Research

Methodology

1.1 Introduction

A large number of research papers are being published (every month) and indexed by a number of systems such as Search Engines, Citation Indexers [1], Digital Libraries [2], Social Bookmarking services which enable researchers to explore through these papers. One of the typical query is to find similar and influential papers by using these services. However, users are frustrated mainly due to the availability of a huge number of results for similar research papers where most of them are not similar at all. A careful analysis of the systems to find similar research papers reveals a major shortcoming which needs to be addressed in order to develop a comprehensive research paper similarity measure with reasonable accuracy. The shortcoming is that a proper definition of research paper similarity measure does not exist, which encompasses different aspects of similarity measure in a comprehensive way. Different researchers provide different views of research paper similarity but a holistic view is missing.

This can be better understood by an analogy of blind (narrow focused) persons observing an elephant. One of them found elephant's leg and said it is like a pillar;

another observed the ears of the elephant as a fan. The third blind person found the trunk and said it was like a snake, for another one after feeling the tail it was like a rope. Actually each person defined the elephant according to the aspect she had observed and was unaware of the bigger picture representing the whole elephant. In the same sense, different researchers in the domain of research paper similarity measure concentrate on specific types of similarity measure. We need to have a comprehensive view (a big picture) which includes all the features and methods to measure similarity.

There are different methods used to find similarity measure between research papers using different features. Some of the similarity measuring techniques are used more frequently with overlapped functionalities and features while others are used less frequently. This also make it a difficult task to identify a gap in the formulation of a similarity measuring technique and then use the missing features and methods to fill that gap.

The features used for research paper similarity can be categorized into content based features and meta data based features. The textual and graphical material within a research paper represents its contents, e.g., title, keywords, abstract, citations, citations' context, authors, venue of publications, section headings, text within sections, conclusions etc. These features are used to compute different informational values by using term frequency (TF), term frequency with inverse document frequency (TF/IDF), term position, probability distribution of terms, etc., to be used to measure similarity of research papers. The similarity measuring methods can be grouped as vector space model-based, probabilistic, citation-based, semantic-based, and other techniques.

Vector space model-based similarity measuring methods use vector space model that represents documents/ research papers as vectors of terms appearing in these papers with weight values like term frequency, TF/IDF, etc. Probabilistic similarity measuring methods such as KL-divergence discussed in [3] and [4], and Point Wise Mutual Information (PMI) [5] use probability distribution lists of terms from the research papers/documents. Citation-based similarity measuring methods use

citation tags, citation context, and bibliographic reference lists from the research papers. Different types of citation-based similarity methods are: direct citation count, co-citation analysis, citation context based, and bibliographic coupling [6]. Other similarity measuring methods are visual, structural, lexical and hybrid. Visual method uses the visual layout of research papers in the form of scanned images [7]. Structural similarity uses XML layout of research papers for finding similarity between them [8]. Hybrid similarity measuring techniques represent a combination of different similarity measuring methods such as short segment based [3] and text documents based [9].

After reviewing these similarity measuring methods and features used in the methods, it can be easily concluded that there is an overlap in the use of features and the methods. Similarity measuring methods like Cosine, Jaccard, Euclidean distance, and other vectors space based methods have overlapping in their computational models. In case of research paper features, there are commonalities among different feature types like term frequency, inverse document frequency, TF/IDF, etc [10]. One can imagine the domain of research papers' similarity as a complex graph of features and methods with some frequently used features and methods as compared to others. There is a need to organize this complex graph into a meaningful structure so that one can easily identify that how to use this graph by intelligently merging multiple components (features and methods) of the graph in order to measure comprehensive similarity among research papers. The organization of the research paper similarity measures in a meaningful structure can be termed as the development of an ontology¹ for the domain of similarity measure.

In this thesis, our main focus is on the development of ontology for modeling of content-based (title, abstract, introduction, and references etc.) research paper similarity methods and the features used in these methods.

The question arises that why we need ontologies or an ontology based approach?

The reasons to develop ontology [11] are:

¹An ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpret able definitions of basic concepts in the domain and relations among them [11]

1. To share common understanding of the structure of information among people or software agents
2. To enable reuse of domain knowledge
3. To make domain assumptions explicit
4. To separate domain knowledge from the operational knowledge
5. To analyze domain knowledge

Therefore developing an ontology to model the domain of research paper similarity measure would help us to achieve the objective of common understanding by people and machines in order to measure overall similarity between the research papers.

Recall the definition of similarity as the measure of likeliness in the features of multiple research papers; there are many features which can be used to measure the similarity. How all the similarity measures can be combined by exploring relationships between them to compute overall document similarity, is not known. There is no precise and complete conceptual model available using which one can compute the research paper similarity comprehensively. If the terms like similarity measuring methods and its sub-categories are classified in the form of ontology, it would become easier for people to understand and use these constituent concepts effectively. If the knowledge base of this ontology (instances) contains different similarity measures which use different features and methods, then an approach to comprehensively integrate all these methods can be devised intelligently.

To develop the ontology for similarity measures, we have adopted a well-structured methodology to develop ontology (Methontology) [12], in which ontology is developed from scratch. The proposed ontology has been named as CORES that stands for Content-based Ontology for Research papers' Similarity. The ontology was evaluated by using automated evaluation tools ODEVAL [13], Hermit reasoner [14], Fact++ [15], and user study based evaluation methods. The evaluation process looked for inconsistency, incompleteness, and redundancy errors [12], their

extensions [16], and evaluation metrics such as accuracy, clarity, and adaptability [17] in CORES. CORES is available as an OWL file (CORES.owl) at following link:

<https://github.com/QamarPC103006/COREs>

An important application of CORES is the computation of comprehensive research papers' similarity measure in a comprehensive way which was illustrated by presenting four use cases. In these use cases different similarity measuring techniques were used in a combined way represented as a weighted sum of vector space based, probabilistic, and citation based similarity measuring techniques modeled in CORES. We have used the knowledge of disjoint and overlap relationships between the similarity measuring techniques (conceptualized in the CORES) for the computation of similarity.

To verify our claim of computing comprehensive research paper similarity using CORES, we have performed an experiment on a gold standard data set of research papers. This data set contained 72 query papers along with 3 to 8 reference papers for each of the query paper, and 368 pair wise combinations of these research papers. This data set already contained the inText citation based and user study based similarity measures among the query papers and their reference papers.

In our experiment we have computed Cosine, Jaccard, and Euclidean distance similarity measures among the research papers. Different binary combinations of these similarity measures were also calculated. Comprehensive similarity measure was computed by combining inText citation based and average vector space based similarity measures by using knowledge about their (overlapping and disjoint) relationships from CORES. User study based similarity measure was used as a benchmark for performance analysis of different similarity measures. For performance analysis fractional regression coefficient (percentage difference) [18] was used. It was found that performance of comprehensive similarity measure was significantly better (with a percentage difference of 47%) than the top performing similarity measure i.e. inText citation based similarity from the data set by use of

fractional regression coefficient. These results were supporting our claim of conceptualization of domain of research paper similarity and computing comprehensive similarity using knowledge.

1.2 Problem Statement

There is a lot of overlappings and redundancies in the research paper similarity measuring techniques as represented in Table 3.39. Multiple similarity measuring techniques are using the same features to find the similarity, for example Cosine, Jaccard, Euclidean, Pearson correlation based similarity measures use same vector space model with a common feature TF/IDF in majority of the published techniques [10]. Important features like term position, phrase depth, common citation/inverse document frequency (CC/IDF) are missing in the computation of these similarity measuring techniques. The vector space based techniques also have commonalities in their similarity measure computational methods such as Cosine, Jaccard, and Euclidean distance techniques use fractional formulas (Table 3.39) which have common values either in their numerators or denominators.

There is no classification of the similarity measuring techniques available in literature, which is based upon the features and methods used. It is not known that how to combine the similarity measures computed by different similarity measuring techniques in order to compute the overall similarity measure among the research papers. We believe that the reason behind all these problems is the non-availability of an ontology of similarity measures among research papers.

1.3 Research Questions

The problem discussed raises at least following research questions:

1. What are the features in a research paper which can be used to find similarity measures? How can these features be classified into logical groupings?

2. What are the methods to compute the similarity measures? How can these methods be classified into logical groupings?
3. Can we organize the features and methods in the form of an ontology for research paper similarity measures?
4. Can the missing features and methods be worked out to develop new techniques to find similarity measures?
5. How the developed ontology is going to be helpful in developing a comprehensive similarity measure by combining and giving weights to different features and methods?

1.4 Research Methodology

We have adopted ontology development methodology (Methontology) [12] as our research methodology, comprising of following steps which will answer the research questions raised in the previous section.

1. Specification: Scope of ontology and scenarios are required to be specified, which is already provided in section 1.1.
2. Literature review for existing ontologies was performed for any possibility of modeling of research paper similarity measures domain.
3. Knowledge Acquisition: Content based similarity measuring techniques were analyzed to find overlapping and disjoint relationships for conceptualization of research paper similarity measures domain.
4. Conceptualization and Implementation: Formation of a comprehensive ontology for similarity measures among research papers using ontology development techniques.
5. Evaluation: Evaluation of the developed ontology using evaluation tools and user study based evaluation.

6. Evaluation: Demonstrate the working of the similarity measuring system by using the proposed ontology to evaluate its adaptability.
7. Publish the results.

1.5 Dissertation Organization

This thesis is organized in the following chapters. Chapter 2 surveys existing ontologies relevant to the proposed ontology and semantic similarity techniques based on ontologies. Chapter 3 describes a detailed analysis of content based research paper similarity techniques in an organized way for knowledge acquisition of domain of research paper similarity measures. This chapter also presents the research gap analysis for surveyed techniques using features and methods from the domain of research paper similarity measures. Chapter 4 presents the conceptualization and implementation of CORES, the proposed ontology. This chapter presents semantic models for features of research papers and for content based similarity measuring techniques. Chapter 5 discusses evaluation of CORES using different ontology evaluation techniques; such as automated tool based and user study based evaluation. Chapter 6 provides application of CORES demonstrated by four use cases to evaluate its adaptability. Chapter 7 describes experimental analysis of comprehensive research paper similarity measure computation using knowledge from CORES. Chapter 8 concludes this thesis along with a number of future tasks.

1.6 Research Contributions

1. Knowledge acquisition from the research paper content based similarity measures for modeling the domain of research papers similarity measures.
2. Conceptualization and development of ontology, CORES, for research papers' similarity measures and evaluation of CORES.

3. Compute a comprehensive research paper similarity measuring technique integrating all content based similarity measuring techniques using knowledge from COrES.

Chapter 2

Literature Review for CORES

When we need to develop an ontology [11], we have to survey the existing ontologies to check if these are modeling the required domain. Therefore in this chapter we will survey different ontologies to see if these ontologies model the domain of research paper similarity measures? Ontologies related to domain of semantic publishing and digital libraries were surveyed in this chapter as they seem more relevant to domain of research paper similarity measures. We have also surveyed ontology based semantic similarity measuring techniques to answer a question that whether these techniques provide a framework for combining different similarity measuring methods? We have also reviewed the literature presenting classifications of different similarity techniques.

2.1 Survey of the Existing Ontologies

No such ontology available in the literature which models the domain of research paper similarity measures. Therefore, we have reviewed those ontologies which involved conceptual modeling of contents of research papers and will be useful for computation of research paper similarity measures. These ontologies are SwetoDblp ontology [19] and ontologies published under a group Semantic Publishing and Referencing (SPAR) [20]. SwetoDblp provides an RDF based model of the DBLP

data set, a digital library of research papers. The focus of SPAR ontologies is on modeling the structure of a scientific document (DoCO), reference lists appearing in these documents (FaBiO) [21] and citation information about these document (CiTO) [21].

2.1.1 SwetoDblp Ontology

An RDF-based ontology [19] named as SwetoDblp is presented in this section, which is populated from DBLP data set [22]. This is a shallow ontology [23] as it comprises of a few terms and it organizes the very large amount of data about research papers stored in DBLP data set. This ontology models the concepts of research papers, their authors social media related information, and meta data of research papers. In SwetoDblp, concepts related to research papers are not further conceptualized according to rhetorical and structural sections. Further, there is no conceptualization of research paper similarity measures and similarity measuring techniques in SwetoDblp.

2.1.2 SPAR Ontologies

Semantic publishing of scientific documents is conceptualized by a set of ontologies called SPAR (Semantic Publishing and Referencing) ontologies. During the development of SPAR ontologies [20], previous work related to the structural and rhetorical conceptualization of scientific documents was analyzed thoroughly to address the structural and rhetorical components of a document. For this purpose, Ontology named DoCO (Document Component Ontology) was introduced. It provides [20] a prearranged vocabulary of document components, both structural (e.g. Block, Inline, container), rhetorical (e.g. Introduction, discussion, acknowledgements, reference list, figure, appendix) and mixed (e.g., paragraph, section, chapter), enabling these components, and documents composed of them, to be described in RDF. The advantage of DoCO is to import different relevant

ontologies such as DEO, Pattern Ontology, and SALT ontology making it rich in structure.

FaBiO is [21] FRBR aligned Bibliographic Ontology for recording and publishing the bibliographic records appearing in the reference section of a scientific document. In particular, DC (Dublin Core) Terms [24], PRISM (Publishing Requirements for Industry Standard Meta data) [25], FRBR (Functional Requirement of Bibliographic Records) [21] and SKOS (Simple Knowledge Organization System) terms [26] are all included in FaBiO. DC terms and vocabularies are most commonly used for modeling of cataloging resources. PRISM is a specification defining a rich set of meta data terms for describing published work. There are different similarity measuring techniques which use bibliographic resources such as references to research papers for finding similar papers.

CiTO (Citation Typing Ontology) presents semantic modeling of citation reasons, which were not previously provided in ontologies like SWAN (Semantic Web Applications in Neuromedicine) and BIBO (Bibliographic Ontology). CiTO holds just two main object properties, cito:cites and its inverse cito:isCitedBy, each of which has 32 sub-properties. CiTO also has two generic object properties to make statements linking two entities that do not constitute formal citation acts: cito:shareAuthorsWith and cito:likes, the latter authorizing social media 'likes' statements to be encoded in RDF. Rhetorical properties of CiTO are being assembled in positive, informative (neutral), and negative.

SPAR ontologies were not found to conceptualize any research paper similarity measures and similarity measuring techniques.

2.2 Survey of Ontology Based Semantic Similarity Measuring Techniques

In this section, we have surveyed ontology based semantic similarity measuring techniques to check whether these techniques offer any framework or ontology

for domain of research paper similarity measures? These techniques have been surveyed to find the answers for the following questions.

1. Does the ontology base semantic similarity measuring techniques are focused on intra ontology or inter ontology similarity?
2. Does the technique is finding similarity measures among research papers or other documents/entities?
3. How this technique works?
4. Which data set does a technique uses or any case studies about usage are presented?
5. How the technique is evaluated on the basis of performance parameters such as: accuracy, precision, f-measure etc.
6. Does the technique provide any semantic model or framework to combine different similarity measuring techniques?

The reason for surveying these techniques using above questions is that we need to analyze the semantic similarity measuring techniques which may be using different ontologies, to see any possibility of modeling of domain of research paper similarity measures.

This research [27] is about a similarity technique to be computed between different concepts. The technique focuses on semantic relationships between the concepts. Ontology based similarity measures are better as the knowledge from ontology is helpful to find unclear relationships between the concepts. It is not clear that whether the proposed technique is intra ontology similarity or inter ontology similarity? The technique uses TF/IDF measures which mean the technique was used for finding similarity measures among the documents. The data set used by experiment are mini newsgroups which is a subset of Yahoo!. The experiment performed in this work uses the Wine ontology available from W3C. The proposed technique was performing text document classification using ontology

based approach. Precision and Recall were considered as performance measure in this experiment. Precision was found more important in order to get the better results. This paper does not provide any semantic model or framework to combine different similarity measuring techniques.

This technique [28] finds the best semantic similarity measure from different available measures. These semantic similarity measures are used to find similarities between the concepts within a single ontology. Author has found the semantic similarity measure which can provide good results with less error rate. The experiment was done on taxonomy of concepts from health domain (ICD10 Taxonomy) not on research papers or general documents. Different semantic similarity measuring techniques were used which perform distance based measures on this taxonomy. The similarity between 30 pairs from health domain has been evaluated using different types of semantic similarity measure equations. This research does not present any semantic model or framework to combine different similarity measuring techniques.

This work [29] proposed a similarity comparison algorithm for documents using ontologies. Authors have used ontologies as graph based model to reflect semantic relationships between concepts and use them for text analysis and comparison. This technique focuses on a single ontology and extraction of sub ontologies. Instead of performing raw document comparison, documents are enhanced using concepts from an ontology. Due to this enhancement documents which were not previously similar may become similar to some extent. This technique is demonstrated to find similarity between general text documents rather than research papers. The use of this technique was demonstrated using only a case study not by any sort of data set and this technique was not evaluated. This work does not provide any semantic model or framework for combining different similarity measuring techniques.

Estimation of semantic relatedness [30] between the terms is an important task to interpret textual data. In this paper, authors have surveyed and classified different ontology based semantic similarity measuring approaches in order to evaluate

their advantages and limitations. They have also compared the performance of these approaches from practical and theoretical point of view. A new ontology based similarity technique is also introduced which uses taxonomical features. A common framework was used to compare the performance of newly proposed approach with other approaches from related work. Proposed approach was found more accurate than the other approaches. The proposed techniques and discussed semantic similarity techniques focused on finding similarity between the concepts of a single ontology. The proposed technique used ontology of concepts rather than documents/research papers to perform its tasks. In this paper different ontology based semantic similarity measures are discussed and these techniques are classified as: edge counting approaches, feature based measures, measures based on information contents. The experiment in this paper uses Miller and Charles and Rubenstein and Goodenough benchmarks. The results of experiment report correlation for different semantic similarity techniques using these benchmarks. This research does not introduce any semantic model or framework for combining different similarity measuring techniques.

A significant and growing amount of semantic data [31] is published on Web as linked open data (LOD). SPARQL provides querying and searching mechanisms for Linked Open Data (LOD), which can help in development of a vast range of semantic web applications. SPARQL is unable to compare, prioritize or rank the search results from queries of users. The examples of systems which provide such search results are recommender systems, matchmaking, and social network analysis. This paper solves this problem related to SPARQL by designing a systematic model to measure semantic similarity between resources in a linked open data. Authors proposed a generalized information content based approach. This approach was previously less perceived in context of usage for linked open data. Authors have validated and evaluated this similarity measure in a recommender system. Authors claim that their approach outperforms the recommender systems using conventional approaches. It is not clear that whether the approach uses a single linked open data set or an interlinked set of LODs. This research does

not introduce any semantic model or framework for combining different similarity measuring techniques.

This paper introduces a new and efficient model [32] for representation of taxonomies called PosetHERep, which is an adaptation of half-edge data structure used for planner graphs. The paper also introduces a new Java library named as Half-Edge Semantic Measure Library (HESML) based on PosetHERep. This library implements most ontology based semantic similarity measures and Information Content models are reported in the literature. This research also provides a set of reproducible experiments on word similarity based on HESML and ReproZip. A replication framework and a data set named as WNSimpRep V1 were also produced by authors to support the replication of methods reported in literature. The proposed approach was adopted due to drawbacks in current semantic measuring libraries, specifically in terms of performance and scalability. PosetHERep proposes a memory efficient representation of taxonomies by providing most taxonomy based algorithms used by semantic measures and Information Content models. This paper has also classified the ontology based semantic similarity measures into different categories. This research work is not about computing any new similarity measure among the documents/research papers. Performance of semantic similarity measures implemented using this technique was found better than the conventional semantic similarity measuring techniques.

This paper provides a framework [33] for unification of different ontology based semantic similarity measures available in the literature. This framework attempts to answer questions such as: which similarity measure should be selected for a concrete application? Are different similarity measures equivalent to each other? Authors have performed an in-depth analysis of different existing semantic similarity measures to identify their core elements. This framework have been tested for hundreds of semantic similarity measures in biomedical context. According to this paper different similarity measures are designed according to specific paradigms. From a paradigm, estimators of commonalities and differences will be defined. These estimators are considered as roots of all existing similarity measures. This

paper provides theoretical definition of a framework dedicated to semantic similarity measures. The similarities tested in this paper used different data sets/ontologies from medical domain like SNOMED-CT (Systematized Nomenclature of Medicine – Clinical Terms). The framework focused on techniques which are used find similarity measures between concepts rather than documents/research papers.

2.3 Survey of Approaches Classifying Different Similarity Measuring Techniques

A number of authors have categorized different content based similarity measuring techniques in their work. Although the techniques they have discussed were not been used to compute similarity measures between research papers. Most of these techniques were used to compute similarity measures between short text segments. These techniques are discussed in this section.

According to [3], short text segment based similarity measures can be computed using different similarity measuring techniques. Authors have categorized similarity techniques into three categories: lexical, probabilistic and hybrid. The most basic similarity measuring techniques are lexical similarities; these techniques are based on matching the terms in surface representation of text. For further classification of these similarity measuring techniques, authors have used categories of text representation. These categories are surface text, expanded text, and stemmed text based similarities. It was observed that authors have classified these similarity measuring techniques according to their perception to find similarities between the short text segments. Figure 2.1 represents a classification of similarity measuring techniques as discussed in [3].

As hybrid techniques uses multiple similarities measuring techniques to get matching results and only the high ranking results are selected. One of these authors has further improved these techniques as discussed in [4]. In this paper, authors

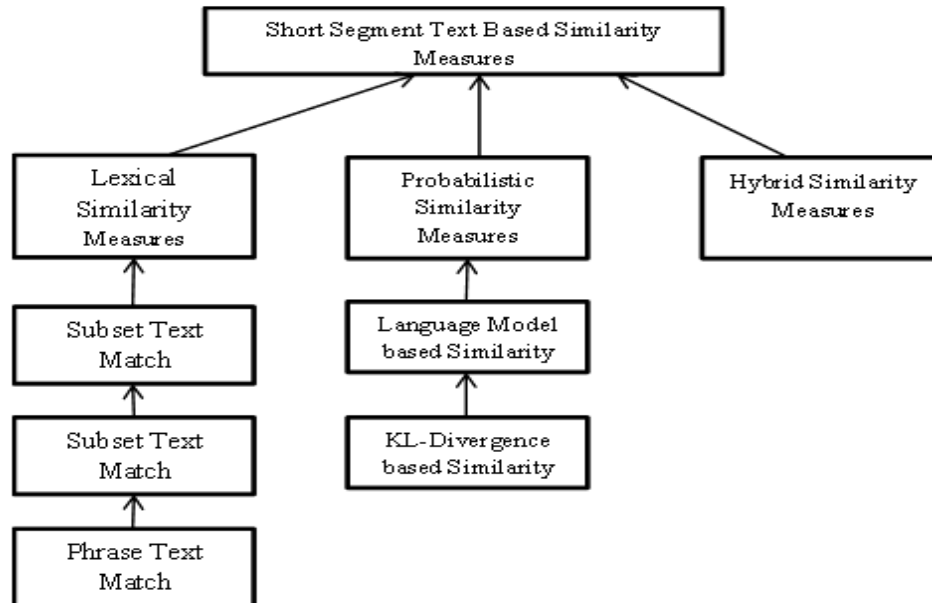


FIGURE 2.1: Classification of similarity techniques by Metzler et al [3]

have described an extended web similarity based ranking scheme [34] using a better term weighting scheme. The web-based similarity ranking function uses term frequency and document frequency to measure the importance of terms in the expanded representation of input text segments. Figure 2.2 represents a similarity hierarchy according to perception of authors of this paper.

A system called UKP is proposed [9] to compute semantic text similarity by combining different content based similarity measuring techniques. Authors have classified different text-based similarity measuring techniques used in their system. The categories of string based similarity measures such as Longest Common Subsequence [35], Greedy string tiling [36], Jaro, Jaro-Winkler, Monge and Elkan, Levenshtein, and Character/Word n-grams were discussed. Another category for text similarity measure is defined as semantic similarity measure with categories such as Pairwise word similarity, Explicit Semantic Analysis, Textual Entailment, and Distributional Thesaurus. Figure 2.3 shows a classification of different similarity measures as discussed by authors in this paper.

After surveying these techniques, we conclude that different authors have classified/grouped the similarity measuring techniques in different ways according to

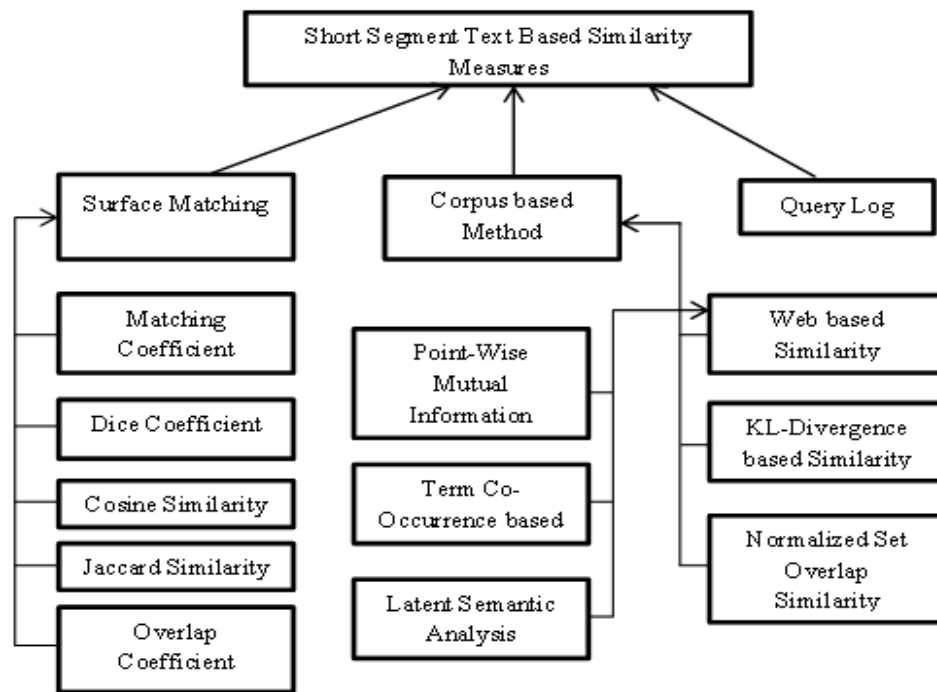


FIGURE 2.2: Classification of similarity techniques by Yih et al [4]

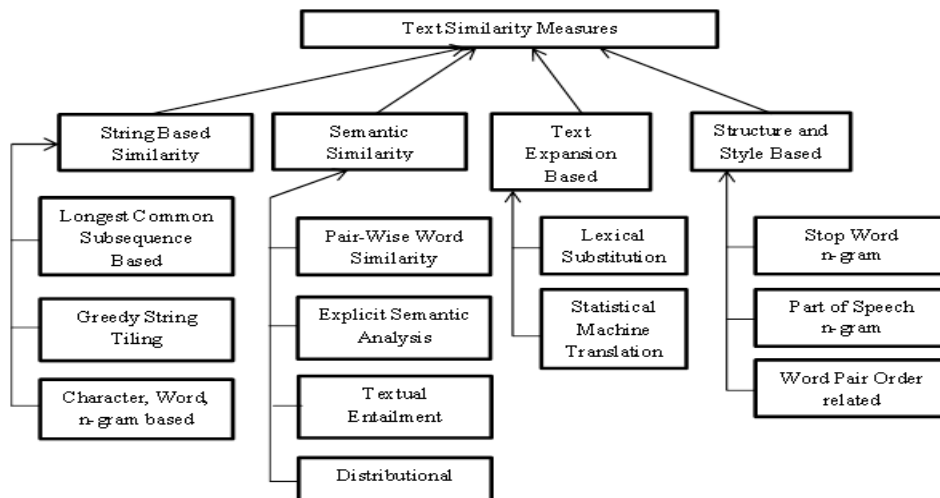


FIGURE 2.3: Classification of similarity techniques by Bar et al [9]

their understanding and experimental needs. Therefore there is a need for a comprehensive classification model of similarity measuring techniques.

2.4 Conclusions

After surveying relevant ontologies, semantic similarity measuring techniques and authors' perceived classifications of similarity measuring methods, we have reached to a number of conclusions. No such ontology available in the literature which models the domain of research paper similarity measures. Proposed ontology (CORS) models the domain of research paper similarity measures. Different semantic similarity measuring techniques were also surveyed and it was found that none of them except one [33] provide any framework to combine different similarity measuring methods. The cited approach which contains a framework was focusing on combining different ontology based semantic similarity measuring techniques rather than research papers based similarity measuring techniques. After surveying approaches, in which authors have classified different similarity measuring techniques according to their understanding and experimental needs, we concluded that each author have their own classification hierarchy of these techniques and there is lack of a comprehensive classification for similarity measuring techniques in the literature.

Chapter 3

Knowledge Acquisition for Development of CORES

3.1 Introduction to the Acquisition Process

To provide an ontology for the domain of research paper similarity measures, we need to acquire the knowledge of this domain. Since it is a very broad domain we have narrowed down our focus by selecting content based similarity techniques, the most dominant category in this domain [10]. The knowledge acquisition task was accomplished by adopting following steps:

1. 67 research papers from the field of content based research paper similarity measuring techniques were collected.
2. These papers were surveyed for any possibility of modeling of domain of the research paper similarity measures.
3. Features of research papers and similarity measuring methods using these features, were analyzed to identify disjoint and overlapping relationships between the similarity measuring techniques to be modeled.
4. This knowledge of relationships between similarity measuring techniques was used in proposed ontology (CORES).

5. A task of research gap analysis for the surveyed similarity measuring techniques was performed as a secondary activity of the knowledge acquisition.

The similarity measuring methods used in the similarity measuring techniques can be categorized as vector space-based, probabilistic, citation based, structural, visual, and lexical. These methods used different features of research papers to find similarity measures between them. A number of vector space based methods are Cosine, Jaccard, Dice, and Overlapping coefficients [37] etc. In these methods features like term frequency, term frequency-inverse document frequencies (TF/IDF) etc. are used. Probabilistic similarity measuring methods [3, 37] use probability distribution of words extracted from the research papers. Two of these methods used term co-occurrence sets built from different sources of research paper repositories. Some examples of Probabilistic similarity methods are KL-Divergence, Average KL-Divergence, and Point Wise Mutual Information (PMI) [5, 38]. In the case of citation based similarity measuring methods: citation tags, citation context information, and bibliographic lists (which are represented as References section) from research papers are used. Many of the citation-based similarity measuring approaches use citation graphs of research papers. Examples of the citation based similarity measuring methods are citation count, co-citation analysis, and bibliographic coupling [6]. Lexical similarity measuring methods (edit distance) use string based text information from the research papers. There are methods which use structural layout information of research papers to find the similarity among them. Structural similarity measuring methods use XML layouts of research papers. Visual similarity measuring methods use the visual layout of research papers in the form of their scanned images.

Content based similarity measuring techniques have been classified by researchers according to their experimental needs and understanding [3, 9], however, there is no comprehensive classification from the nature of the algorithms and feature used (**operational semantics classification**) available. Further different content based similarity measuring techniques use different weighting schemes of the

research papers. Therefore, a survey of different content based similarity measuring techniques and a model for classification of content based similarity measuring methods have been presented from the (**semantics-of-their-operations**) point of view in this chapter.

This survey was accomplished to achieve the objective of formulation of an comprehensive scheme (ontology) to find similar papers from a huge repository of research papers in order to provide the answers to the following questions:

1. Can we categorize a particular concept related to research paper features and similarity measuring methods as disjoint or overlapping, by analyzing content based similarity measuring techniques?
2. Is there an ontology (semantic model) available for a given similarity measuring technique based on the operational semantics and features used?
3. How can a particular similarity measuring technique be integrated with other techniques into the formulation of a hybrid similarity measuring technique without redundancies and overlapping in the methods and features?

3.2 Basic Terminology about Content Based Similarity Measuring Techniques:

Before starting the survey of content based similarity measuring techniques, we will introduce reader with basic terminology from this domain. For this purpose, we have provided definitions of basic terms from domain of content based similarity measuring techniques in this section.

Vector Space Model: Vector space model or term vector model is an algebraic model for the demonstration of text documents (and any objects, in general) as vectors of identifiers for example: index terms. It is used in information filtering, information retrieval, indexing, and relevancy rankings.

Document Vector: Documents and queries are denoted as vectors in equation (3.1) and (3.2).

$$d_j = (w_{\{1,j\}}, w_{\{2,j\}}, \dots, w_{\{t,j\}}) \quad (3.1)$$

$$q = (w_{\{1,q\}}, w_{\{2,q\}}, \dots, w_{\{n,q\}}) \quad (3.2)$$

where d_j and q are document vectors and $w_{\{1,j\}}, w_{\{2,j\}}, \dots, w_{\{t,j\}}$ and $w_{\{1,q\}}, w_{\{2,q\}}, \dots, w_{\{n,q\}}$ are weights of terms appearing in document d_j and q respectively.

Term Frequency (TF): The number of times a term appears in a document. Term frequency is used as weight in term vector for a document in Vector Space model.

Document Frequency (DF): Document frequency df_t represents the number of documents in a collection that contain a term t . This measure provides a collection wide statistics of term t , which is helpful in discriminating between the documents for purpose of scoring.

Inverse document frequency (IDF): If N is total number of documents in a collection and t is a term used in the document, then inverse document frequency is represented by following equation (3.3)

$$idf_t = \log \frac{N}{df_t} \quad (3.3)$$

IDF of a rare term is high whereas the IDF of a frequent term is likely to be low.

TF/IDF: Definitions of Term Frequency and Inverse Document Frequency can be combined to produce a composite weight for each term in each document. TF/IDF assigns a term t a weight in a document d using following equation (3.4).

$$tf-idf_{t,d} = tf_{t,d} \times idf_t \quad (3.4)$$

CC/IDF: Cite seer uses common citations to make an estimation of which documents in the downloaded database of research papers are the most closely related to a document picked by the user. This measure, “Common Citation Inverse Document Frequency” (CC/IDF) is analogous to the word oriented TF/IDF word weights. As in the use of TF/IDF, CC/IDF assumes that if a very uncommon citation is shared by two documents, this should be weighted more highly than a citation made by a large number of documents.

Term position: In which field of input of a research paper, whether a term of query appears. Many fields are considered more important than others. The terms of query which appear in important fields of research paper should be ranked better.

N -gram: The words in a text document with N number of characters in it, where N is an arbitrary integer value. Specific cases of N -gram are unigram (single character based), digram (two character based), trigram (three character based) text.

Document space: The matrix in following table 3.1 represents words as vectors in a document space. Each row of the table represents a document and columns represent terms which are required to be checked for their presence in these documents. A cell value 1 represents the presence of a term in a document and 0 represents its absence.

TABLE 3.1: Document Space

	cosmonaut	astronaut	moon	car	truck
d_1	1	0	1	1	0
d_2	0	1	1	0	0
d_3	1	0	0	0	0
d_4	0	0	0	1	1
d_5	0	0	0	1	0
d_6	0	0	0	0	1

Word space: The matrix in following table 3.2 represents words as vectors in word space. An entry in this matrix contains the number of times a word in

a column co-occurs with word in a row. The numeric value in cells of matrix represents the number of co-occurrences of a word/term.

TABLE 3.2: Word Space

	cosmonaut	astronaut	moon	car	truck
cosmonaut	2	0	1	1	0
astronaut	0	1	1	0	0
moon	1	1	2	1	0
car	1	0	1	3	1
truck	0	0	0	1	2

Co-occurrence: Words are similar to the extent that they co-occur with the same words. The table 3.1 for document space represents the co-occurrence of a term within multiple documents. The figure for word space represents the co-occurrence of a term with other term in the documents observed.

Binary vector: The binary vectors contain entries containing either 1 or 0. The simplest way to define a binary vector is as the set of dimensions on which it has non zero values. For example vector for cosmonaut in following table 3.3 can be represented as the set {Soviet, Spacewalking}

TABLE 3.3: Binary Vector

	cosmonaut	astronaut	moon	car	truck
Soviet	1	0	0	1	1
American	0	1	0	1	1
spacewalking	1	1	0	0	0
red	0	0	0	1	1
full	0	0	1	0	0
old	0	0	0	1	1

Matching coefficient: The matching coefficient simply counts the number of dimensions on which both vectors are non-zero. For binary vectors X and Y this coefficient is represented by following equation (3.5).

$$X \cap Y \tag{3.5}$$

Dice coefficient: The Dice coefficient normalizes for length by dividing by the total number of non-zero entries for binary vectors. We multiply by 2 so that we get a measure that ranges from 0.0 to 1.0 with 1.0 indicating identical vectors. For binary vectors X and Y it is represented by following equation (3.6).

$$\frac{2|X \cap Y|}{|X| + |Y|} \quad (3.6)$$

Jaccard coefficient: The Jaccard coefficient penalizes a small number of shared entries (as a proportion of all non-zero entries) more than the Dice coefficient does. Both measures range from 0.0 (no overlap) to 1.0 (perfect overlap), but the Jaccard coefficient gives lower values to low-overlap cases. For binary vectors X and Y this coefficient is represented by following equation (3.7).

$$\frac{|X \cap Y|}{|X \cup Y|} \quad (3.7)$$

Overlap coefficient: The Overlap coefficient has the flavor of a measure of inclusion. It has a value of 1.0 if every dimension with a non-zero value for the first vector is also non-zero for the second vector or vice versa. For binary vectors X and Y this coefficient is represented by following equation (3.8).

$$\frac{|X \cap Y|}{\min(|X|, |Y|)} \quad (3.8)$$

Cosine coefficient: The cosine is identical to the Dice coefficient for vectors with the same number of non-zero entries, but it penalizes less in cases where the number of non-zero entries is very different. For binary vectors X and Y the cosine coefficient is represented by following equation (3.9).

$$\frac{|X \cap Y|}{\sqrt{|X| \times |Y|}} \quad (3.9)$$

Surface text: Surface representation of text is text of a document itself. The

surface text of documents/research papers is commonly used by different similarity measuring techniques. Surface text usage in the similarity measures may cause vocabulary mismatch problem which may reduce the recall for a similarity measuring method.

Expanded text: When relevant text from outer source is appended with text of a document, it is called the expanded text. Expanded text is used to overcome the vocabulary mismatch problem.

Stemmed text: Stemming is most common way to generalize a text. It is commonly used in information retrieval systems to overcome the vocabulary mismatch problem. Stemming may cause a problem to introduce noise in the text which may cause difficulties in computing the similarity measures.

Vector Space based similarity techniques: Such similarity techniques which use vector space model of the documents to find similarity between them. In vector space model, a document is represented by a vector containing features for different dimensions of that document. Normally a dimension represents a term that appears in the document. Common vector space model based similarity measuring techniques are Cosine, Jaccard, and Euclidean distance similarity measures, etc.

Cosine similarity: When documents are presented as vectors, the degree of similarity measure between the two documents is measured as correlation between their corresponding vectors. This can be further quantified as cosine of the angle between the two vectors. Given two documents \vec{t}_a and \vec{t}_b , their cosine similarity measure is represented in equation (3.10):

$$SIM(\vec{t}_a, \vec{t}_b) = \frac{\vec{t}_a \cdot \vec{t}_b}{|\vec{t}_a| \times |\vec{t}_b|} \quad (3.10)$$

where \vec{t}_a and \vec{t}_b are m -dimensional vectors over the term set $T = \{t_1, \dots, t_{mm}\}$. Each dimension represents a term with its weight in the document, which is non-negative.

L-Cosine similarity: According to authors [39], each structural part of a document/research paper (such as title, abstract, etc.) contributes with a different importance. Therefore cosine similarity metric is not applied on whole text of the document but applied to each field (in this case title and abstract) separately. This is named as linear combination of different cosine values or L-Cosine which is computed using following equation (3.11).

$$L - Cosine(i, j) = \alpha \times Cosine(t_i, t_j) + \beta \times Cosine(a_i, a_j) \quad (3.11)$$

Soft Cosine similarity: The traditional cosine similarity measure ignores similarity between features of two vectors to compute similarity between them. Soft Cosine similarity measure uses similarity weights of feature vectors to compute similarity between them. The similarity measure between features of two vectors can be found using stemming in Natural Language Processing (NLP) technique.

Cosine index: It is also called Salton's cosine index. It is computed as ratio of items contained in the documents i and j , normalized by square root of product of items from document i and j represented by following equation (3.12).

$$Cosine\ Index = \frac{items_{ij}}{\sqrt{items_i \cdot items_j}} \quad (3.12)$$

Jaccard similarity: This similarity measure uses intersection of two objects divided by their union. Following equation (3.13) represents Jaccard similarity.

$$SIM_J(\vec{t}_a, \vec{t}_b) = \frac{\vec{t}_a \cdot \vec{t}_b}{|\vec{t}_a|^2 + |\vec{t}_b|^2 - \vec{t}_a \cdot \vec{t}_b} \quad (3.13)$$

Jaccard index: This index is calculated as the ratio of items (e.g. words, citation, etc.) being contained in documents i and j , normalized by the sum of the items in documents i and j minus the nominator as shown in equation (3.14):

$$Jaccard\ Index = \frac{items_{ij}}{items_i + items_j - items_{ij}} \quad (3.14)$$

Inclusion Index: The Inclusion Index takes into account the common items between two documents based on the minimum number of items from document i or j , represented by following equation (3.15) :

$$Jaccard\ Index = \frac{items_{ij}}{\min(items_i, items_j)} \quad (3.15)$$

Pearson Correlation similarity: Pearson's correlation coefficient is another measure of the extent to which two vectors are related. There are different forms of the Pearson correlation coefficient formula. Given the term set $T = \{t_1, \dots, t_{mm}\}$, a commonly used form is presented below in equation (3.16).

$$SIM_P(\vec{t}_a, \vec{t}_b) = \frac{m \sum_{t=1}^m w_{t,a} \times w_{t,b} - TF_a \times TF_b}{\sqrt{\left[m \sum_{t=1}^m w_{t,a}^2 - TF_a^2 \right] \left[m \sum_{t=1}^m w_{t,b}^2 - TF_b^2 \right]}} \quad (3.16)$$

Edit distance similarity: Edit distance is a way of quantifying how dissimilar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other. The similarity measuring technique which uses edit distance to find the similarity between the string values is called edit distance similarity.

Research paper sections (Meta data of Research paper): Different sections of a research paper are Title, Author Names, Affiliations, Abstract, Introduction, Related Work, Results and discussions, conclusions etc. These are also called meta data of research paper. Different similarity measuring techniques use these sections to measure similarity among the research papers.

Entities outside the research paper: Such contents or material which is not directly part of a research paper but represents important information related to the paper. Examples of such entities are citation graphs of research papers, social bookmarking tags for research papers, user's reviews about research papers, etc. These entities are useful to measure similarity among the research papers.

Bag of Words: The bag-of-words model is a simplifying representation used in natural language processing (NLP) and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag (multi set) of its words, disregarding grammar and even word order but keeping multiplicity.

Recommender system: A recommender system or a recommendation system is a subclass of information filtering system that seeks to predict the “rating” or “preference” a user would give to an item.

Recency of research papers: This is a metric based on date of publication of a research paper. It provides information about a research paper that how recently it has been published.

Probabilistic similarity techniques: Such similarity measuring techniques, which consider documents as probability distribution of terms are called probabilistic similarity techniques. These similarity measuring techniques are computed by measuring distance between probability distribution of terms of two documents. Normally these similarity measuring techniques are used for document clustering.

Topic Distributions: According to probabilistic model based representation of documents, a document can be represented by a set of topics. The terms appearing in a document can be associated with a topic on the basis of probability score.

Explicit Semantic analysis (ESA): A vector based representation of text (which can be individual words or complete documents) that uses document corpse as a knowledge base. In this scheme, a word is represented as a column vector in TF/IDF matrix of the text/document corpse. A document is represented as centroid of the vectors representing its words. Most commonly used text corpse for this scheme is Wikipedia.

Topic Model: A topic model is a type of statistical model for determining the abstract “topics” that occur in a collection or group of documents. Topic modeling is a commonly used text-mining technique for discovery of hidden semantic structures in text body of documents.

Latent Dirichlet Allocation (LDA): Latent Dirichlet allocation (LDA) is a statistical model that allows sets of observations to be explained by unobserved groups that explain why other parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics.

KL (Kullback Leibler)-Divergence: In information theory based clustering, a document is considered as a probability distribution of terms. The similarity of two documents is measured as the distance between the two conforming probability distributions. In the document scenario, the divergence between two distributions of words is represented in equation (3.17):

$$D_{KL} \left(\vec{t}_a \parallel \vec{t}_b \right) = \sum_{t=1}^m w_{t,a} \times \log \left(\frac{w_{t,a}}{w_{t,b}} \right) \quad (3.17)$$

As KL-Divergence is not symmetric so it is not a true measure of similarity, as a similarity measure should have symmetric property.

Average KL-Divergence: For documents, the averaged KL divergence can be computed with the following equation (3.18):

$$D_{AvgKL} \left(\vec{t}_a \parallel \vec{t}_b \right) = \sum_{t=1}^m (\pi_1 \times D(w_{t,a} \parallel w_t) + \pi_2 \times D(w_{t,b} \parallel w_t)) \quad (3.18)$$

where $\pi_1 = \frac{w_{t,a}}{w_{t,a}+w_{t,b}}$, $\pi_2 = \frac{w_{t,b}}{w_{t,a}+w_{t,b}}$ and $w_t = \pi_1 \times w_{t,a} + \pi_2 \times w_{t,b}$

The average weighting between two vectors ensures symmetry, that is, the divergence from document i to document j is the same as the divergence from document j to document i .

Information Radius: Information radius is a probabilistic similarity measure which is another name for Average KL-Divergence measure.

Point-wise Mutual Information (PMI): It is a measure of association used in information theory and statistics. The PMI of a pair of outcomes x and y belonging to discrete random variables X and Y quantifies the discrepancy between the

probability of their coincidence given their joint distribution and their individual distributions, assuming independence.

Normalized Point-wise Mutual Information: Point-wise mutual information can be normalized between $[-1, +1]$ resulting in -1 (in the limit) for never occurring together, 0 for independence, and $+1$ for complete co-occurrence.

Second order co-occurrence Point-wise Mutual Information: In computational linguistics, second-order co-occurrence point-wise mutual information is a semantic similarity measure. To assess the degree of association between two given words, it uses point-wise mutual information (PMI) to sort lists of important neighbor words of the two target words from a large corpus. These words will not be co-occurring directly in a sentence, therefore their neighbor words will be required to be used to find their co-occurrence.

Term co-occurrence: In linguistics, co-occurrence or co-occurrence of terms refers to an above-chance frequency of occurrence of two terms (also known as coincidence or concurrence) from a text corpus alongside each other in a certain order.

Co-word analysis: A content analysis technique that is used to map the strength of association between keywords in textual data. This technique measures the co-occurrence of keywords to examine content in the textual data.

Topic graph: A graph in which nodes represent different topics. Different ontologies are also examples of topic graph such as scientific subject ontology. Weighted keyword graphs can be generated from the ontologies based on topic graphs as discussed by [27].

Citation context based similarity measuring technique: Such similarity measuring techniques which use sentences of research papers in which citation tags appear to find similarity between the papers.

Citation tags: The markers/tags used to cite other research papers in a research paper are called citation tags. There are different formats for these tags: such as square brackets [], author first name with year of publication, etc.

Bibliography lists: The reference list at the end of a research paper which contains the research papers cited by this paper. These lists can be in different formats depending on the reference style used by a research paper. Most popular and commonly used styles are Chicago, APA (American Psychological Association), MLA (Modern Language Association), Harvard, and Vancouver etc.

Citation graphs of research papers: In information science and bibliometric, a citation graph (or citation network) is a directed graph in which each vertex represents a document and in which each edge represents a citation from the current publication to another.

Bibliographic coupling: Bibliographic coupling occurs when two works/research papers reference a common third work/research paper in their bibliographies. It is an indication that a probability exists that the two works treat a related subject matter. The “coupling strength” of two given documents is higher the more citations to other documents they share.

Co-citation analysis: Co-citation is defined as the frequency with which two documents are cited together by other documents. If at least one other document cites two documents in common these documents are said to be co-cited. The more co-citations two documents receive, the higher their co-citation strength, and the more likely they are semantically related.

Direct citation analysis: In case of direct citation, the similarity relationship between research papers/articles is computed if a research paper is cited in other paper.

Citation matrix: A matrix which contains the information about the citation information of documents. If C denotes a citation matrix if C_{ij} is an element at intersection of a row representing a document d_i and column representing d_j . Value of $C_{ij} = 1$ if d_i cites d_j , and zero otherwise.

Cosine similarity matrix: A cosine similarity matrix represents the cosine similarity measures between each pair of documents from a set of documents. This set is represented by following equation (3.19).

$$S_{ij} = \cos(d_i, d_j) = \frac{d_i^T \cdot d_j}{\|d_i\| \cdot \|d_j\|} \quad (3.19)$$

where S_{ij} represents an entry in cosine matrix for cosine similarity measure between document d_i and d_j .

Authority metric: Given two research papers A and B , for each of these papers a location citation graph can be constructed. These graphs are labelled as Graph A and Graph B . The similarity between paper A and B can be computed using their citation graphs. From citation graphs of each of the papers only “Authoritative” papers can be used to represent the citation environments. Authority metric for finding an authority paper is by computing the in-degree measure for its node in the citation graph.

Structural similarity: Finding similarity between two documents by using their structure. A research using structural similarity was using XML layout of documents to find similarity between them. Tree edit distance based similarity approach was used to find similarity between the documents.

Visual similarity: The visual layout of research papers in the form of scanned images is used to find similarity between them. Lexical similarity represents edit distance and tree edit distance similarity measuring techniques using strings and concept trees as weighting schemes respectively.

XML tag frequency: The XML tag frequencies of two documents in comparison can be computed. A distance based similarity measure between the XML tags frequencies of two documents can be used to find the similarity measures between them.

Document clustering: Document clustering (or text clustering) is the application of cluster analysis to textual documents. It has applications in automatic document organization, topic extraction and fast information retrieval or filtering.

Short text segment similarity: A similarity measuring technique which is used to find similarity between the short text segments. These segments can be phrases, sentences, or paragraphs of text.

Pairwise word similarity: The measure of computing word similarity on a semantic level functions on a graph based representation of words and the semantic relations among them within a lexical semantic resource. For this purpose normally graphs like WordNet are used.

Longest Common Subsequence (LCS) measure: The longest common subsequence (LCS) problem is the problem of discovering the longest subsequence common to all sequences in a set of sequences (often just two sequences).

Textual Entailment: In natural language processing, it is a directional relation between text fragments. The relation holds whenever the truth of one text fragment follows from another text. In the Textual Entailment framework, the entailing and entailed texts are termed text (t) and hypothesis (h), respectively. An example of textual entailment is: Text (t): If you help the needy, God will reward you. Hypothesis (h): Giving money to a poor man has good consequences.

3.3 Survey of Content Based Similarity Measuring Techniques

In this section, we have surveyed different content based similarity measuring techniques. We have considered following categories of these techniques:

1. Vector Space-based similarity measuring techniques
2. Probabilistic similarity measuring techniques
3. Citation-based similarity measuring techniques
4. Miscellaneous similarity measuring techniques

During the survey for each of these categories, we will seek answers to the three questions raised in Section 3.1 of this chapter. For the first three categories in the above list, we have organized the surveyed techniques in the form of tables, shown in Tables 3.4 to 3.38

3.4 Vector Space Based Similarity Measuring Techniques

Different types of Vector Space based similarity measuring methods are Cosine, Jaccard, Pearson Correlation, Overlapping, and Dice coefficients, etc. Content-based similarity measuring techniques are commonly used to find similarity among the research papers [10] in recommender systems. Therefore, in this section, our main focus will be on a survey of those similarity measuring techniques which are majorly used in the recommender systems.

3.4.1 Finding Concepts With Disjoint and Overlapping Relationships

To answer the first question in Section 3.1, we have analyzed the vector space based similarity measuring techniques using a multidimensional way, to classify a particular concept either with disjoint or overlapping relationships with other concepts. A secondary outcome from this analysis was identification of different research gaps in the surveyed similarity measuring methods. We have selected five parameters for this analysis listed below. These parameters are selected to get maximum information contents of a research paper which can be used by vector space based similarity measuring techniques.

1. Research Paper Sections: representing different sections of research paper like Title, Authors, Abstract, etc.

2. Text Representation Schemes: these schemes represent different text layouts of text contained by research papers. We have used surface text, expanded text and stemmed text categories for analysis.
3. Text Extraction Schemes: these schemes represent terms or phrases which are extracted from text of research papers/documents.
4. Research Paper Weighting Schemes: they represent different weighting schemes which are used by different Vector Space based similarity methods. These schemes are term frequency, TF/IDF, term position, term weight, n-gram, binary vector, phrase depth etc.
5. Vector Space based and String based Similarity Methods: these methods represent different categories such as Cosine, Jaccard, Pearson Correlation, Matching Coefficient, Edit Distance etc.

We have analyzed different vector space based similarity measuring techniques in the form of two dimensional tables. Each table analyses the techniques by using the two parameters from above list. For five above mentioned parameters there were 10 different combinations between each of two parameters. For each of these combinations a table was designed to perform analysis on similarity measuring techniques. The cells of these tables represent the different surveyed similarity measuring techniques. The grey colored cells of tables represent the research gaps found during this analysis.

Table 3.4 represents the analysis of Vector Space based similarity techniques by using “Research Paper Sections” and “Text Representation Schemes” dimensions. This table contains 12 rows and 3 columns. The rows represent different sections of a research paper whereas columns represent different categories of Text Representation Schemes.

By observing Table 3.4, conclusions were made for vector space based similarity measuring techniques which are listed below:

TABLE 3.4: Analysis on basis of Research Paper Sections and Text Representation Schemes

Research Paper Sections	Text Representation Schemes		
	Surface Text	Expanded Text	Stemmed Text
Title	[1, 40]		
Authors	[1, 40]		
Affiliations/Venue	[1]		
Abstract	[1, 41, 42]		
Authors Provided Keywords	[41, 43, 44]		
Introduction	[1, 43]		
Related Work			
Methodology			
Results			
Conclusions			
Bibliography	[45]		
All contents (except Bibliography)	[1, 37, 39, 46-57]	[58]	

1. No such similarity measuring techniques were discovered which use stemmed text representation of research papers to find similarity among them.
2. No similarity measuring techniques were discovered which used “Related Work”, “Methodology”, “Results”, and “Conclusion” sections of research papers for finding similarity.
3. There was only one technique found which was using expanded text of research papers for finding similarity measures.

Table 3.5 represents the analysis of Vector Space based similarity techniques by using “Research Paper Sections” and “Text Extraction Schemes” dimensions. This table contains 12 rows and 2 columns. The rows represent different sections of a research paper whereas columns represent different categories of Text Extraction Schemes.

By observing Table 3.5, conclusions were made for Vector Space-based similarity techniques which are listed below:

1. There was only one technique found which was using “Text Phrases” for Abstract section of research papers for finding similarity.

TABLE 3.5: Analysis on basis of Research Paper Sections and Text Extraction Schemes

Research Paper Sections	Text Representation Schemes	
	Bag of Words/Extracted Keywords	Text Phrases
Title	[1]	
Authors	[1]	
Affiliations/Venue	[1]	
Abstract	[1, 41, 42]	[1]
Authors Provided Keywords	[41, 43, 44]	
Introduction	[1, 43]	
Related Work		
Methodology		
Results		
Conclusions		
Bibliography	[41]	
All contents (except Bibliography)	[1, 37, 39, 46–55, 57]	

2. No similarity techniques were discovered which used “Related Work”, “Methodology”, “Results”, and “Conclusion” sections of research papers for finding similarity.

Table 3.6 represents the analysis of Vector Space based similarity techniques by using “Research Paper Sections” and “Research Paper Weighting Schemes” dimensions. This table contains 12 rows and 11 columns. The rows represent different sections of a research paper whereas columns represent different categories of Research Paper Weighting Schemes.

By observing the Table 3.6, conclusions were made for Vector Space-based similarity techniques which are listed below:

1. “Related Work”, “Methodology”, and “Bibliography” sections of research paper were not used by any of the discovered similarity measuring techniques.
2. For “Title”, “Authors”, “Affiliation/Venue”, “Abstract”, and “Introduction” sections of research paper only “Term Frequency”, “CC/IDF”, and “Concept Graph” weighting schemes were used in similarity techniques.

TABLE 3.6: Analysis on basis of Research Paper Sections and Research Paper Weighting Schemes

Research Paper Sections	Research Paper Weighting Schemes										
	Term Weight	Term Position	Term Frequency	IDF	TF/IDF	CC/IDF	n-gram	Binary Vector	Phrase Depth	Concept Tree	Concept Graph
Title						[1]					[40]
Authors						[1]					[40]
Affiliations/ Venue						[1]					
Abstract			[1, 41, 42]			[1]					
Authors Provided			[41, 43, 44]			[1]					
Keywords											
Introduction			[1, 43]			[1]					
Related Work											
Methodology											
Results											
Conclusions											
Bibliography											
All contents (except Bibliography)	[39]	[39]	[39, 47, 48]	[39]	[1, 37, 39, 47-51, 53-55, 57, 58]	[1, 44, 52]	[39, 47]	[59]	[47]	[49]	[46]

3. “TF/IDF” was most commonly used weighting scheme in different similarity measuring techniques.

Table 3.7 represents the analysis of vector space based similarity measuring techniques by using “Research Paper Sections” and “Vector Space based and String based Similarity Methods” dimensions. This table contains 12 rows and 9 columns. The rows represent different sections of a research paper whereas columns represent different categories of vector space based and string based similarity measuring methods.

By observing Table 3.7, conclusions were made for Vector Space-based similarity techniques which are listed below:

1. “Related Work”, “Methodology”, “Results”, “Conclusions”, and “Bibliography” sections of research paper were not used by any of the surveyed similarity technique.
2. Edit Distance measure was used in 1998 but after that it was not adopted to compute research paper similarity.
3. L-Cosine, Soft Cosine, Jaccard, Pearson Correlation, Matching Coefficient, Overlapping Coefficient, and Euclidean Distance were not used for “Title”, “Authors”, “Affiliations”, “Author Provided Keywords”, “Abstract”, and “Introduction” sections of research papers on individual basis.
4. Cosine similarity method was most commonly used in different similarity measuring techniques.

Table 3.8 represents the analysis of vector space based similarity measuring techniques by using “Text Representation Schemes” and “Text Extraction Schemes” dimensions. This table contains 3 rows and 2 columns. The rows represent different categories of Text Representation Schemes whereas columns represent different categories of Text Extraction Schemes.

By observing the Table 3.8, conclusions were made for Vector Space-based similarity techniques which are listed below:

TABLE 3.7: Analysis on basis of Research Paper Sections and Vector Space based Similarity Methods

Research Paper Sections	Vector Space based and String based Similarity Methods									
	Cosine	L-Cosine	Soft Cosine	Jaccard	Pearson Correlation	Matching Coefficient	Overlapping Coefficient	Euclidean Distance	Edit Distance	
Title	[1]								[1]	
Authors	[1]								[1]	
Affiliations/ Venue	[1]								[1]	
Abstract	[1, 41, 42]									
Authors Provided Keywords	[41, 43, 44]									
Introduction	[1, 43]									
Related Work										
Methodology										
Results										
Conclusions										
Bibliography										
All contents (except Bibliography)	[39]	[55]		[37, 53, 54]	[37, 53, 54]	[53, 59]	[53, 59]	[37, 53]		

TABLE 3.8: Analysis on basis of Text Representation Schemes and Text Extraction Schemes

Text Representation Schemes	Text Extraction Schemes	
	Bag of Words/Extracted Keywords	Text Phrases
Surface Text	[1, 37, 39–42, 42, 43, 45–57, 57]	[1]
Expanded Text	[58]	
Stemmed Text		

1. No similarity techniques were found which used stemmed text of research paper and text phrases to find similar papers.
2. Majority of the similarity techniques were using Surface Text of research papers in the form of Extracted Keywords or Bag of Words.
3. Text Phrases were not used by majority of similarity techniques.

Table 3.9 represents the analysis of vector space based similarity techniques by using “Text Representation Schemes” and “Research Paper Weighting Schemes” dimensions. This table contains 3 rows and 10 columns. The rows represent different categories of Text Representation Schemes whereas columns represent different categories of Research Paper Weighting Schemes.

By observing the Table 3.9, conclusions were made for Vector Space-based similarity measuring techniques which are listed below:

1. Expanded Text representation of research paper was used with TF/IDF weighting scheme only. No other weighting schemes used expanded text representation.
2. Stemmed Text representation of research paper was not used by any of the weighting scheme for research papers for surveyed techniques.

Table 3.10 represents the analysis of vector space based similarity measuring techniques by using “Text Representation Schemes” and “Vector Space based and String based Similarity Methods” dimensions. This table contains 3 rows and 9

columns. The rows represent different categories of Text Representation Schemes whereas columns represent different categories of Vector Space based and String based Similarity Methods.

By observing the Table 3.10, conclusions were made for Vector Space-based similarity techniques which are listed below:

1. Beside Cosine similarity, other similarity measuring methods were not using Expanded Text representation of research papers.
2. Stemmed Text representation of research paper was not used by any of the surveyed techniques.

Table 3.11 represents the analysis of vector space based similarity measuring techniques by using “Text Extraction Schemes” and “Research Paper Weighting Schemes” dimensions. This table contains 2 rows and 10 columns. The rows represent different categories of Text Extraction Schemes whereas columns represent different categories of Research Paper Weighting Schemes.

By observing the Table 3.11, conclusions were made for Vector Space-based similarity measuring techniques which are listed below:

1. Text Phrases were used in a single technique with weighting scheme of CC/IDF.
2. TF/IDF was most commonly used research paper weighting scheme with extracted keywords/bag of words.
3. Term Weight, Term Position, and Phrase Depth were used in many of the similarity measuring techniques as compared to TF/IDF.

Table 3.12 represents the analysis of vector space based similarity measuring techniques by using “Text Extraction Schemes” and “Vector Space based and String based Similarity Methods” dimensions. This table contains 2 rows and 9 columns. The rows represent different categories of Text Extraction Schemes

TABLE 3.1.1: Analysis on basis of Text Extraction Schemes and Research Paper Weighting Schemes

Text Extraction Schemes	Research Paper Weighting Schemes									
	Term Weight	Term Position	Term Frequency	TF/IDF	CC/IDF	n-gram	Binary Vector	Phrase Depth	Concept Tree	Concept Graph
Bag of Words/Extracted Keywords	[39]	[39]	[1, 39, 41-44, 47, 48]	[1, 37, 39, 46-51, 53, 54, 57, 58, 60, 61]	[1]	[39, 47]	[59]	[47]	[49]	[40, 46, 60, 62-65]
Text Phrases					[1]					

whereas columns represent different categories of Vector Space based and String based Similarity Methods.

By observing the Table 3.12, conclusions were made for Vector Space-based similarity techniques which are listed below:

1. Text Phrases are used in Cosine similarity measuring method but not in other methods.
2. Most of the similarity measuring methods used Extracted Keywords/Bag of words from research papers.

Table 3.13 represents the analysis of Vector Space based similarity measuring techniques by using “Research Paper Weighting Schemes” and “Vector Space based Similarity Methods” dimensions. This table contains 8 rows and 8 columns. The rows represent different categories of Research Paper Weighting Schemes whereas columns represent different categories of Vector Space based Similarity Methods.

By observing the Table 3.13, conclusions were made for Vector Space-based similarity measuring techniques which are listed below:

1. N-gram weighting scheme was not used in the surveyed Vector Space based similarity measuring methods.
2. Term Weight and Term Position were only used in Cosine and L-Cosine similarity measuring methods.
3. Term Frequency was only used in Cosine similarity measuring method but not in other methods.
4. Matching Coefficient and Overlapping Coefficient methods used only Binary Vector scheme.
5. Phrase depth scheme was only used with Euclidean distance.

3.4.2 Conclusions from Analysis of Vector Space Based Similarity Measuring Techniques

While performing knowledge acquisition for different Vector Space based similarity measuring techniques, we have reached to conclusions which are discussed in this section.

No such similarity measuring techniques were discovered which used stemmed text representation of research papers to find similarity between them. “Related Work”, “Methodology”, “Results”, and “Conclusion” sections of research papers were not used in any Vector Space based similarity measuring techniques. There was a technique found using expanded text of research papers. One of the techniques was using “Text Phrases” for Abstract section of research papers for finding similarity measures.

“Related Work”, “Methodology”, and “Bibliography” sections of research paper were not used by any of the surveyed similarity measuring technique. For “Title”, “Authors”, “Affiliation/Venue”, “Abstract”, and “Introduction” sections of research paper only “Term Frequency”, “CC/IDF”, and “Concept Graph” weighting schemes were used in the similarity techniques. “TF/IDF” was most commonly used weighting schemes in different similarity measuring techniques. “Related Work”, “Methodology”, “Results”, “Conclusions”, and “Bibliography” sections of research paper were not used by any of the surveyed similarity techniques. Edit Distance measure was used in 1998 but after that it was not adopted to compute research paper similarity measures.

L-Cosine, Soft Cosine, Jaccard, Pearson Correlation, Matching Coefficient, Overlapping Coefficient, and Euclidean Distance were not used for “Title”, “Authors”, “Affiliations”, “Author Provided Keywords”, “Abstract”, and “Introduction” sections of research papers on individual basis. Cosine similarity measuring method was most commonly used in different similarity measuring techniques. No similarity measuring techniques were found using stemmed text and text phrases of

research papers. Majority of the similarity measuring techniques were using Surface Text of research papers in the form of Extracted Keywords or Bag of Words. Text Phrases were not used by majority of similarity techniques. Expanded Text representation was used with TF/IDF weighting scheme only, no other weighting schemes used expanded text representation.

Stemmed Text representation of research paper was not used by any of the weighting scheme. Beside Cosine Similarity, other similarity measuring methods were not using Expanded Text representation of research papers. Stemmed Text representation of research paper was not used by any of the surveyed techniques. Text Phrases were used in a single technique with weighting scheme of CC/IDF. TF/IDF was most commonly used research paper weighting scheme with extracted keywords/bag of words. Term Weight, Term Position, and Phrase Depth were used in many other of the similarity measuring techniques as compared to TF/IDF. Text Phrases are used in Cosine similarity measuring method but not in other methods.

Most of the similarity measuring methods use Extracted Keywords/Bag of words from research papers. N-gram weighting scheme was not used in the surveyed similarity measuring methods. Term Weight and Term Position were only used in Cosine and L-Cosine similarity measuring methods. Term Frequency was only used in Cosine similarity measuring method but not in other methods. Matching Coefficient and Overlapping Coefficient methods used only Binary Vector scheme. Phrase depth scheme was only used with Euclidean Distance.

3.4.3 Survey of Vector Space Based Similarity Measuring Techniques

Each of the surveyed similarity measuring technique is briefly discussed in this section. Further the answers to the second (Semantic Model) and third questions (Integration with other techniques) discussed in the Section 3.1 were given for all these surveyed techniques at the end of this section.

Sugiyama's approach used plain text term frequency as a weighting scheme [48] for finding similarity measures among the research papers. In this scheme, a researcher's profile was constructed from his past research papers and feature vectors for candidate research papers were built. Then cosine similarity measure was computed between researcher's profile and candidate research papers' features. This technique used TF (Term Frequency) only instead of TF/IDF and ignored other weighting schemes.

Docear's research paper recommender system [58] uses mind maps. Mind maps are collections of papers, references, and annotations. The mind maps use weighting schemes TF (Term Frequency) and TF/IDF (Term Frequency \times Inverse Document Frequency) to compute Cosine similarity. This approach used TF and TF/IDF combined, but it ignored other weighting schemes such as CC/IDF, term weight, and term position etc.

Ferrara's technique used phrase-based similarity measures [47], in which a user is associated with a set of documents which were usually identified by tagging. These tagged documents are exploited by a Key Phrase Extraction Module (KPEM). The similarity measure between key phrases of user's papers and repository papers is computed using cosine similarity measuring technique. This technique used n-grams, term frequencies but it ignored other weighting schemes such as TF/IDF, term weight, term position etc.

Jiang et al. [42] explained that other similarities measuring techniques provide just a list of papers which were similar to a paper provided by a researcher. The user is required to separate problem and solution based papers from this list by herself. The proposed technique used cosine similarity on feature vectors of research papers to provide two lists of recommended papers: papers representing problem and papers representing a solution. Similarity models were built for abstracts of research papers using TF-IDF and topic models, ignoring other weighting schemes.

Bethard et al. [50] used feature model of research papers based on weighting schemes like TF/IDF, citation count, recency of research paper, citation contexts, topic models, and social habits. All these features are included in the feature

vectors, making it rich enough to improve the recall. The similarity measuring technique used in this approach is a combination of Cosine and LDA (Latent Dirichlet Allocation) while ignoring other similarity measuring techniques.

Ozono's approach models the users of research papers as graphs [66]. A search facility named as "parties" was developed in this approach, which found information about a person and uses "a know who" search mechanism accessing information from distributed resources. Similarity measures used in this technique are keyword based matching scheme only, ignoring other similarity techniques such as Cosine, Jaccard etc. No other weighting schemes like TF/IDF, TF etc. is used in this approach.

Citeseer is a known citation indexer [1] which uses a combination of string based and cosine similarities. According to authors, TF/IDF can be affected by noise data and cannot effectively be used for finding similar documents. Another problem with TF/IDF is that it ignores semantics of terms. Citeseer used common citations of research papers and formulated it as a new measure named as common citation \times inverse document frequency (CC/IDF). Citeseer used CC/IDF and TF/IDF as a combined weighting scheme with cosine and string based similarity measures. No other weighting schemes and similarity measuring techniques are used in this approach.

Vector Space-based similarity measuring techniques were discussed in the book "Foundations of Statistical Natural Language Processing" [59]. These techniques were named as Cosine, Jaccard, Overlapping, Dice, and matching coefficients. Cosine and Jaccard were used on Vector Space-based model of documents. The remaining techniques used Binary Vectors of documents. According to authors, Cosine similarity measure was most commonly used technique because it uses feature vectors of documents, which can be made rich by use of different weighting schemes.

Huang discussed the content based similarity measuring techniques [37] in the

context of clustering the similar papers. Different similarity measuring techniques such as Cosine, Jaccard, Pearson correlation coefficient, and average KL-Divergence are used, while ignoring other similarity techniques. Only TF/IDF weighting scheme and the probability distribution of words are used.

In a recommender system [39], several potential queries using terms from a query research paper are generated and posted on the Web to find the candidate research papers. After collecting these papers, content-based similarity techniques are applied on these papers to find their relevance with the query paper. These papers are ranked according to their relevance score. Cosine similarity is used to find the relevant papers using weighting schemes n-gram, term weight, and term position. No other similarity techniques and weighting schemes were considered by the authors.

A recommender system [49] also used tree edit distance similarity for finding research papers related to an author's publication interest. In this approach authors' profiles were built by using their previous publications. By comparing the profile of author with other profiles in a collection database, research papers in those profiles were recommended to that author. These profiles were maintained as a tree of concepts and tree edit distance is used as a similarity measuring approach.

A document clustering algorithm is proposed in which authors have devised a new similarity technique to compute the pairwise similarity of text-based documents using the suffix tree document model [56]. This similarity technique is utilized to devise a new document clustering algorithm. This algorithm is applied to the web page based documents for clustering. The results obtained from the author's proposed algorithm are better than the traditional TF/IDF based measures. This technique does not involve the use of any combinations of similarity measuring techniques to perform the clustering operation.

Another study [53] utilizes different vector space based similarity measuring approaches in a recommender system to generate recommendations for E-Commerce and Social Web sites. Authors surveyed different similarity measuring algorithms

like Cosine, Pearson Correlation based, Euclidean distance, and their effectiveness in a recommender system. No combinations of these similarity measuring techniques have been used in the proposed approach.

In another scheme [57] authors have used lexical similarity measures using dependency graph structures. Different features have been employed to compute the similarities such as, a bag of words, topic distribution, and dependency structures, named entities, and expansion features. The approach uses cosine similarity measuring technique for vectors based on the above-described features. No combinations of other similarity measuring techniques were analyzed in this approach.

In this approach [54] authors have investigated the utility of Inclusion Index, the Jaccard Index, and the Cosine Index for calculating the similarity measures of documents. According to the authors, Inclusion Index provides a better similarity measure in particular when computing similarity using citation data. In this scheme, the comparison is performed between other similarity measuring techniques like co-word analysis, Subject-Action-Object (SAO) structures, bibliographic coupling, Co-citation analysis, and self-citation links. However, this research does not provide any similarity measuring technique as combinations of multiple similarity techniques.

A new [55] similarity measuring technique is proposed in this research by using Cosine similarity named as Soft Cosine similarity when there is no similarity measure between the features, the proposed soft similarity measuring technique becomes equal to the standard similarity. Soft cosine similarity measuring technique is a generalized model of cosine similarity measure. Authors have proposed different formulas for exact or approximate calculations for the soft cosine similarity measure. This research does not use different similarity measuring techniques in a combined way.

Another [51] approach used clustering of keywords for extending scientific subject ontology. The ontology used by this technique was based on a topic graph in which nodes were representing different topics. Different graph based similarity

measures were used in this technique. This research technique does not use different similarity measuring techniques in a combined way.

A technique [40] which used to implement random walk on graphs of research paper was published. This technique was used to solve the cold start problems in Recommender Systems. This technique was not using any combinations of different similarity measuring techniques.

An approach [43] combined content based and collaborative filtering approaches for finding similar research papers. This approach used the concept of graph based recommender system. Different vector space based similarity techniques were used in this approach.

We analyzed all these approaches to find that whether they use a semantic based conceptual model for a given similarity measuring technique based on the operational semantics and features used. According to our findings, no such similarity approach found which used such a semantic model. We could not find such similarity techniques which can be integrated into the formulation of a hybrid approach without redundancies and overlapping in the methods and features.

3.5 Probabilistic Similarity Measuring Techniques

Different Probabilistic similarity measuring methods used probability distribution of words/terms from research papers, named as KL-Divergence, Average KL-Divergence, Point-wise Mutual Information (PMI), Normalized PMI, Information Radius, and Manhattan Norm. The techniques based on Probabilistic similarity measuring methods are used in applications for document clustering and recommender systems.

3.5.1 Finding Concepts with Disjoint and Overlapping Relationships

To answer the first question in Section 3.1, we have analyzed the probabilistic similarity techniques by a multidimensional way, to design/classify a particular concept either with disjoint or overlapping relationships with other concepts. A secondary outcome from this analysis was identification of different research gaps in surveyed similarity methods. We have selected six parameters for this analysis listed below. Five of these parameters were defined in Section 3.4.1 for vector space based similarity measuring techniques.

1. Research Paper Sections: Representing different sections of research paper like Title, Authors, Abstract, etc.
2. Text Representation Schemes: they representation different text layouts of text contained by research papers. We have used surface text, expanded text and stemmed text categories for analysis.
3. Text Extraction Schemes: these schemes represent terms or phrases which are extracted from text of research papers/documents and bag of words/extracted keywords.
4. Research Paper Weighting Schemes: they represent different weighting schemes which are used by different Probabilistic similarity methods. These schemes are Probability distribution of Terms and Term Co-occurrence sets.
5. Probabilistic Similarity Methods: these methods represent different probabilistic similarities such as KL-Divergence, Average KL-Divergence, Information Radius, Manhattan norm, Pointwise Mutual Information, and Normalized Point-wise Mutual Information.
6. Entities related to Research Papers: these are divided into two categories: Research Paper contents and Entities outside the Research Paper which are used in different similarity methods.

The reason for using last parameter in above list is usage of entities/contents defined outside the research papers for computation of probabilistic similarity measures. We have analyzed different Probabilistic similarity measuring techniques in the form of two dimensional tables. Each table analyses the techniques by using the two parameters from the above list. For six above mentioned parameters there were 15 different combinations between each of the two parameters. For each of these combinations a table was designed to perform analysis on similarity measuring techniques.

Table 3.14 represents the analysis of Probabilistic similarity measuring techniques by using “Research Paper Sections” and “Text Representation Schemes” dimensions. This table contains 12 rows and 3 columns. The rows represent different sections of a research paper whereas columns represent different categories of Text Representation Schemes.

TABLE 3.14: Analysis using Research Paper Sections and Text Representation Schemes

Research Paper Sections	Text Representation Schemes		
	Surface Text	Expanded Text	Stemmed Text
Title			
Authors			
Affiliations/Venue			
Abstract			
Authors Provided Keywords			
Introduction			
Related Work			
Methodology			
Results			
Conclusions			
Bibliography			
All contents (except Bibliography)	[3, 5, 37, 38, 54, 59]	[3]	[3]

By observing Table 3.14, conclusions were made for Probabilistic similarity measuring techniques which are listed below:

1. No surveyed probabilistic similarity measuring technique was working on individual sections for research paper. All these techniques were using all contents of research paper except bibliography.
2. Only a single technique was found which was using expanded text and Stemmed text to find similarity measures. Although the technique was working on short segments of text. No such similarity technique was found which used expanded or stemmed representation of text for research papers to find similarity measures between them.

Table 3.15 represents the analysis of Probabilistic similarity measuring techniques by using “Research Paper Sections” and “Research Paper Weighting Schemes” dimensions. This table contains 12 rows and 2 columns. The rows represent different sections of a research paper whereas columns represent different categories of Research Paper Weighting Schemes.

TABLE 3.15: Analysis using Research Paper Sections and Research Paper Weighting Schemes

Research Paper Sections	Research Paper Weighting Schemes	
	Probability Distribution of Terms	Term Co-occurrence Set
Title		
Authors		
Affiliations/Venue		
Abstract		
Authors Provided		
Keywords		
Introduction		
Related Work		
Methodology		
Results		
Conclusions		
Bibliography		
All contents (except Bibliography)	[3, 5, 37, 54, 59]	[38]

By observing Table 3.15, conclusions were made for Probabilistic similarity measuring techniques which are listed below:

1. No such technique found which used individual sections of research papers to find similarity measures between them.
2. Probability distribution of terms was more commonly used in probabilistic similarity measuring techniques rather than co-occurrence based term sets.

Table 3.16 represents the analysis of Probabilistic similarity measuring techniques by using “Research Paper Sections” and “Text Extraction Schemes” dimensions. This table contains 12 rows and 2 columns. The rows represent different sections of a research paper whereas columns represent different categories of Text Extraction Schemes. By observing Table 3.16, conclusions were made for Probabilistic

TABLE 3.16: Analysis using Research Paper Sections and Text Extraction Schemes

Research Paper Sections	Text Extraction Schemes	
	Bag of Words/Extracted Keywords	Text Phrases
Title		
Authors		
Affiliations/Venue		
Abstract		
Authors Provided		
Keywords		
Introduction		
Related Work		
Methodology		
Results		
Conclusions		
Bibliography		
All contents (except Bibliography)	[3, 5, 37, 38, 54, 59]	

similarity measuring techniques which are listed below:

1. Bag of Words/Extracted Keywords scheme was not used for individual sections of research papers for Probabilistic similarity techniques.
2. Text phrases were not used for contents of research papers for Probabilistic similarity techniques.

Table 3.17 represents the analysis of Probabilistic similarity measuring techniques by using “Research Paper Sections” and “Entities Related to Research Paper” dimensions. This table contains 12 rows and 2 columns. The rows represent different sections of a research paper whereas columns represent different categories of Entities Related to Research Paper.

TABLE 3.17: Analysis using Research Paper Sections and Entities Related to Research Paper

Research Paper Sections	Entities related to Research Paper	
	Research Paper Contents	Entities Outside the Research Paper
Title		
Authors		
Affiliations/Venue		
Abstract		
Authors Provided Keywords		
Introduction		
Related Work		
Methodology		
Results		
Conclusions		
Bibliography		
All contents (except Bibliography)	[3, 5, 37, 54, 59]	[38]

By observing Table 3.17, conclusions were made for Probabilistic similarity measuring techniques which are listed below:

1. No such techniques found which use contents of research papers for individual sections of research papers. Most of probabilistic similarity measuring techniques were using all contents of research papers to find similarity among them.
2. No such techniques found which use entities outside the contents of research papers for individual sections of research papers.

Table 3.18 represents the analysis of Probabilistic similarity measuring techniques by using “Research Paper Sections” and “Probabilistic Similarity Methods” dimensions. This table contains 12 rows and 6 columns. The rows represent different sections of a research paper whereas columns represent different categories of Probabilistic Similarity Methods.

By observing Table 3.18, conclusions were made for Probabilistic similarity measuring techniques which are listed below:

1. None of the mentioned probabilistic similarity measuring methods were used for individual sections (Title, Authors, Abstract, and Introduction etc.) of research papers.
2. All the surveyed techniques were using all contents of research papers. The techniques such as Point-wise Mutual Information (PMI) and Normalize PMI were using term co-occurrence sets to find similarity between documents.

Table 3.19 represents the analysis of Probabilistic similarity measuring techniques by using “Text Representation Schemes” and “Text Extraction Schemes” dimensions. This table contains 3 rows and 2 columns. The rows represent Text Representation Schemes whereas columns represent different categories of Text Extraction Schemes.

By observing Table 3.19, conclusions were made for Probabilistic similarity measuring techniques which are listed below:

1. Text Phrases were not used in any of surveyed probabilistic similarity measuring techniques for any category of Text Representation Schemes.
2. A minor number of Probabilistic similarity measuring techniques were found that were using Expanded and Stemmed Text representations with Bag of Words from Extracted Text Schemes.

Table 3.20 represents the analysis of Probabilistic similarity measuring techniques by using “Text Representation Schemes” and “Research Paper Weighting Schemes”

TABLE 3.18: Analysis using Research Paper Sections and Probabilistic Similarity Methods

Research Paper Sections	Probabilistic Similarity Methods					
	KL-Divergence	Average KL-Divergence	Information Radius	Manhattannorm	Pointwise Mutual Information	Normalized Pointwise Mutual Information
Title						
Authors						
Affiliations/Venue						
Abstract						
Authors Provided						
Keywords						
Introduction						
Related Work						
Methodology						
Results						
Conclusions						
Bibliography						
All contents (except Bibliography)	[3, 37, 59]	[3, 37, 54, 59]	[59]	[59]	[5, 38]	[5, 38]

TABLE 3.19: Analysis using Text Representation Schemes and Text Extraction Schemes

Text Representation Schemes	Text Extraction Schemes	
	Bag of Words/Extracted Keywords	Text Phrases
Surface Text	[3, 5, 37, 38, 54, 59]	
Expanded Text	[3]	
Stemmed Text	[3]	

dimensions. This table contains 3 rows and 2 columns. The rows represent Text Representation Schemes whereas columns represent different categories of Research Paper Weighting Schemes.

TABLE 3.20: Analysis using Text Representation Schemes and Research Paper Weighting Schemes

Text Representation Schemes	Research Paper Weighting Schemes	
	Probability Distribution of Terms	Term Co-occurrence Set
Surface Text	[3, 37, 38, 54, 59] [5]	[5, 38]
Expanded Text	[3]	
Stemmed Text	[3]	

By observing Table 3.20, conclusions were made for Probabilistic similarity measuring techniques which are listed below:

1. Term Co-occurrence set was used for those similarity measuring techniques which were using Surface Text category from Text Representation Scheme.
2. Expanded Text and Stemmed Text categories from Text Representation Schemes were not commonly used in different probabilistic similarity measuring techniques.

Table 3.21 represents the analysis of Probabilistic similarity measuring techniques by using “Text Representation Schemes” and “Entities related to Research Paper” dimensions. This table contains 3 rows and 2 columns. The rows represent Text

TABLE 3.21: Analysis using Text Representation Schemes and Research Entities related to Research Paper

Text Representation Schemes	Entities related to Research Paper	
	Research Paper Contents	Entities Outside the Research Paper
Surface Text	[3, 5, 37, 38, 54, 59]	[5, 38]
Expanded Text	[3]	
Stemmed Text	[3]	

Representation Schemes whereas columns represent different categories of Entities related to Research Paper.

By observing Table 3.21, conclusions were made for Probabilistic similarity measuring techniques which are listed below:

1. Entities outside the research paper were not used in Probabilistic similarity measuring techniques for Expanded and Text Representation of text for research papers/documents.
2. Majority of the techniques for using Surface Text Representation of Research Papers as contents of research paper in Probabilistic Similarity measuring techniques.

Table 3.22 represents the analysis of Probabilistic similarity measuring techniques by using “Text Representation Schemes” and “Probabilistic Similarity Methods” dimensions. This table contains 3 rows and 6 columns. The rows represent Text Representation Schemes whereas columns represent different categories of Probabilistic Similarity Methods.

By observing Table 3.22, conclusions were made for Probabilistic similarity measuring techniques which are listed below:

1. Expanded Text representation was not used in Information Radius, Manhattan norm, Point-wise Mutual Information, and Normalized Point-wise Mutual Information similarity techniques.

TABLE 3.22: Analysis using Text Representation Schemes and Probabilistic Similarity Methods

Text Representation Schemes	Probabilistic Similarity Methods					
	KL-Divergence	Average KL-Divergence	Information Radius	Manhattannorm	Pointwise Mutual Information	Normalized Pointwise Mutual Information
Surface Text	[3, 37, 54, 59]	[3, 37, 59]	[59]	[59]	[5, 38]	[5, 38]
Expanded Text	[3]	[3]				
Stemmed Text	[3]	[3]				

2. KL-Divergence and Average KL-Divergence used Surface, Expanded, and Stemmed Text representations to find the similarity measures among the research papers/documents.

Table 3.23 represents the analysis of Probabilistic similarity measuring techniques by using “Term Extraction Schemes” and “Research Paper Weighting Schemes” dimensions. This table contains 2 rows and 2 columns. The rows represent Term Extraction Schemes whereas columns represent different categories of Research Paper Weighting Schemes.

TABLE 3.23: Analysis using Term Extraction Schemes and Research Paper Weighting Schemes

Term Extraction Schemes	Research Paper Weighting Schemes	
	Probability Distribution of Terms	Term Co-occurrence Set
Bag of Words/Extracted Keywords	[3, 5, 37, 38, 54, 59]	[5, 38]
Text Phrases		

By observing Table 3.23, conclusions were made for Probabilistic similarity measuring techniques which are listed below:

1. Text Phrases were not used either by Probability distribution of terms or by term co-occurrence sets.
2. Majority of the probabilistic similarity measuring techniques used probability distribution of terms rather than term co-occurrence sets.

Table 3.24 represents the analysis of Probabilistic similarity measuring techniques by using “Term Extraction Schemes” and “Entities related to Research Paper” dimensions. This table contains 2 rows and 2 columns. The rows represent Term Extraction Schemes whereas columns represent different categories of Entities related to Research Paper.

By observing Table 3.24, conclusions were made for Probabilistic similarity measuring techniques which are listed below:

TABLE 3.24: Analysis using Term Extraction Schemes and Entities related to Research Paper

Term Extraction Schemes	Entities related to Research Paper	
	Research Paper Contents	Entities Outside the Research Paper
Bag of Words/Extracted Keywords	[3, 5, 37, 38, 54, 59]	[5, 38]
Text Phrases		

1. Text Phrases were not used either by Research Paper Contents or by Entities Outside the Research Paper.
2. Bag of words/Extracted Keywords scheme was most commonly used Research Paper Contents scheme.

Table 3.25 represents the analysis of Probabilistic similarity measuring techniques by using “Term Extraction Schemes” and “Probabilistic Similarity Methods” dimensions. This table contains 2 rows and 6 columns. The rows represent Term Extraction Schemes whereas columns represent different categories of Probabilistic Similarity Methods.

By observing Table 3.25, conclusions were made for Probabilistic similarity measuring techniques which are listed below:

1. Text Phrases were not used any of the surveyed Probabilistic similarity measuring method.
2. Almost all the surveyed techniques were using Bag of Words/Extracted Keywords.

Table 3.26 represents the analysis of Probabilistic similarity measuring techniques by using “Research Paper Weighting Schemes” and “Entities related to Research Paper” dimensions. This table contains 2 rows and 2 columns. The rows represent Research Paper Weighting Schemes whereas columns represent different categories of Entities related to Research Paper.

TABLE 3.25: Analysis using Term Extraction Schemes and Probabilistic Similarity Methods

Term Extraction Schemes	Probabilistic Similarity Methods					
	KL-Divergence	Average KL-Divergence	Information Radius	Manhattannorm	Pointwise Mutual Information	Normalized Pointwise Mutual Information
Bag of Words/Extracted Keywords	[3, 37, 59]	[3, 37, 54, 59]	[59]	[59]	[5, 38]	[5, 38]
Text Phrases						

TABLE 3.26: Analysis using Research Paper Weighting Schemes and Entities related to Research Paper

Research Paper Weighting Schemes	Entities related to Research Paper	
	Research Paper Contents	Entities Outside the Research Paper
Probability Distribution of Terms	[3, 5, 37, 38, 54, 59]	
Term Co-occurrence Set		[5, 38]

By observing Table 3.26, conclusions were made for Probabilistic similarity measuring techniques listed below:

1. Entities Outside the Research Paper were note used in Probability Distribution of Terms scheme.
2. Research Paper Contents were not used in Term Co-occurrence Set.

Table 3.27 represents the analysis of Probabilistic similarity measuring techniques by using “Research Paper Weighting Schemes” and “Probabilistic Similarity Methods” dimensions. This table contains 2 rows and 6 columns. The rows represent Research Paper Weighting Schemes whereas columns represent different categories of Probabilistic Similarity Methods.

By observing Table 3.27, conclusions were made for Probabilistic similarity measuring techniques which are listed below:

1. Term Co-occurrence Set is not used in KL-Divergence, Average KL-Divergence, Information Radius and Manhattannorm probabilistic similarity measuring techniques.
2. Point-wise Mutual Information and Normalized Point-wise Mutual Information is used in both Probability Distribution of Terms and Term Co-occurrence Set schemes.

TABLE 3.27: Analysis using Research Paper Weighting Schemes and Probabilistic Similarity Methods

Research Paper Weighting Schemes	Probabilistic Similarity Methods					
	KL-Divergence	Average KL-Divergence	Information Radius	Manhattannorm	Pointwise Mutual Information	Normalized Pointwise Mutual Information
Probability Distribution of Terms	[3, 37, 59]	[3, 37, 54, 59]	[59]	[59]	[5, 38]	[5, 38]
Term Co-occurrence Set					[5, 38]	[5, 38]

Table 3.28 represents the analysis of Probabilistic similarity measuring techniques by using “Entities Related to Research Paper” and “Probabilistic Similarity Methods” dimensions. This table contains 2 rows and 6 columns. The rows represent Entities Related to Research Paper whereas columns represent different categories of Probabilistic Similarity Methods.

By observing Table 3.28, conclusions were made for Probabilistic similarity measuring techniques which are listed below:

1. KL-Divergence, Average KL-Divergence, Information Radius, and Manhattan norm do not use Entities Outside the Research Paper.
2. Point-wise Mutual Information and Normalized Point-wise Mutual Information were using both Research Paper Contents and Entities Outside the Research Paper.

3.5.2 Conclusions from Analysis of Probabilistic Similarity Measuring Techniques

In this section we have reached to conclusions after analyzing and surveying Probabilistic based similarity measuring techniques. These conclusions are discussed further in this section.

Different Probabilistic similarity measuring techniques use contents of all sections of research paper except bibliography. Only a single technique was found which was using expanded text and Stemmed text representations. Probability distribution of terms was more commonly used in probabilistic similarity measuring techniques rather than co-occurrence based term sets. Bag of Words/Extracted Keywords scheme was not used for individual sections of research papers for Probabilistic similarity measuring techniques. Text phrases were not used for contents of research papers in case of Probabilistic similarity measuring techniques. No such techniques found which use entities outside the contents of research papers

TABLE 3.28: Analysis using Entities Related to Research Paper and Probabilistic Similarity Methods

Entities Related to Research Paper	Probabilistic Similarity Methods					
	KL-Divergence	Average KL-Divergence	Information Radius	Manhattannorm	Pointwise Mutual Information	Normalized Pointwise Mutual Information
Research Paper Contents	[3, 37, 59]	[3, 37, 54, 59]	[59]	[59]	[5, 38]	[5, 38]
Entities Outside the Research Paper					[5, 38]	[5, 38]

for individual sections (Title, Authors, Abstract, and Introduction etc.) of the research papers. None of the mentioned probabilistic similarity measuring methods were used for individual sections (Title, Authors, Abstract, and Introduction etc.) of research papers. Techniques such as Point-wise Mutual Information (PMI) and Normalize PMI were using term co-occurrence sets to find similarity among the documents.

A minor number of Probabilistic similarity measuring techniques were found that were using Expanded and Stemmed Text representations with Bag of Words Extracted Text Scheme. Term Co-occurrence set was used for those Probabilistic similarity measuring techniques, which were using Surface Text Representation Scheme. Entities outside the research paper were not used in Probabilistic similarity measuring techniques for Expanded and Stemmed Text Representation. Expanded Text representation was not used in Information Radius, Manhattannorm, Point-wise Mutual Information, and Normalized Point-wise Mutual Information similarity measuring techniques. KL-Divergence and Average KL-Divergence used Surface, Expanded, and Stemmed Text representations to find the similarity between documents/short segments of text. Text Phrases were not used either by Probability distribution of terms or by term co-occurrence sets. Majority of the probabilistic similarity measuring techniques used probability distribution of terms rather than term co-occurrence sets.

Text Phrases were not used either by Research Paper Contents or by Entities Outside the Research Paper. Almost all the surveyed techniques were using Bag of Words/Extracted Keywords extraction scheme. Term Co-occurrence Set is not used in KL-Divergence, Average KL-Divergence, Information Radius and Manhattannorm techniques. Point-wise Mutual Information and Normalized Point-wise Mutual Information were used in both Probability Distribution of Terms and Term Co-occurrence Set schemes. KL-Divergence, Average KL-Divergence, Information Radius, and Manhattannorm were not used in Entities Outside the Research Paper. Point-wise Mutual Information and Normalized Point-wise Mutual Information were using both Research Paper Contents and Entities Outside the Research Paper.

3.5.3 Survey of Probabilistic Similarity Measuring Techniques

Each of the surveyed similarity measuring technique is briefly discussed in this section. Further the answers to the second (Semantic Model) and third questions (Integration with other techniques) discussed in the Section 3.1 were answered for all these surveyed techniques at the end of this section.

An approach [37] used Average KL-Divergence for document clustering. The author discussed KL-Divergence and Average KL-Divergence. According to the author, KL-Divergence is not a true metric to find similarity measure as it is not symmetric. Average KL-Divergence should be used which resolved the problems with KL-Divergence. Other approaches like Information Radius, PMI, and Manhattan norm were ignored by this approach.

In another approach [3] authors are of opinion that probabilistic similarity measuring methods focused on the use of an expanded representation of texts to find similarity between them. These methods included KL-divergence [67], which was based on a ranking function using surface and expanded representation of text. In this approach again authors had ignored the usage of other probabilistic similarity measuring techniques.

In the book “Foundations of Statistical Natural Language Processing” [59] authors have discussed different Probabilistic similarity measuring techniques such as KL-Divergence, Information Radius, and Manhattan norm. According to the authors, KL-Divergence was not a reliable measure because it computes undefined results in case of maximum likelihood between distribution lists of terms from documents; and also it is not symmetric. Information Radius is a measure which solves these two problems. Manhattan norm is also a symmetric and is a measure of expected proportion of different events. According to the authors, Cosine similarity measure can also be used on a probability distribution of words from the documents between those, in which the similarity measure is required to be found.

Point-wise Mutual Information and Normalized PMI are also probabilistic [38] similarity measuring techniques which use term co-occurrences. A technique had offered a variation of PMI and Normalized PMI in the context of collocation extraction. According to authors Point-wise, Mutual Information is a measure of how much the actual probability of occurrence of particular events differs from what we would expect it to be on basis of probability of individual events with the assumption that both events are independent. Point-wise Mutual Information [5] results are in the range of -1 to +1. Normalized PMI is its normalized form to convert the result in a range of 0 to 1. The proposed approach does not use any other probabilistic similarity measure.

A technique [45] performs classification of journal papers by using their Abstracts and Bibliography lists. This technique focuses on topic models based similarity methods. These methods are based on probabilistic similarity measuring approach. The proposed technique does not use any combination of probabilistic similarity measuring techniques.

Another approach [68] uses topic model based probabilistic similarity measuring technique. The name of this technique is PMRA and it is used to search related articles from PubMed data set. Focus of this technique is on relatedness rather than relevance. An experiment performed by authors suggests that PMRA model provides an effective ranking algorithm for related articles search.

In the case of survey of probabilistic similarity measuring techniques, we have not found such a technique which used a semantic based conceptual model for a given similarity measuring technique based on the operational semantics and features used. We were also unable to find any technique which can be integrated into the formulation of a hybrid technique without redundancies and overlapping in the methods.

3.6 Citation Based Similarity Measuring Techniques

Contents of research papers representing the citation information are citation tags, citation context-based text, and bibliographic lists of research papers to find the relationships between features of research papers and methods using these features. Finding different research gaps in surveyed similarity measuring methods during this experiment was a byproduct outcome. We had surveyed different citation based similarity measuring techniques which were categorized as Citation Count based, Citation Context based, and Citation Graph-based similarity measuring techniques. Citation related contents of research papers are thought to be more reliable measure than other contents of research papers for finding similarity measures among these papers [10].

3.6.1 Finding Concepts with Disjoint and Overlapping Relationships

To answer the first question in Section 3.1, we have analyzed the citation based similarity measuring techniques by a multidimensional way, to design/classify a particular concept either with disjoint or overlapping relationships with other concepts. A secondary outcome from this analysis was identification of different research gaps in surveyed similarity measuring methods. We have selected six parameters for this analysis listed below.

1. Research Paper Sections: Representing different sections of research paper like Title, Authors, Abstract, etc.
2. Text Representation Schemes: they represent different text layouts of text contained in research papers. We have used surface text, expanded text and stemmed text categories for analysis.

3. Research Paper Weighting Schemes: they represent different weighting schemes which are used by different Citation based similarity methods. These schemes are Citation Tags, Citation Context Sentences, Bibliography Lists, and Citation Graphs of Research Paper.
4. Citation based Similarity Methods: these methods represent different citation based similarities such as Citation Graph based, Citation Context based, Direct Citation Count, Bibliographic Coupling, and Hybrid Similarity approach.
5. Entities related to Research Papers: these are divided into two categories: Research Paper contents and Entities outside the Research Paper which are used in different similarity methods.

We have analyzed different Citation based similarity measuring techniques in the form of two dimensional tables. Each table analyses the techniques by using the two parameters from the above list. For five above mentioned parameters there were 10 different combinations between each of the two dimensions. For each of these combinations a table was designed to perform analysis on similarity techniques. The cells of these tables represent the different surveyed similarity measuring techniques analyzed on the basis of two parameters selected from the above list. The grey colored cells of tables represent the research gaps found during this analysis.

Table 3.29 represents the analysis of Citation based similarity measuring techniques by using “Research Paper Sections” and “Text Representation Schemes” dimensions. This table contains 12 rows and 3 columns. The rows represent Research Paper Sections whereas columns represent different categories of Text Representation Schemes.

By observing Table 3.29, conclusions were made for Citation based similarity techniques which are listed below:

1. Text Representation schemes were not used for different individual sections of research papers for finding citation based similarity measures among them.

TABLE 3.29: Analysis using Research Paper Sections and Text Representation Schemes

Research Paper Sections	Text Representation Schemes		
	Surface Text	Expanded Text	Stemmed Text
Title			
Authors			
Affiliations/Venue			
Abstract			
Authors Provided			
Keywords			
Introduction			
Related Work			
Methodology			
Results			
Conclusions			
Bibliography	[6, 69]		
All contents (except Bibliography)	[6, 43, 60, 69, 69]		

2. Surface text based representation of Bibliography section and All contents of research papers were used by different citation based similarity measuring techniques.

Table 3.30 represents the analysis of Citation based similarity measuring techniques by using “Research Paper Sections” and “Entities related to Research Paper” dimensions. This table contains 12 rows and 2 columns. The rows represent Research Paper Sections whereas columns represent different categories of Entities related to Research Paper.

By observing Table 3.30, conclusions were made for Citation based similarity measuring techniques, which are listed below:

1. Different Citation based similarity techniques were not using Entities related to research papers for different individual sections of research papers except Bibliography.
2. Most of the Citation based similarity techniques were using all contents of research papers/documents to find similarity between them.

TABLE 3.30: Analysis using Research Paper Sections and Entities related to Research Paper

Research Paper Sections	Entities related to Research Paper	
	Research Paper Contents	Entities Outside the Research Paper
Title		
Authors		
Affiliations/ Venue		
Abstract		
Authors Provided		
Keywords		
Introduction		
Related Work		
Methodology		
Results		
Conclusions		
Bibliography	[6, 69]	[70–72]
All contents (except Bibliography)	[6, 60, 60, 69]	[1, 61–65, 70–72]

Table 3.31 represents the analysis of Citation based similarity measuring techniques by using “Research Paper Sections” and “Research Paper Weighting Schemes” dimensions. This table contains 12 rows and 4 columns. The rows represent Research Paper Sections whereas columns represent different categories of Research Paper Weighting Schemes.

By observing Table 3.31, conclusions were made for Citation based similarity measuring techniques, which are listed below:

1. Different Research Paper Weighting Schemes were not used with individual sections of Research Paper except Bibliography.
2. Research Paper Weighting Schemes were commonly used for All contents of Research Paper.

Table 3.32 represents the analysis of Citation based similarity measuring techniques by using “Research Paper Sections” and “Citation based Similarity Methods” dimensions. This table contains 12 rows and 5 columns. The rows represent

TABLE 3.31: Analysis using Research Paper Sections and Research Paper Weighting Schemes

Research Paper Sections	Research Paper Weighting Schemes			
	Citation Tags	Citation Context Sentences	Bibliography Lists	Citation Graphs of Research Papers
Title				
Authors				
Affiliations/ Venue				
Abstract				
Authors Provided				
Keywords				
Introduction				
Related Work				
Methodology				
Results				
Conclusions				
Bibliography	[6]		[6]	
All contents (except Bibliography)	[6]	[43, 60, 69]	[6]	[70–72]

Research Paper Sections whereas columns represent different categories of Citation based Similarity Methods.

By observing Table 3.32, conclusions were made for Citation based similarity measuring techniques, which are listed below:

1. Citation based similarity measuring methods were not using individual sections of research papers to find similarity measures among them.
2. Most of the Citation based similarity measuring methods were using all contents of research paper and a few were using Bibliography section of research papers.

Table 3.33 represents the analysis of Citation based similarity measuring techniques by using “Text Representation Schemes” and “Entities Related to Research Paper” dimensions. This table contains 3 rows and 2 columns. The rows represent Text Representation Schemes whereas columns represent different categories of Entities Related to Research Paper.

By observing Table 3.33, conclusions were made for Citation based similarity measuring techniques, which are listed below:

1. Surface Text Representation was used by different citation based similarity measuring techniques for finding similarity among research papers.
2. Expanded Text and Stemmed Text was not used by the surveyed Citation based similarity measuring techniques.

Table 3.34 represents the analysis of Citation based similarity measuring techniques by using “Text Representation Schemes” and “Research Paper Weighting Schemes” dimensions. This table contains 3 rows and 4 columns. The rows represent Text Representation Schemes whereas columns represent different categories of Research Paper Weighting Schemes.

By observing Table 3.34, conclusions were made for Citation based similarity measuring techniques, which are listed below:

TABLE 3.32: Analysis using Research Paper Sections and Citation based Similarity Methods

Research Paper Sections	Citation based Similarity Methods				
	Citation Graph Based	Citation Context Based	Direct Citation Count	Bibliographic Coupling	Hybrid
Title					
Authors					
Affiliations/ Venue					
Abstract					
Authors Provided Keywords					
Introduction					
Related Work					
Methodology					
Results					
Conclusions					
Bibliography			[6]	[6]	
All contents (except Bibliography)	[1, 61–65, 70–72]	[69]	[6]	[6]	

TABLE 3.33: Analysis using Text Representation Schemes and Entities Related to Research Paper

Text Representation Schemes	Entities Related to Research Paper	
	Research Paper Contents	Entities Outside the Research Paper
Surface Text	[6, 43, 60, 69]	[1, 61–65, 70–72]
Expanded Text		
Stemmed Text		

TABLE 3.34: Analysis using Text Representation Schemes and Research Paper Weighting Schemes

Text Representation Schemes	Research Paper Weighting Schemes			
	Citation Tags	Citation Context Sentences	Bibliography Lists	Citation Graphs of Research Papers
Surface Text	[6]	[6, 43, 60, 69]	[6]	[1, 61–65, 70–72]
Expanded Text				
Stemmed Text				

1. Surface Text Representation scheme is used in Citation Tags, Citation Context Sentences, Bibliography Lists, and Citation Graphs of Research Papers weighting schemes.
2. Expanded and Stemmed Text was not used in any of the mentioned Research Paper Weighting schemes.

Table 3.35 represents the analysis of Citation based similarity measuring techniques by using “Text Representation Schemes” and “Citation Based Similarity Methods” dimensions. This table contains 3 rows and 5 columns. The rows represent Text Representation Schemes whereas columns represent different categories of Citation Based Similarity Methods.

TABLE 3.35: Analysis using Text Representation Schemes and Citation Based Similarity Methods

Text Representation Schemes	Citation Based Similarity Methods				
	Citation Graph Based	Citation Context Based	Direct Citation Count	Bibliographic Coupling	Hybrid
Surface Text	[1, 61–65, 70–72]	[69]	[6]	[6]	
Expanded Text					
Stemmed Text					

By observing Table 3.35, conclusions were made for Citation based similarity measuring techniques which are listed below:

1. Surface Text based Representation of text from Research Papers was used commonly by majority of Citation based Similarity techniques.
2. Expanded and Stemmed Text representation was not used by surveyed Citation based similarity techniques.

Table 3.36 represents the analysis of Citation based similarity measuring techniques by using “Entities Related to Research Paper” and “Research Paper Weighting Schemes” dimensions. This table contains 2 rows and 4 columns. The rows represent Entities Related to Research Paper whereas columns represent different categories of Research Paper Weighting Schemes.

TABLE 3.36: Analysis using Entities Related to Research Paper and Research Paper Weighting Schemes

Entities Related to Research Paper	Research Paper Weighting Schemes			
	Citation Tags	Citation Context Sentences	Bibliography Lists	Citation Graphs of Research Papers
Research Paper Contents	[6]	[69]	[6]	
Entities Outside the Research Paper				[1, 61–65, 70–72]

By observing Table 3.36, conclusions were made for Citation based similarity measuring techniques, which are listed below:

1. Research Paper Contents are not used in Citation Graphs of Research Paper citation weighting scheme.
2. Citation Tags, Bibliographic Lists, and Citation Context Sentences belong to Research Paper Contents.

Table 3.37 represents the analysis of Citation based similarity measuring techniques by using “Entities Related to Research Paper” and “Citation based Similarity Methods” dimensions. This table contains 2 rows and 5 columns. The rows represent Entities Related to Research Paper whereas columns represent different categories of Citation based Similarity Methods.

By observing Table 3.37, conclusions were made for Citation based similarity measuring techniques which are listed below:

TABLE 3.37: Analysis using Entities Related to Research Paper and Citation based Similarity Methods

Entities Related to Research Paper	Citation based Similarity Methods				
	Citation Graph Based	Citation Context Based	Direct Citation Count	Bibliographic Coupling	Hybrid
Research Paper Contents		[6, 43, 60, 69]	[6]	[6]	
Entities Outside the Research Paper	[1, 61-65, 70-72]				

1. Citation Graph based and Hybrid similarity measuring methods were not used with Research Paper Contents.
2. Citation Context based, Direct Citation Count, and Bibliographic Coupling have used Research Paper Contents.
3. Entities Outside the Research Paper was not used in Citation Context based, Direct Citation Count, and Bibliographic Coupling methods.

Table 3.38 represents the analysis of Citation based similarity measuring techniques by using “Research Paper Weighting Schemes” and “Citation based Similarity Methods” dimensions. This table contains 2 rows and 5 columns. The rows represent Research Paper Weighting Schemes whereas columns represent different categories of Citation based Similarity Methods.

TABLE 3.38: Analysis using Research Paper Weighting Schemes and Citation based Similarity Methods

Research Paper Weighting Schemes	Citation based Similarity Methods				
	Citation Graph Based	Citation Context Based	Direct Citation Count	Bibliographic Coupling	Hybrid
Citation Tags		[69]	[6]	[6]	
Citation Context Sentences		[6, 43, 60, 69]			
Bibliography Lists			[6]	[6]	
Citation Graphs of Research Papers	[1, 61-65, 70-72]				

By observing Table 3.38, conclusions were made for Citation based similarity measuring techniques, which are listed below:

1. Citation Tags are not used in Citation Graph based and Hybrid similarity measuring techniques.

2. Citation Context Sentences were not used in Citation Graph based, Direct citation count, Bibliographic Coupling, and Hybrid similarity measuring techniques.
3. Bibliography Lists were used only in Direct Citation Count and Bibliographic Coupling similarity measuring techniques.
4. Citation Graphs of Research Paper were used in only Citation Graph based similarity measuring techniques.

3.6.2 Conclusions from Analysis of Citation Based Similarity Measuring Techniques

Conclusions were drawn from survey of Citation based similarity measuring techniques, which are discussed in below paragraphs.

Surface text based representation of Bibliography section and contents all sections of research papers were used by different citation based similarity measuring techniques. Surface Text Representation scheme is used in Citation Tags, Citation Context Sentences, Bibliography Lists, and Citation Graphs of Research Papers weighting schemes. Research Paper Weighting Schemes and Citation based similarity measuring methods were commonly used for contents all sections of Research Paper. Expanded Text and Stemmed Text was not used by the Citation based similarity measuring techniques.

Citation Tags, Bibliographic Lists, and Citation Context Sentences are Research Paper Contents. Citation Graphs of Research Paper do not belong to Research Paper Contents. Citation Context based, Direct Citation Count, and Bibliographic Coupling have used Research Paper Contents, but Citation Graph based and Hybrid similarity measuring methods have not used these contents. Entities Outside the Research Paper was not used in Citation Context based, Direct Citation Count, and Bibliographic Coupling methods.

Citation Tags are not used in Citation Graph based and Hybrid similarity measuring techniques. Citation Context Sentences were not used in Citation Graph based, Direct citation count, Bibliographic Coupling, and Hybrid similarity techniques. Bibliography Lists were used only in Direct Citation Count and Bibliographic Coupling similarity techniques. Research Paper Weighting Schemes, Text Representation schemes, and Citation based similarity methods were not used for different individual sections of research papers for finding citation based similarity measures among them.

3.6.3 Survey of Citation Based Similarity Measuring Techniques

Each of the surveyed similarity measuring technique is briefly discussed in this section. Further the answers to the second (Semantic Model) and third questions (Integration with other techniques) discussed in the Section 3.1 were answered for all these surveyed techniques at the end of this section.

Boyac et al. [6] have compared different citation count based similarity measuring techniques to cluster the biomedical literature. Four similarity techniques are discussed: co-citation analysis, bibliographic coupling, direct citation, and a bibliographic coupling-based citation-text hybrid technique. According to authors, bibliographic coupling slightly outperformed the co-citation analysis in terms of accuracy for clustering of selected repository. These techniques used the citation tags and bibliography lists of research papers, ignoring other weighting schemes.

Citation context represents text around a citation tag in a research paper, which is used by different citation based similarity measuring techniques. In this approach [69] authors have proposed a context-aware citation recommendation system. This approach represents a hybrid technique for finding similarity measure between citation contexts of two research papers. The proposed approach uses citation information, title, and abstracts of research papers, ignoring other weighting

schemes. A probabilistic model was also developed to measure the context based relevance between a citation context and a document in this work.

Citation graph of research papers represents a graph whose nodes are different research papers from a repository. There is a directed edge from the node representing a citing paper to the node which represents a cited paper. Therefore that edge is said to be a citation link and the graph is called a citation graph. Different graph algorithms were used on these citation graphs to find similarity measure among the research papers. The similarity measuring techniques surveyed below are using citation graphs of research papers.

In [70] authors analyzed the citation graph of SCOW (Computer-Supported Cooperative Work and Social Computing) conference to identify core and prominent clusters of research papers. Different research papers of this conference reside in the areas of computer and social sciences. A paper ignored in computer science can be useful in the field of social sciences. Such papers are called chasm papers and this approach found these papers. No other weighting scheme than citation graphs were used in this approach.

In an algorithm [71], a citation matrix for all documents in a corpus is defined which contains information about the documents citing other documents. A similarity matrix is generated using Cosine similarity for these documents. Similarity for documents with missing citations is difficult to compute by this approach. This technique uses citation vectors of documents rather than content vectors for computing similarity, which represents more reliable information as a feature [10].

Lu et al. [72] have proposed two graph-based metrics named as maximum flow metric and authority metric to compute the similarity between documents. Weights of edges between source and destination documents represent the maximum flow which is used to compute similarity. Each paper is then represented by a vector whose elements are authority weights of nodes in its local citation graph. The similarity is computed by using Cosine between these vectors, ignoring other similarity measuring techniques and weighting schemes.

In this approach [73] authors have proposed a supervised learning framework for citation recommendation. This approach uses Title, Abstract, Introduction sections of research papers in the form of surface text. Classifier used in this approach is based on features like citation counts of candidate papers author aware and context aware features. The proposed approach is integrated in Citeseer digital library.

Another citation based approach [69] have used abstract of research paper using bag of words along with citation context. Authors have built a context aware citation recommender system in this research, a probabilistic model was developed to measure the context based relevance between a citation context and a document.

Kataria et al. proposed a technique [46] using citation context in interlinked corpse of documents associating terms in context to the cited document. Authors proposed a document generation approach incorporating context in which a document links to another document, combines context information with link of documents.

In the case of a survey of Citation based similarity measuring techniques, we have not found such a technique which used a semantic based conceptual model for a given similarity measuring technique based on the operational semantics and features used. We were also unable to find any technique which can be integrated into the formulation of a hybrid technique without redundancies and overlapping in the methods.

3.7 Survey of Miscellaneous Similarity Measuring Techniques

The similarity measuring techniques surveyed in this section are categorized as structural, visual, and lexical similarity measuring techniques. Structural similarity represents those techniques which use XML layout of research papers. In visual similarity, the visual layout of research papers in the form of scanned images is

used to find similarity measures among them. Lexical similarity measures represents edit distance and tree edit distance similarity measuring techniques using strings and concept trees as weighting schemes respectively.

A similarity measuring technique using XML layout of documents [74] was proposed in this paper. A pair of XML documents was matched using the exact matching technique. The XML documents were also validated on basis of Document Type Definition (DTD) and XML schema. If tags were matched partially, it represented information overlapping between these documents. Tag frequency can be an additional feature to be used in process of finding similarity. This approach did not use any other weighting schemes besides XML tags.

An approach using tree edit distance algorithm [8] on XML-based research document was introduced in this paper. As XML represents a hierarchical structure, a pair of XML documents was represented as XML trees. Tree edit distance technique found similarity measures between these trees and hence computed a similarity measure between the two documents. This approach did not use any other weighting scheme beside XML tags.

A recommender system [49] also used tree edit distance similarity for finding research papers related to an author's publication interest. In this approach authors' profiles were built by using their previous publications. By comparing the profile of author with other profiles in a collection database, research papers in those profiles were recommended to that author. These profiles were maintained as a tree of concepts and tree edit distance is used as a similarity approach.

In this approach authors [7] have used the visual features of research document for their classification. Authors have devised a technique named as Visual Similarity measure which works on scanned images of documents. Visual similarity measure uses image processing techniques to extract features from a scanned document, ignoring text-based similarity measuring approaches. Weighting schemes like TF/IDF are not used by Visual Similarity measuring techniques.

A document clustering algorithm is proposed in which authors have devised a new similarity measuring technique to compute the pairwise similarity measure of text-based documents using the suffix tree document model [56]. This similarity measuring technique is utilized to devise a new document clustering algorithm. The results obtained from the author's proposed algorithm are better than the traditional TF/IDF based similarity measures. This technique does not involve the use of any combinations of similarity measuring techniques to perform the clustering operation.

In another scheme, a document classification approach is proposed by using the Wikipedia semantic space [75]. According to authors in traditional similarity measuring approaches, a document is treated as a set of words without considering the semantics between these words. Each document is represented as a concept vector in the Wikipedia semantic space. The proposed approach is used for classification of documents. This approach does not involve any combinations of different similarity measuring techniques for classification.

Another study [53] uses different vector space based similarity measuring approaches in a recommender system to generate recommendations for E-Commerce and Social web sites. Authors surveyed different similarity algorithms like Cosine, Pearson Correlation based, and Euclidean distance and their effectiveness in a recommender system. No combinations of these similarity measuring techniques have been used in the proposed approach.

In a technique [76] authors have discussed the use of Wikipedia as an external knowledge source for document clustering. According to authors, traditional approaches focus on a bag of words representation of documents without considering semantic relationships between these words. Due to the usage of such representations, documents are assigned to the wrong group/cluster. One solution to this problem is the use of ontology to enrich the document with background knowledge, but this approach has certain issues like limited knowledge base and information loss. Authors in their proposed approach have addressed these two problems by using Wikipedia as a knowledge source to cluster the documents. The similarity

measuring approach adopted uses contents, semantic, and category information of documents for clustering operation. The proposed similarity measuring technique uses exact matching and TF/IDF based matching of terms of documents within term set of Wikipedia articles. No other similarity measuring technique is harnessed in the proposed approach.

In another scheme [57] authors have proposed lexical similarity measures using dependency graph structures. Different features have been employed to compute the similarity measures such as, a bag of words, topic distribution, and dependency structures, named entities, and expansion features. The approach uses cosine similarity for vectors based on the above-described features. No combinations of other similarity measuring techniques were analyzed in this approach.

In this approach [54] authors have investigated the utility of Inclusion Index, the Jaccard Index, and the Cosine Index for calculating the similarity of documents. According to the authors, Inclusion Index provides a better similarity measure in particular when computing similarity measures using citation data. In this scheme, the comparison is performed between other similarity measuring techniques like co-word analysis, Subject-Action-Object (SAO) structures, bibliographic coupling, Co-citation analysis, and self-citation links. However, this research does not provide any similarity measuring technique as combinations of multiple techniques.

A new [55] similarity measuring technique is proposed in this research by using Cosine similarity named as Soft Cosine similarity when there is no similarity measures between the features, the proposed soft similarity becomes equal to the standard cosine similarity measure. Soft cosine similarity measuring technique is a generalized model of cosine similarity measure. Authors have proposed different formulas for exact or approximate calculations for the soft cosine measure. This research does not use different similarity measuring techniques in a combined way.

In this work [77] the fourteen existing text similarity measures are evaluated on text sentences. The evaluation was conducted on three data sets. Authors have used different vector space based similarity measures with different combinations

of features such as TF/IDF, Word Overlapping etc. These measures were evaluated on the basis of different parameters like precision, recall, rejection, accuracy etc. Similar to previous discussed technique, this scheme does not combine the similarity measures in such a way that a comprehensive similarity measure could be identified.

In this technique [78] authors have introduced a new similarity measure which combines lexical and semantic similarity measures using machine learning techniques. According to authors, the results of their experiment are close to the human judgments. The proposed technique was used for clustering of a large set of documents covering different genres and topics. This research does not involve multiple combinations of similarity measures.

3.8 Finding Relationships Between Content Based Similarity Measures

We have identified different missing features of research papers and the similarity measuring methods using these features by knowledge acquisition through Tables 3.4 to 3.38. In this section we have devised a criteria for finding relationships between content based similarity measures using this knowledge. The relationships found using this criteria are further used in conceptualization of CORES.

We have considered two parameters of a content-based similarity measuring technique to identify relationships among these techniques.

1. Which features of a research paper, this technique uses for computing similarity measure? (Document Model)
2. What is a computational method, this technique adopted for computing similarity measure? (Similarity Measure Computation Method)

Table 3.39 represents an analysis of different content based similarity measuring techniques available in the literature. This table provides information about technique name, features of a research paper this technique uses, its computational formula. This information helped us in conceptual modeling of content-based similarity measuring techniques in the CORES as discussed in Chapter 4.

Table 3.39 provides information about different content based similarity measures. We have used this information to the model domain of content based document similarity. These techniques were categorized under categories such as Vector Space-based, citation based, probabilistic, lexical, and structural similarity measures etc. By using the information from Table 3.39, overlapping and disjoint relationships between different types of content based similarity measures were identified and modeled in the proposed ontology (CORES). For identification of disjoint and overlapping relationships between similarity techniques, we have used following criteria.

We have devised an algorithm to find the relationships between different similarity techniques. Inputs for this algorithm are two similarity measuring techniques and the document weighting schemes these techniques are using. If the two similarity measuring techniques use same document weighting scheme and same computational method then these techniques will have overlapping relationship otherwise they will have a disjoint relationship. Output of algorithm will be overlapping or disjoint relationships between similarity measuring techniques. We will use this algorithm to define disjoint or overlapping relationships between similarity techniques in CORES.

3.9 Conclusions

After performing the survey of different content based similarity measuring techniques, we have reached to a number of conclusions. We have surveyed vector space based, probabilistic, citation based, and miscellaneous similarity measuring methods for knowledge acquisition of research paper similarity measures domain,

to build the proposed ontology (COReS). We have also found and reported research gaps in these similarity measuring methods by considering parameters related to different research paper features and weighting schemes, as a secondary task. After surveying these similarity measuring methods, we have concluded that none of these methods have semantically modeled the domain of research paper similarity measures. None of these methods were integrated with other techniques to formulate a hybrid technique without redundancies and overlapping in the methods and features. We have also concluded that by observing the currently available similarity measuring methods, we can identify (disjoint and overlapping) relationships between the similarity measuring techniques by observing their computational methods and document weighting schemes being used. These relationships will help us in the conceptual modeling of similarity measuring techniques in the proposed ontology (COReS).

TABLE 3.39: Analysis of Content-Based Similarity Measuring Techniques

Name of Similarity Technique	Cosine Similarity	Binary Cosine
Result Value Range	0 to 1	0 to 1
Research Paper Features used	Term Vectors in Vector Space, TF/IDF	Binary Vectors
Computational Formula	$SIMc(\vec{ta}, \vec{tb}) = \frac{\vec{ta} \cdot \vec{tb}}{ \vec{ta} \times \vec{tb} }$	$\frac{ X \cap Y }{\sqrt{ X \times Y }}$
Other Overlapping Similarity Techniques	Jaccard, Pearson correlation, Euclidean	Binary Jaccard, Matching Coefficient, Overlap Coefficient, Dice Coefficient

Name of Similarity Technique	Jaccard Coefficient	Binary Jaccard
Result Value Range	0 to 1	
Research Paper Features used	Term Vectors in Vector Space, TF/IDF	Binary Vectors
Computational Formula	$SIMc(\vec{ta}, \vec{tb}) = \frac{\vec{ta} \cdot \vec{tb}}{ \vec{ta} ^2 + \vec{tb} ^2}$	$\frac{ X \cap Y }{ X \cup Y }$
Other Overlapping Similarity Techniques	Cosine, Pearson Correlation Coefficient, Euclidean	Binary Cosine, Matching Coefficient, Overlap Coefficient, Dice Coefficient

Name of Similarity Technique	Pearson Correlation Coefficient
Result Value Range	-1 to +1
Research Paper Features used	Term Vectors in Vector Space, TF/IDF
Computational Formula	$SIMc(\vec{ta}, \vec{tb}) = \frac{m \sum_{t=1}^m w_{t,a} \times w_{t,b} - TF_a \times TF_b}{\sqrt{\left[m \sum_{t=1}^m w_{t,a}^2 - TF_a^2 \right] \left[m \sum_{t=1}^m w_{t,b}^2 - TF_b^2 \right]}}$ <p>where</p> $TF_a = \sum_{t=1}^m w_{t,a}^2 \text{ and } TF_b = \sum_{t=1}^m w_{t,b}^2$

Other Overlapping Similarity Techniques	Cosine, Jaccard Coefficient, Euclidean
--	--

Name of Similarity Technique	Euclidean Distance
Result Value Range	0 to 1
Research Paper Features used	Term Vectors in Vector Space, TF/IDF
Computational Formula	$D_E(\vec{t}_a, \vec{t}_b) = \left(\sum_{t=1}^m w_{t,a} - w_{t,b} ^2 \right)^{1/2}$
Other Overlapping Similarity Techniques	Cosine, Pearson Correlation coefficient, Jaccard

Name of Similarity Technique	Matching coefficient
Result Value Range	0 to 1
Research Paper Features used	Binary Vector
Computational Formula	$X \cap Y$

Other Overlapping Similarity Techniques	Binary Cosine, Binary Jaccard, Matching Coefficient, Overlap Coefficient, Dice Coefficient
--	--

Name of Similarity Technique	Dice Coefficient
Result Value Range	0 to 1
Research Paper Features used	Binary Vector
Computational Formula	$\frac{2 X \cap Y }{ X + Y }$
Other Overlapping Similarity Techniques	Binary Cosine, Binary Jaccard, Matching Coefficient, Overlap Coefficient, Dice Coefficient

Name of Similarity Technique	Overlap Coefficient
Result Value Range	0 to 1
Research Paper Features used	Binary Vector
Computational Formula	$\frac{ X \cap Y }{\min(X , Y)}$

Other Overlapping Similarity Techniques	Binary Cosine, Binary Jaccard, Matching Coefficient, Dice Coefficient
--	---

Name of Similarity Technique	Tree Edit Distance
Result Value Range	0 to 1
Research Paper Features used	Syntactic n -gram Trees
Computational Formula	$\delta(\theta, \theta) = 0$ $\delta(F_1, \theta) = \delta(F_1 - v, \theta) + \gamma(v \rightarrow \lambda)$ $\delta(\theta, F_2) = \delta(\theta, F_2 - v) + \gamma(\lambda \rightarrow \omega)$ $\delta(F_1, F_2) = \min \left\{ \begin{array}{l} \delta(F_1 - v, \theta) + \gamma(v \rightarrow \lambda), \\ \delta(F_1, F_2 - \omega) + \gamma(\lambda \rightarrow \omega), \\ \delta(F_1(v), F_2(\omega)) + \\ \delta(F_1 - T_1(v), T_2 - F_2(\omega)) + \\ \gamma(v \rightarrow \omega) \end{array} \right.$
Other Overlapping Similarity Techniques	Edit Distance

Name of Similarity Technique	Edit Distance
---	---------------

Result Value Range	0 to 1
Research Paper Features used	String representation of text
Computational Formula	$d_{0j} = \sum_{k=1}^j w_{ins}(a_k) \text{ for } 1 \leq i \leq m$ $d_{i0} = \sum_{k=1}^j w_{del}(b_k) \text{ for } 1 \leq j \leq n$ $d_{ij} = \begin{cases} d_{i-1,j-1} & \text{for } a_j = b_j \\ \min \begin{cases} d_{i-1,j} + w_{del}(b_i), \\ d_{i,j-1} + w_{ans}(a_j), \\ d_{i-1,j-1} + w_{sub}(a_j, b_i) \end{cases} & \text{for } a_j \neq b_j \end{cases}$ <p style="text-align: center;">for $1 \leq i \leq m$ and for $1 \leq j \leq n$</p>
Other Overlapping Similarity Techniques	

Name of Similarity Technique	KL-Divergence	KL-Divergence (General)
Result Value Range	0 to 1	0 to 1
Research Paper Features used	Probability distribution of words in documents	General Distributions
Computational Formula	$D_{KL}(\vec{t}_a \parallel \vec{t}_b) = \sum_{t=1}^m w_{t,a} \times \log\left(\frac{w_{t,a}}{w_{t,b}}\right)$	$D_{KL}(P \parallel Q) = P \log\left(\frac{P}{Q}\right)$

Other Overlapping Similarity Techniques	Average KL-Divergence, Information Radius	Average KL-Divergence, Information Radius
--	---	---

Name of Similarity Technique	Average KL-Divergence (Symmetric)	Average KL-Divergence (General)
Result Value Range	0 to 1	0 to 1
Research Paper Features used	Probability distribution of words in documents	General Distributions
Computational Formula	$D_{AvgKL}(\vec{t}_a \parallel \vec{t}_b) = \sum_{t=1}^m (\pi_1 \times D(w_{t,a} \parallel w_t) + (\pi_2 \times D(w_{t,b} \parallel w_t)))$ <p>where $\pi_1 = \frac{w_{t,a}}{w_{t,a} + w_{t,b}}$, $\pi_2 = \frac{w_{t,b}}{w_{t,a} + w_{t,b}}$ and $w_t = \pi_1 \times w_{t,a} + \pi_2 \times w_{t,b}$</p>	$D_{AvgKL}(P \parallel Q) = \pi_1 \times D_{KL}(P \parallel M) + (\pi_2 \times D_{KL}(Q \parallel M))$ <p>where $\pi_1 = \frac{P}{P+Q}$, $\pi_2 = \frac{Q}{P+Q}$ and $M = \pi_1 \times P + \pi_2 \times Q$</p>
Other Overlapping Similarity Techniques	KL-Divergence (Non-Symmetric)	KL-Divergence (Non-Symmetric)

Name of Similarity Technique	Information Radius (IR)
Result Value Range	0 to 1

Research Paper Features used	General Distributions
Computational Formula	$D(p \parallel \frac{p+q}{2}) + D(q \parallel \frac{p+q}{2})$
Other Overlapping Similarity Techniques	

Name of Similarity Technique	L1 Norm (Manhattan)
Result Value Range	0 to 1
Research Paper Features used	General Distributions
Computational Formula	$\sum_i p_i - q_i $
Other Overlapping Similarity Techniques	

Name of Similarity Technique	Pointwise Mutual Information (PMI)	Second Order Co-occurrence PMI
Result Value Range	0 to 1	0 to 1

Research Paper Features used	Co-occurring words in scientific documents	Co-occurring words in scientific documents
Computational Formula	$pmi(x; y) = \log \frac{p(x,y)}{p(x)p(y)}$ $= \log \frac{p(x y)}{p(x)} = \log \frac{p(y x)}{p(y)}$	$f(w_1, w_2, \beta_1) = \sum_{i=1}^{\beta_1} (f^{pmi}(X_i^{w_1}, w_2))^\gamma$ $f(w_2, w_1, \beta_2) = \sum_{i=1}^{\beta_2} (f^{pmi}(X_i^{w_2}, w_1))^\gamma$ $sim(w_1, w_2) = \frac{f(w_1, w_2, \beta_1)}{\beta_1} + \frac{f(w_2, w_1, \beta_2)}{\beta_2}$
Other Overlapping Similarity Techniques	Normalized Pointwise Mutual Information (NPMI)	Pointwise Mutual Information (PMI)

Name of Similarity Technique	Co-Citation Analysis	Bibliographic Coupling	Direct Citation Count
Result Value Range	0 to 1	0 to 1	0 to 1
Research Paper Features used	Citation Tags/Bibliography List/ Co-citation Frequencies	Citation Tags/Bibliography List/ Bibliographic Frequencies	Citation Tags/Bibliography List/ Direct Citation Frequencies

<p>Computational Formula</p>	$F_{ij} = \frac{1}{\log(p(C_{i,j}+1))}$ <p>where</p> $p(C_{i,j} + 1) = C_{i,j} \frac{(C_{i,j}+1)}{2}$ $K50_{i,j} = \max \left\{ \frac{F_{i,j} - E_{i,j}}{\sqrt{S_i S_j}}, \frac{F_{j,i} - E_{j,i}}{\sqrt{S_i S_j}} \right\}$	<p>Same Method used as for Co-citation analysis, by using Bibliographic frequencies</p>	<p>Same Method used as for Co-citation analysis by using Direct Citation Frequencies</p>
<p>Other Overlapping Similarity Techniques</p>	<p>Cosine Similarity</p>	<p>Cosine Similarity</p>	<p>Cosine Similarity</p>

Chapter 4

Conceptualization and Implementation of CORES

4.1 Introduction to Development of CORES

In this chapter we have discussed the development of CORES, the proposed ontology. We have used Methontology [12] technique for development of CORES. We have used the knowledge from the survey of content based similarity [12] techniques and existing ontologies from Chapter 3 and 2 to build the CORES. There are three major concept hierarchies in this ontology: a hierarchy of different content based document similarity measuring techniques, a conceptual model for pair-wise content based document similarity measures computed using these techniques, and a conceptual model for weighting schemes of research papers used in different similarity measuring techniques. First of all we have discussed the abstract level definition of these hierarchies in CORES. After that we have presented the definition (class hierarchies, properties, individuals, etc.) of CORES in Protégé environment under the guidelines by authors in their work [11]. We have also discussed SPARQL queries, ontology metrics and rules of CORES. At the end of this chapter, we have compared CORES with surveyed ontologies: SwetoDblp and SPAR ontologies on the basis of structure and domain coverage.

4.2 Abstract Definition of CORES

Abstract level representation of CORES presents the description of its major layers of concept hierarchies. A semantic model of content based similarity measuring techniques is further discussed, which represents different semantic relationships between content based similarity measuring techniques while surveying them in Chapter 3.

4.2.1 Abstract Layers of CORES

The CORES is described using a layered approach in Figure 4.1. There are two layers named as “Content Based Similarity Measuring Techniques” and “Pair-wise Research Paper Similarity Measures”. “Content Based Similarity Measuring Techniques” describes a class hierarchy to model the content based document similarity measuring techniques. The layer “Pair-wise Research Paper Similarity Measures” represents different content based similarity measures, which are computed between two pairs of documents. These measures are computed using the techniques modeled in layer “Content-Based Similarity Measuring Techniques”.

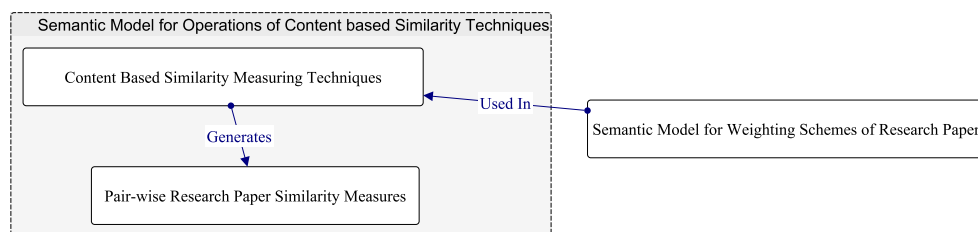


FIGURE 4.1: Abstract layers of CORES Ontology

Another layer named as “Semantic Model for Weighting Schemes of Research Paper” in this architecture describes different weighting schemes computed from contents of research paper such as TF/IDF, Term Vectors, and Citation Frequencies etc. [10]. These schemes are used by techniques modeled in “Content Based Similarity Measuring Techniques” layer, represented by “Used In” property between these layers. This layer also represents those weighting schemes which are

based on entities outside the contents of research papers. Such schemes are citation graphs of research papers, topic models, and visual layouts of research papers. These entities do not represent contents of the research paper, but they are based on external features using contents of multiple research papers in a combined way. The conceptual models of abstract layers of the CORES are explained in detail in coming sections.

4.2.2 Semantic Model for Operations of Content Based Similarity Measuring Techniques

We have proposed a semantic model for operations of content based similarity measuring techniques in this section. Figure 4.2 shows this model which is in the form of ontology and different content base similarity operations are represented as classes of this ontology. In this ontology “SubclassOf” relationships are used to represent the relationships between different superclass and subclass concepts. An example of such a relationship is between “Cosine” subclass and “Vector Space Based” super class concepts as shown in Figure 4.2. Therefore, different content based similarity measuring operations have been classified in this ontology as shown in Figure 4.2.

These similarity measuring techniques were classified as Vector Space-based, Probabilistic, Citation based, Lexical, Structural, and Visual similarity techniques. For each of these categories, their subtypes are further modeled as subclasses. For example, Vector Space-based similarity measuring technique has subclasses: Cosine, Jaccard, Dice coefficient, Matching coefficient, Overlap coefficient, Distance based, Pearson Correlation etc. In this ontology, different relationships are also defined between the classes. For example “Vector Space-based”, “Probabilistic”, and “Citation based” classes have a disjoint relationship between them. All subclasses defined for a super class have overlapping relationships between them. These relationships represent the content based similarity approaches from the **semantics-of-their-operation** point of view. Current version of CORES implements the three layers of abstract model shown in Figure 4.1 and the focus of CORES is to

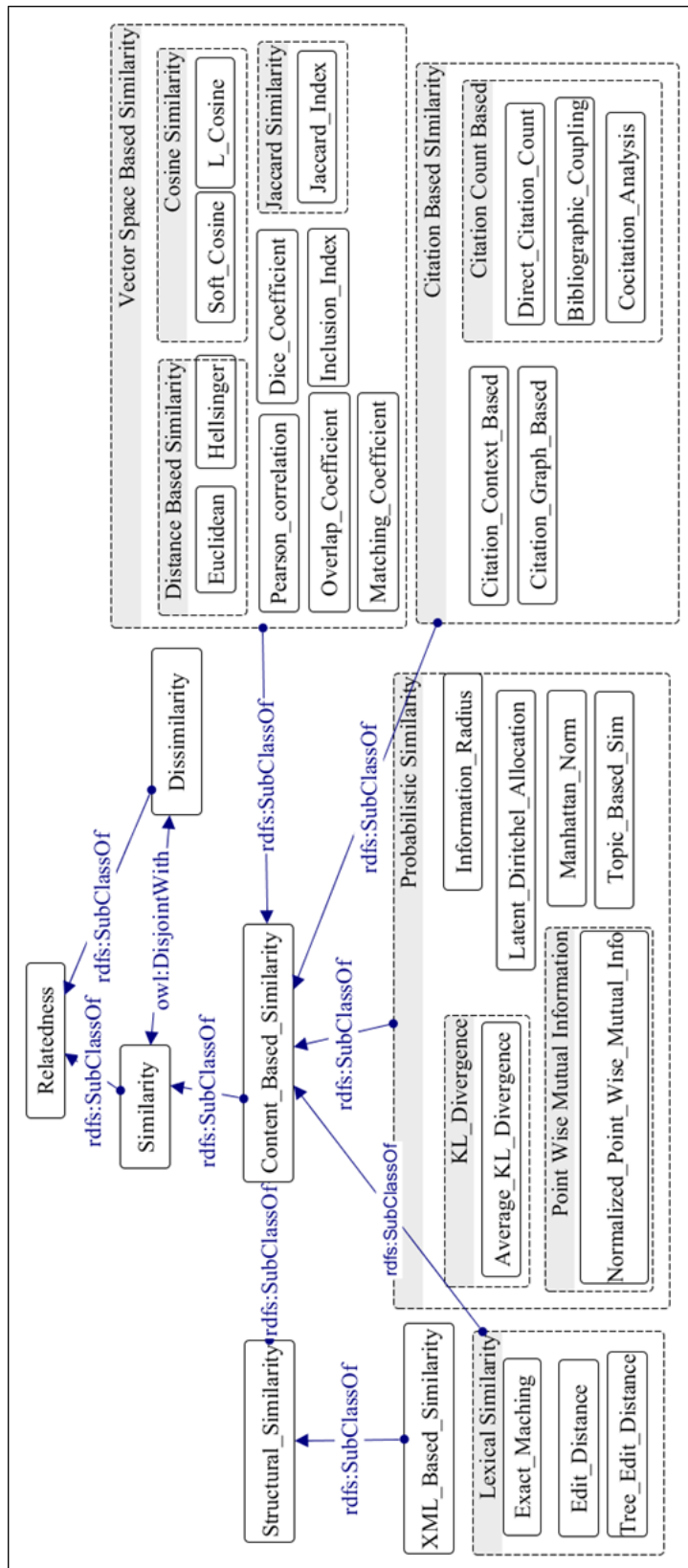


FIGURE 4.2: Semantic model for classification operations of content-based similarity measuring techniques

provide the application for computation of comprehensive similarity measures. A link to complete documentation of CORES is available in Annexure IV of Appendix section.

4.2.3 Semantic Model for Contents of Research Paper

We have proposed another semantic model to formulate the different weighting schemes for research papers which are used by content-based similarity measuring techniques. Figure 4.3 shows this model in the form of a hierarchy.

Sections of a research paper are modeled under a group named “Research Paper Contents”. Model for the text representation of contents of research papers is defined under a group named “Text Representation of Contents”. A group of concepts “Terms/Word Set” models a set of terms/words which are extracted from the different text representations of contents of research papers. From the terms/words sets, different weighting schemes are computed, used by different content based document similarity measuring techniques. For example, TF/IDF is a weighting scheme of research papers commonly used [10] in Cosine similarity measuring technique.

These different schemes are grouped in the form of concepts under a group named “Weighting Schemes”. “Entities outside the Research Paper” represents those concepts which model the different features not representing the contents encapsulated in research paper, but these features are used to compute research paper similarity measures. Such features are Citation Graphs, Synonyms of terms/keywords, research papers clusters, XML tags, Writing Style, Topic Models, Social Tags, Pseudo Documents, User Graphs, and Layout Information. Citation graphs of research papers are used in the citation based similarity measuring techniques combined with the graph algorithms.

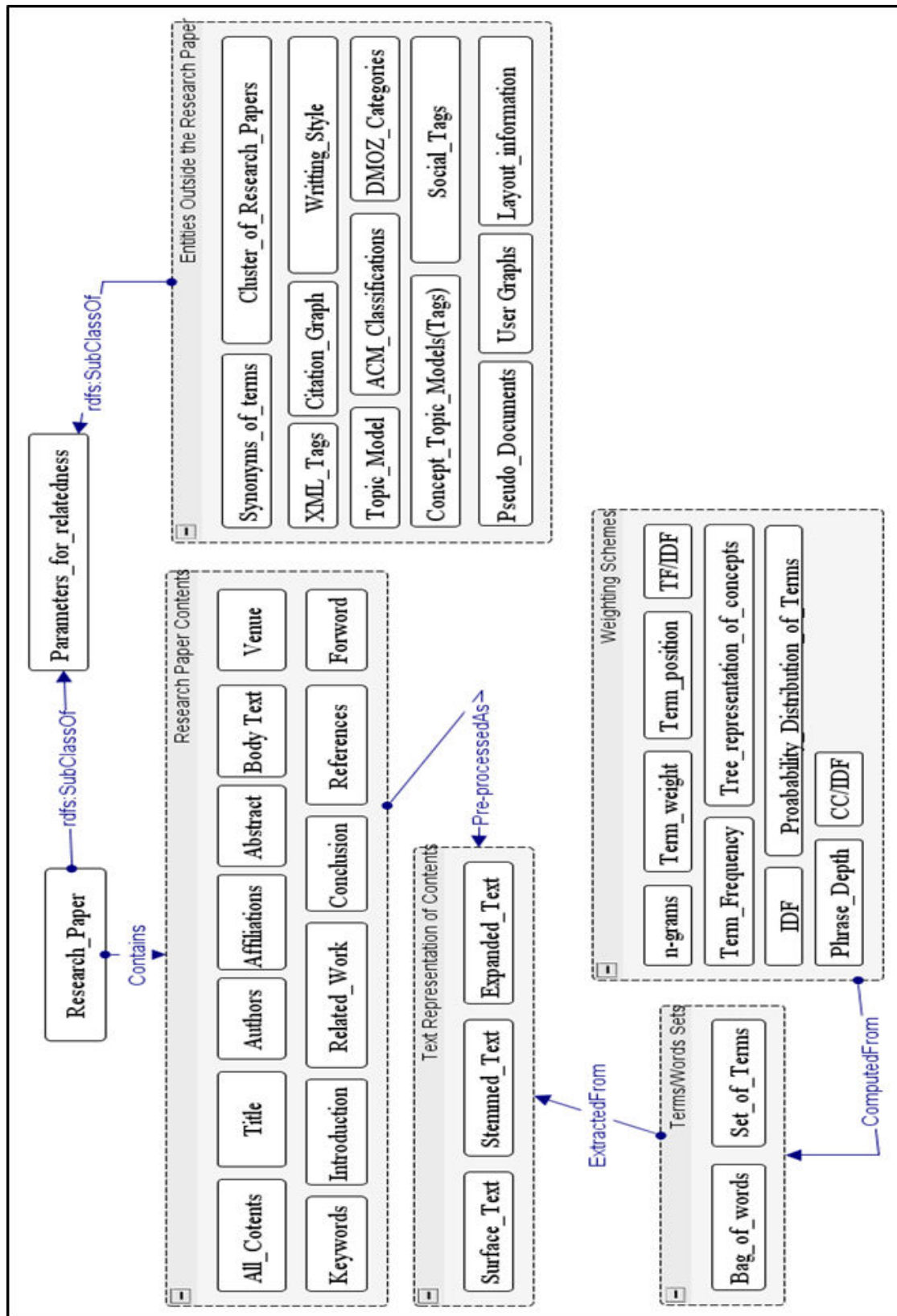


FIGURE 4.3: Semantic model for representation of contents and weighting schemes of a research paper

4.3 Definition of CORES Using Protégé

We have used Protégé 4.3 to develop CORES using OWL. We have developed class hierarchies by keeping the abstract definition of CORES in mind. These classes were defined on the basis of survey performed on the content based similarity measuring techniques and by identifying the differences and commonalities between them as discussed in Chapter 3.

Figure 4.4 represents the classes “Parameters_of_relatedness” and “Relatedness” which are subclasses of the super class “Thing”. The “Relatedness” class is further classified as “Dissimilarity” and “Similarity”.

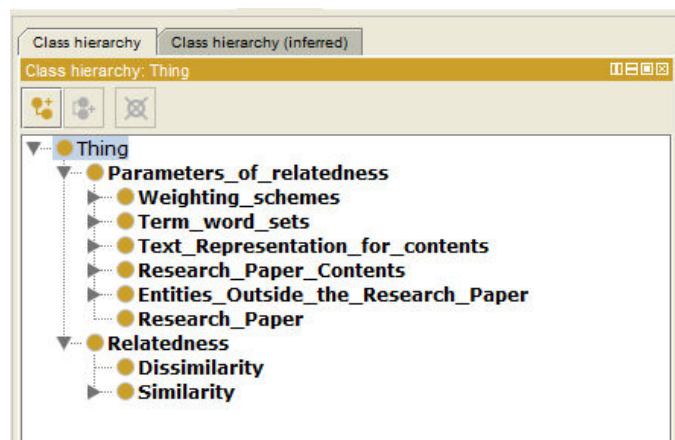


FIGURE 4.4: Top level classes in CORES as viewed in Protégé

Figure 4.5 represents the classes under the class “Content_based_similarity”. The main classes under this category are “Citation_based_similarity”, “Lexical_similarity”, “Probabilistic_similarity”, “Vector_Space_based_similarity”, “Structural_similarity”, and “Hybrid_content_based_similarity”. Classes for “Vector_Space_based_similarity” are defined to model different vector space based similarity methods like Cosine, Jaccard, Dice, and Matching coefficients. “Probabilistic_similarity” class further classifies the probabilistic similarity techniques like KL-Divergence and Pointwise Mutual Information etc. “Hybrid_content_based_similarity” class represents different content based similarity measuring techniques in which techniques from different categories are combined. “Citation_based_similarity” class represents the different techniques which use citation information of research papers to compute similarity measures between them.

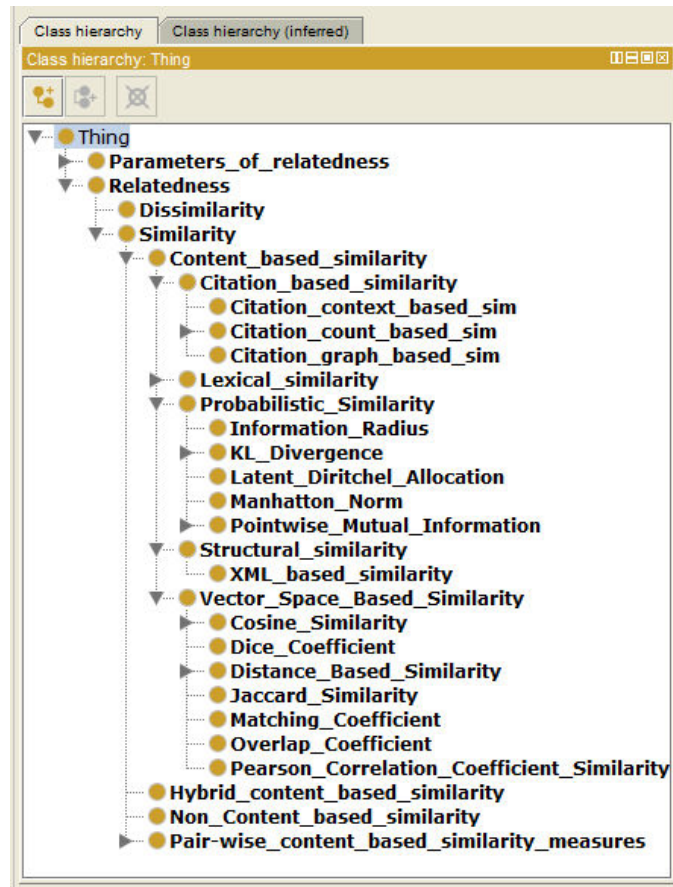


FIGURE 4.5: Content based similarity modelling classes in CORES as viewed in Protégé

Figure 4.6 represents the class hierarchy under the class “Parameters_of_relatedness”. There are five classes which are defined under this class. “Weighting_schemes”, “Research_Paper_Contents”, “Term_Word_Sets”, “Text_Representation_for_contents”, and “Entities_outside_the_Research_Paper”. “Weighting_schemes” class represents different weighting schemes of research papers which are used to compute similarity between them. “Research_Paper_Contents” class represents the contents from different sections of research papers, which are used to compute similarity measures. “Term_Word_Sets” is used to represent the different ways in which terms or keywords are extracted from research papers and are further used by weighting schemes. The class “Entities_outside_the_Research_Paper” models the information contents defined outside the structure of research papers which are used to compute similarity measures.

Figure 4.7 and 4.8 represents the object properties defined in CORES beside the

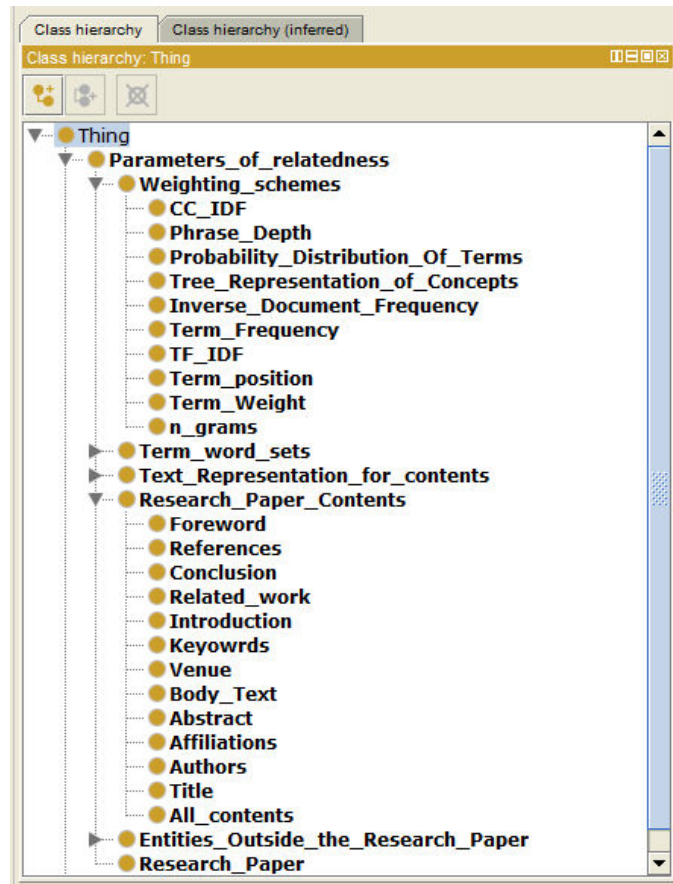


FIGURE 4.6: Research Paper content based modelling classes in CORES as viewed in Protégé

“SubclassOf” properties. There are four object properties named as: “Used_In”, “Extracted_From”, “Represented_As”, “Contains”, and “Generates”. The domain of “Used_In” is “Weighting_schemes” class and range is “Content_based_similarity”. It represents a relationship that different weighting schemes are used in different content based similarity measuring techniques. The domain of “Extracted_From” is “Term_word_sets” class and range is “Text_representation_for_contents”. This property models a relationship that term word sets are extracted from different text representations for contents of research papers. “Generates” property has the class “Content_based_similarity” as its domain whereas the range class of it is “Pair-wise_content_based_similarity_measure” as its range class. This property represents the fact that pair-wise content based similarity measures are generated by using content based similarity methods. “Represented_As” property has domain class “Research_paper_contents” and range class “Text_representation_for_contents”.

This property represents the fact that research paper contents are represented by different text representation schemes.

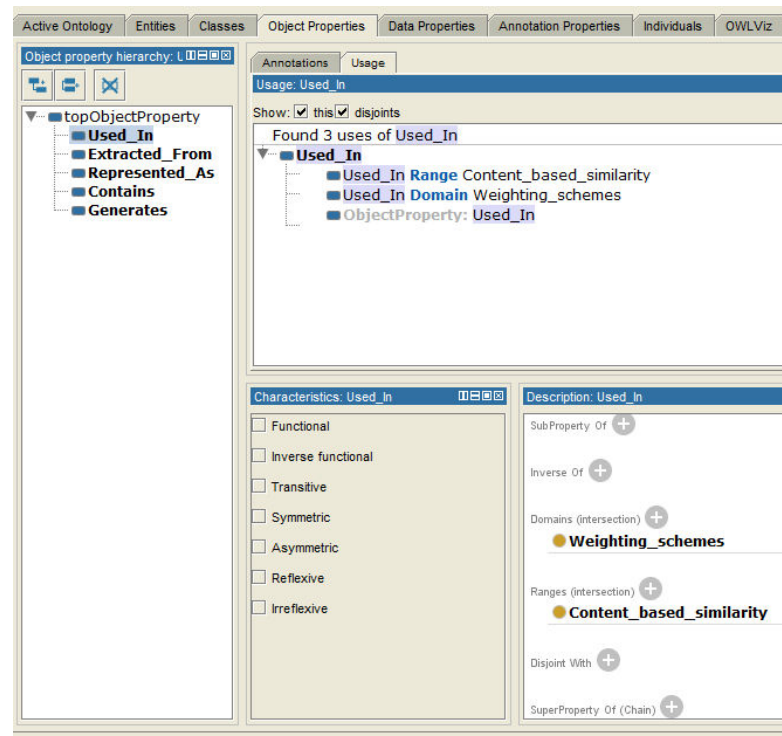


FIGURE 4.7: “Used_in” object property in CORES as viewed in Protégé

Figure 4.9 shows the data properties defined for class “Pairwise_content_based_similarity_measures”, which are “Source_Paper_Id”, “Destination_Paper_Id”, “Weighted_Sum_Parameter” and “Similarity_Value”. These properties represent the source and destination research papers and the similarity measures computed between them as numeric value from 0 to 1. The weighted sum parameter represents a parameter value for a similarity measure which is used to compute comprehensive similarity measure as discussed in Chapter 6.

Figure 4.10 shows individuals or instances defined for concepts of CORES. These individuals are divided into two categories: one represents the published similarity measuring techniques belonging to a similarity measuring method and other are generic similarity measuring algorithms. CORES is currently instantiated manually.

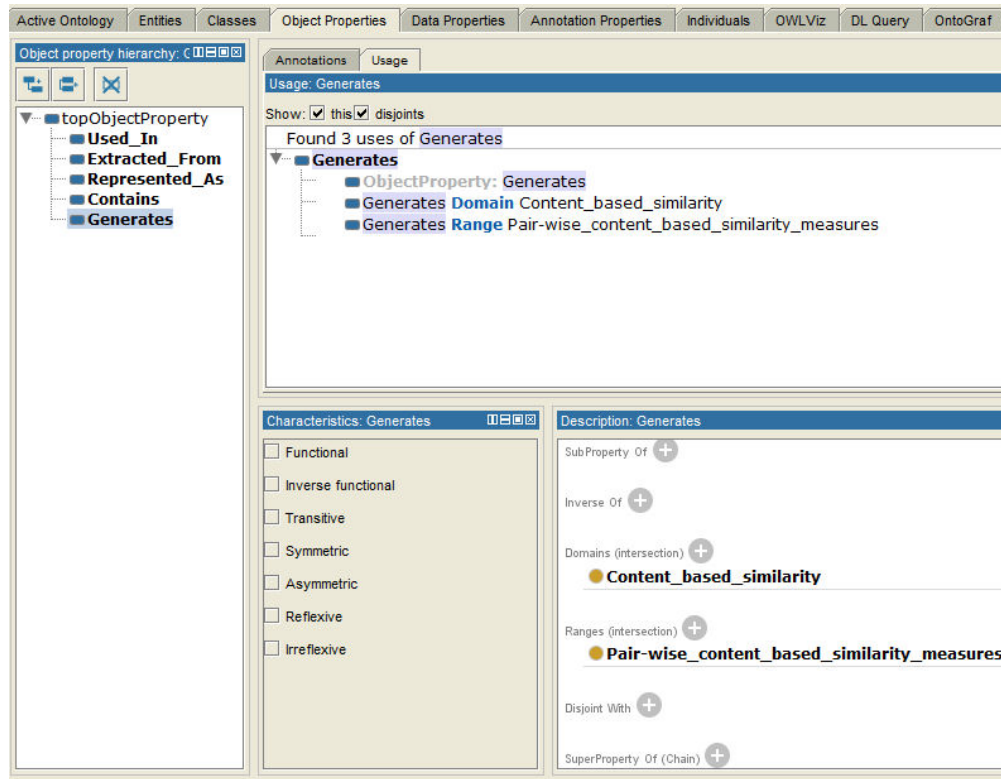


FIGURE 4.8: “Generates” object property in CORES as viewed in Protégé

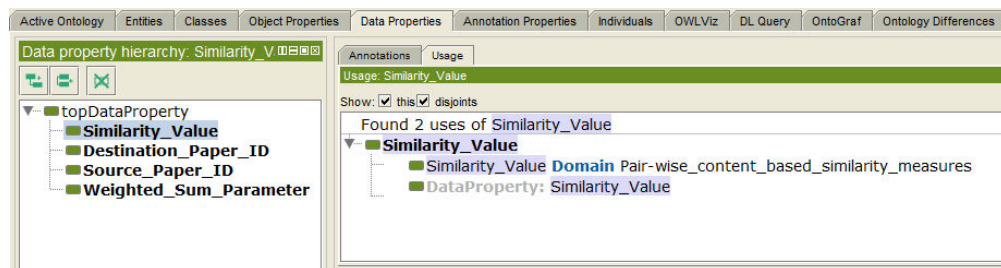


FIGURE 4.9: Data properties in CORES as viewed in Protégé

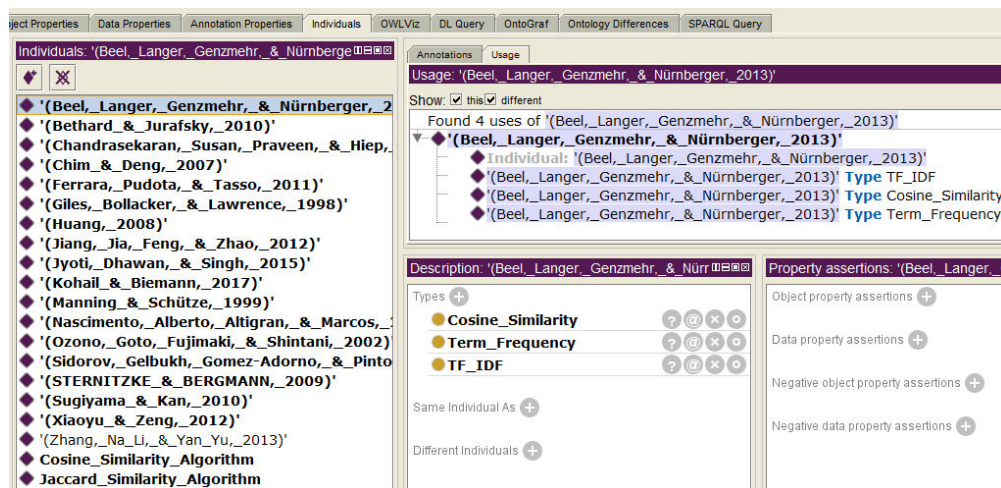


FIGURE 4.10: Individuals for concepts in CORES as viewed in Protégé

Figure 4.11 represent the visualization of COREs using ontology visualization facility available in Protégé. These visualizations represent the hierarchies of content based similarity measuring methods defined in COREs.

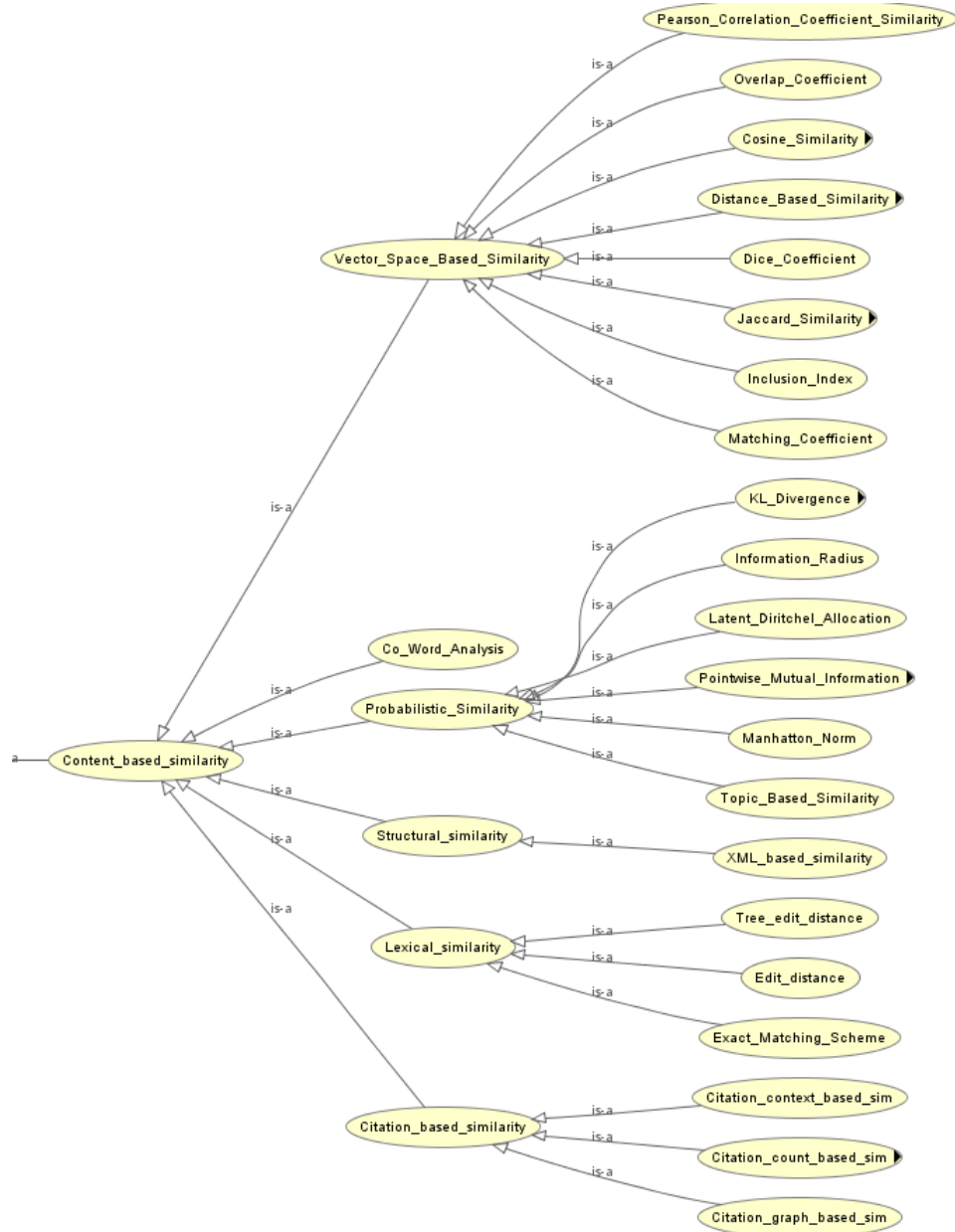


FIGURE 4.11: Visualization of content based similarity methods in COREs as viewed in Protégé

4.4 SPARQL Queries for CORES

We have performed SPARQL queries on the CORES to retrieve information from this ontology. These queries and their results have been reported in this section. Figure 4.12 represents a query to view the sub classes of a super class “Vector_Space_Based_Similarity”. The results show different sub classes for this class. Figure 4.13 represents a query to view the sub classes of a super class “Weighting_schemes”. The results show different sub classes for this class.

SPARQL query:			
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>			
PREFIX owl: <http://www.w3.org/2002/07/owl#>			
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>			
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>			
PREFIX cores: <http://purl.org/net/COREs#>			
SELECT ?subject ?object			
WHERE {?subject rdfs:subClassOf ?object. filter regex(str(?object),"Vector_Space_Based_Similarity","i")}			
subject			
Cosine Similarity	Vector	Space	Based Similarity
Overlap Coefficient	Vector	Space	Based Similarity
Inclusion Index	Vector	Space	Based Similarity
Pearson Correlation Coefficient Similarity	Vector	Space	Based Similarity
Dice Coefficient	Vector	Space	Based Similarity
Matchina Coefficient	Vector	Space	Based Similarity
Jaccard Similarity	Vector	Space	Based Similarity
Distance Based Similarity	Vector	Space	Based Similarity

FIGURE 4.12: SPARQL Query for Vector Space based similarity methods in CORES as viewed in Protégé

SPARQL query:	
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> PREFIX owl: <http://www.w3.org/2002/07/owl#> PREFIX xsd: <http://www.w3.org/2001/XMLSchema#> PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> PREFIX cores: <http://purl.org/net/COREs#>	
SELECT ?subject ?object WHERE {?subject rdfs:subClassOf ?object. filter regex(str(?object),"Weighting_schemes","i")}	
subject	
n grams	Weighting schemes
Probability Distribution Of Terms	Weighting schemes
Self Citation Link	Weighting schemes
Citation Count	Weighting schemes
Recency of Research Paper	Weighting schemes
CC IDF	Weighting schemes
Inverse Document Frequency	Weighting schemes
Subject Action Object Structure	Weighting schemes
Tree Representation of Concepts	Weighting schemes
Term position	Weighting schemes
Term Frequency	Weighting schemes
Citation Context	Weighting schemes
TF IDF	Weighting schemes

FIGURE 4.13: SPARQL Query for research paper weighting schemes in COREs as viewed in Protégé

4.5 Ontology Metrics for COREs

The ontology metrics represents entity and axiom counts for the axioms in an ontology and its imports closure. Figure 4.14 and 4.15 represents the ontology metrics for COREs. These metrics were taken from Protégé ontology metrics tab. According to these metrics there are 130 classes in this ontology along with 5 object properties and 3 data properties. There are 128 subclass axioms and 61 class assertion axioms. There are also 59 annotation assertion axioms. In this version of COREs we have not imported any other ontology yet therefore these axioms are totally related to concepts purely defined for COREs.

Ontology metrics:	
Metrics	
Axiom	428
Logical axiom count	205
Class count	130
Object property count	5
Data property count	3
Individual count	20
DL expressivity	ALC(D)
Class axioms	
SubClassOf axioms count	128
EquivalentClasses axioms count	0
DisjointClasses axioms count	3
GCI count	0
Hidden GCI Count	0

FIGURE 4.14: Ontology Metrics for CORES (First View) as viewed in Protégé

Ontology metrics:	
ObjectPropertyDomain axioms count	5
ObjectPropertyRange axioms count	5
SubPropertyChainOf axioms count	0
Data property axioms	
SubDataPropertyOf axioms count	0
EquivalentDataProperties axioms count	0
DisjointDataProperties axioms count	0
FunctionalDataProperty axioms count	0
DataPropertyDomain axioms count	3
DataPropertyRange axioms count	0
Individual axioms	
ClassAssertion axioms count	61
ObjectPropertyAssertion axioms count	0
DataPropertyAssertion axioms count	0
NegativeObjectPropertyAssertion axioms count	0
NegativeDataPropertyAssertion axioms count	0
SameIndividual axioms count	0
DifferentIndividuals axioms count	0
Annotation axioms	
AnnotationAssertion axioms count	59

FIGURE 4.15: Ontology Metrics for CORES (Second View) as viewed in Protégé

4.6 Rules for CORES

In this section, we have presented the selected axioms from TBox (terminological component) of CORES. The terminologies from T-Box are represented as Major Class Axioms and Property Axioms in the CORES. These axioms are represented using DL (Description Logic) notation which makes relationships between classes of ontology clear.

Following are T-Box axioms related to major classes of CORES. The axioms discussed below represent subsumption property between the classes. These classes are shown in semantic model for content based similarity techniques as shown in Figure 4.2.

$$\begin{aligned}
 &\text{Vector Space Based Sim} \sqsubseteq \text{Content Based Similarity} \\
 &\text{Probabilistic Sim} \sqsubseteq \text{Content Based Similarity} \\
 &\text{Citation Based Sim} \sqsubseteq \text{Content Based Similarity} \\
 &\text{Lexical Sim} \sqsubseteq \text{Content Based Similarity} \\
 &\text{Structural Sim} \sqsubseteq \text{Content Based Similarity} \\
 &\text{Hybrid Similarity} \sqsubseteq \text{Content Based Similarity}
 \end{aligned}$$

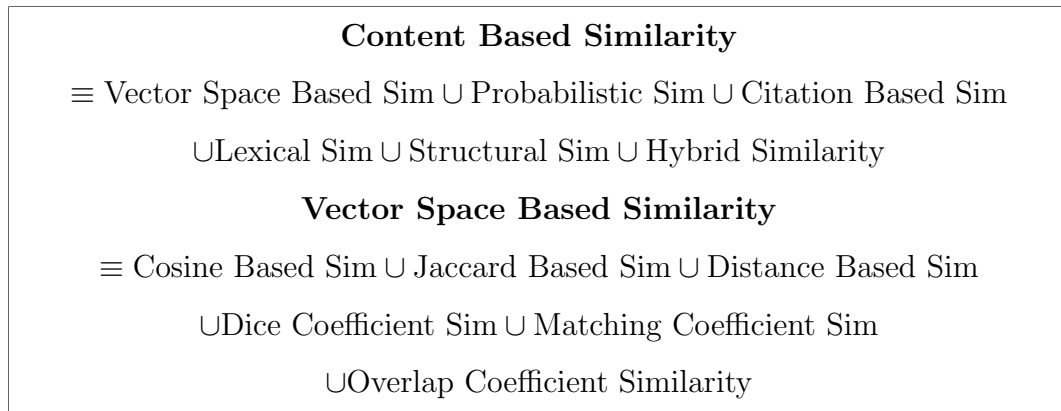
Following axioms represent the disjoint relationships between classes of CORES.

$$\begin{aligned}
 &\text{Vector Space Based Sim} \sqsubseteq \text{Content Based Similarity} \\
 &\text{Probabilistic Sim} \sqsubseteq \text{Content Based Similarity} \\
 &\text{Citation Based Sim} \sqsubseteq \text{Content Based Similarity} \\
 &\text{Lexical Sim} \sqsubseteq \text{Content Based Similarity} \\
 &\text{Structural Sim} \sqsubseteq \text{Content Based Similarity} \\
 &\text{Hybrid Similarity} \sqsubseteq \text{Content Based Similarity}
 \end{aligned}$$

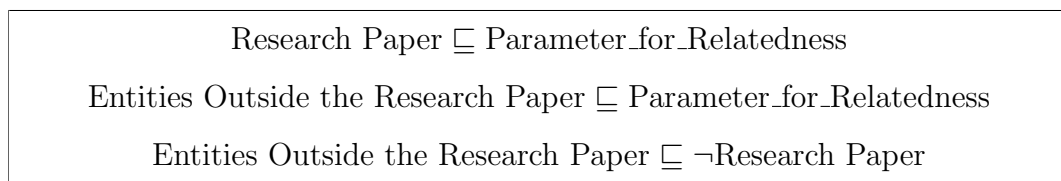
Following axioms represent the disjoint relationships between classes of CORES.

$$\begin{aligned}
 &\text{Vector Space Based Sim} \cap \neg \text{Probabilistic Sim} \\
 &\text{Vector Space Based Sim} \cap \neg \text{Citation Based Sim} \\
 &\text{Probabilistic Sim} \cap \neg \text{Citation Based Sim}
 \end{aligned}$$

Following axiom represent the composition classes of CORES.



Following axioms are related to classes represented in semantic model for weighting schemes of research papers in Figure 4.3.



4.7 Comparison of CORES With the Surveyed Ontologies

We have identified parameters for comparison of ontologies from literature to compare CORES with other ontologies. These parameters are: modeling domain, imported/reused ontologies, and research field of ontology. Other parameters [79] are also considered for this comparison. These parameters are richness, formalization, clarity criteria, extendibility criteria, implementation tools used for ontology, maximize the useful information quantity. According to clarity criteria, all the terms of ontology should be well defined using natural language and the term names of ontology should also be self-descriptive. Extendibility criteria addresses that, one should try to use the terms in an ontology from an existing vocabulary, without revising their meanings. Maximizing the useful information quantity criteria describe that information of an ontology should be complete and should not have redundancies.

Table 4.1 represents a comparison of CORES with other surveyed ontologies on the basis of discussed parameters.

TABLE 4.1: Comparison of CORES the ontology with other surveyed ontologies

Parameters used for Comparison	SPAR (Semantic Publishing and Referencing) ontologies			SwetoDblp ontology	CORES ontology
Comparison	DoCO (Document Component Ontology)	FaBiO (FRBR-aligned Bibliographic Ontology)	CiTO (Citation Typing Ontology)		
Modeling domain	Structural and Rhetorical components of document	Bibliography section of research documents	Characterization of Bibliographic Citations	Modeling the DBLP dataset of research papers	Modeling of document similarity methods and measures
					Modeling of research paper weighting schemes
Imported/reused ontologies	DEO, Pattern ontology, SALT ontology	DC Terms, PRISM, SKOS	Moved into FaBiO	FOAF, Dublin Core	No ontologies imported in the current version
Research field of ontology	Semantic Publishing			Digital Library	Document Similarity
Richness (No. of classes, properties)	No. of classes=54, No. of properties=9	No. of classes = 250, No. of properties=28	No. of Classes =9, No. of properties=96	No. of classes=7, No. of properties=9	No. of classes = 130, No. of properties =8
Formalization (Language used)	OWL 2 DL	OWL 2 DL	OWL 2 DL	OWL 2 DL	OWL 2 DL
Clarity criteria	All terms are well defined	All terms are well defined	All terms are well defined	All terms are well defined	All terms are well defined
Extendibility criteria	Using terms from existing vocabularies	Using terms from existing vocabularies	Using terms from existing vocabularies	Using terms from existing vocabularies	Not using terms from existing vocabularies
Implementation tool used for development of ontology	Not known	Not known	Not known	D2RQ for conversion of XML to RDF format	Protégé 4.3
Maximize the useful information quantity criterion	This parameter is dependent on richness parameters	This parameter is dependent on richness parameters	This parameter is dependent on richness parameters	This parameter is dependent on richness parameters	This parameter is dependent on richness parameters

From Table 4.1, we can make following conclusions about the CORES by comparing it with other surveyed ontologies.

- Neither of the surveyed ontologies was modelling the domain of research paper similarity measures, a number of these are conceptualization of semantic publishing while one is about modelling a digital library. Therefore CORES is the only ontology modelling the domain of research paper similarity measures.

- Current version of CORES is not importing/reusing any of the available ontologies as compared to surveyed ontologies.
- All the ontologies including CORES have well defined concepts.
- CORES has a rich structure as compared to DoCO, FaBiO, CiTO, and SwetoDblp.
- Since SwetoDblp is a shallow ontology which contains a few concepts and a huge number of instances in its knowledge base, so it will not be suitable to be imported in CORES.
- DoCO will be reused in future versions of CORES because modeling of document similarity measures needs concepts representing structural and rhetorical representation of research papers, which are modeled in DoCO.
- FaBiO is useful if imported in CORES, because of different similarity measuring techniques using the bibliographic information for research papers. Examples of such techniques are bibliographic coupling and co-citation analysis [6]. Therefore, FaBiO was decided to be imported in future versions of CORES.
- As citation information and citation reasoning information can be useful in computing citation based research paper similarity measures, so it would be good option to add CiTO in future versions of CORES.

4.8 Conclusions

After presenting the development of CORES we have reached to following conclusions. CORES is a richly defined ontology to model the domain of research paper similarity measures (mainly focusing on content based similarity measuring techniques). CORES have followed the proper guidelines of ontology development using a standard tool Protégé. CORES has been presented with a well-defined abstract model which makes one to easily understand its concept hierarchies. Classes,

object and data properties, and individual for CORES were defined with clarity. Proper description for all classes and properties were provided in the ontology definition. By using Protégé we have provided detailed description of CORES and visualization of its major concepts. We have also performed different SPARQL queries on CORES to explore information about its concepts and individuals. Since the CORES is defined in Protégé we can easily import different ontologies into CORES for its future enhancement. CORES was compared with SwetoDblp and SPAR ontologies to find which of these ontologies should be imported in future versions of CORES? CORES has a rich structure as compared to other surveyed ontologies which means it attempts to model the domain of research paper similarity measures with keeping completeness in mind.

Chapter 5

Evaluation of CORES

There are different ontology evaluation approaches and criteria available in literature. Raad and Cruz [17] have comprehensively discussed different ontology evaluation approaches and criteria in a survey. In order to use ontologies effectively in different applications, we need to check that whether these ontologies are “good ontologies”? For this identification we need to understand the ontology evaluation criteria and evaluation approaches. According to these criteria we have to evaluate the CORES by using ontology evaluation tools. Since these tools were unable to cover all the evaluation metrics during the evaluation process, therefore, CORES was evaluated by user study based evaluation method as well. This user study was performed to check the accuracy, completeness, clarity of CORES. Results of these evaluations are discussed in the coming sections.

5.1 Ontology Evaluation Criteria

There are different characteristics of ontologies and information provided by ontologies can be subjective because they are defined to model a specific domain. Size of ontologies can be another concern while evaluating them. Large ontologies need more processing cost and complexity which is required to be evaluated under

a criteria. Several criteria for ontology evaluation has been discussed by different approaches [12, 80–82], which is presented below.

Accuracy: This criteria state if the definitions, descriptions of classes, properties, and individuals in an ontology are correctly representing a domain. It means that axioms of ontology should obey the domain knowledge and classes of ontology should be correctly defined and described.

Completeness: This metric measure if the domain of interest is properly covered in an ontology and compare the ontology with available corpuses or gold ontologies.

Conciseness: If the ontology covers irrelevant elements with regard to domain to be modelled. For this metric redundant representation of concepts are checked and compared with gold ontologies or corpus.

Adaptability: How better an ontology is used for the tasks for which it is defined? It is recommended to use ontology in new circumstances to evaluate its performance.

Clarity: It measures how effectively the ontology communicates the intended meaning of the defined terms. Definitions should be objective and independent of the context. Concepts of ontology should be documented sufficiently and fully labelled in all necessary languages.

Computational efficiency: It measures the ability of the used tools to work with the ontology, in particular the speed that reasoners need to accomplish the required tasks for which ontology is defined.

Consistency: It describes that the ontology does not include or allow for any contradictions. There should not be any contradiction found in ontology either manually or by reasoner tools.

In principle it can be stated the ontology evaluation is a problem of assessing an ontology by the point of view of these previously mentioned criteria. Therefore to evaluate CORES the proposed ontology, we need to assess this ontology on the basis of above described metrics.

5.2 Ontology Evaluation Approaches

Ontology evaluation approaches vary on the basis that how many criteria discussed in Section 5.1 are used in process of evaluation. These approaches are divided into four categories: gold standard, corpus-based, task-based, and criteria based.

In gold standard approaches an ontology is compared with existing ontology [83] modelling the same domain. Maedche and Staab [84] consider ontologies as two layered models consisting of a lexical and conceptual layer. In corpus-based approaches which are also called data driven approaches, an ontology is evaluated about its coverage of a domain. The concept of this approach is to compare an ontology with a text corpus significantly covering a given domain. An approach [85] assess the coverage of the ontology by mining textual data from it, such as names of concepts and relations. Jones and Alani [86] use the Google search engine to find a corpus based on a user query for ranking of ontologies. After encompassing the user query using WordNet, the first 100 pages from Google results are measured as the corpus for assessment. Gold standard and corpus-based approaches practically cover the same evaluation criteria: accuracy, completeness, and conciseness.

Task based approaches try to find how an ontology helps in improving the results of a certain task. This type of evaluation considers that an ontology is intended for performing a specific type of task and is evaluated on the basis of its performance for this task. If someone designs an ontology for refining the performance of a web search engine, she may accumulate several example queries and match whether the search results contain more relevant documents if a certain ontology is used [87]. Haase and Sure [88] evaluate the worth of an ontology by defining how efficiently it allows users to obtain relevant individuals in their exploration. Task based approaches help in evaluating the adaptability of an ontology.

Criteria based approaches measure that how far an ontology or taxonomy follows a certain needed criteria. There are two types for the criteria based approaches: structure based and complex and expert based. In structure based approach an ontology is evaluated on base of its breadth, depth, and richness. Complex and expert

based approach uses complex ontology evaluation measures to assess an ontology from different aspects. As an example of structure based approach, Fernandez et al. [89] study the effect of several structural ontology measures on the ontology quality. An example of complex and expert approach, Alani and Brewster include several measures of ontology evaluation in the prototype system AKTiveRank, like class match measure, density and betweenness which are described in details in [90]. These approaches focus on evaluating clarity of an ontology. The clarity could be measured on basis of simple structure. Criteria based approaches are also helpful in detection of presence of contradictions in ontology to be evaluated, by evaluating axioms in the ontology.

5.3 Ontology Evaluation Tools

Since the complex and expert based approach suggests the usage of evaluation tools for evaluation of an ontology therefore we will discuss ontology evaluation tools which will be used for evaluation of CORES. These tools mainly focus on checking the syntax of languages to describe an ontology, such as RDFS [91], OWL [92] etc. But they also check for ontology evaluation criteria discussed in Section 5.1.

OWL 2 Validator [93] is a tool by University of Manchester used for evaluation of OWL ontologies. This tool can accept ontologies written in different formats such as RDF/XML (RDF 1.1 XML Syntax, 2014), OWL/XML [94], OWL Functional Syntax [95], Manchester OWL Syntax [96], etc. Reports from this tool can be generated in different formats such as Manchester OWL Syntax [96], DL Syntax and Functional Syntax [92]. OWL Validator uses OWL API version 3.4.5.

Hermit reasoner [14] supports OWL 2 all features. It supports object and data property classification. It also supports DL SWRL [97] rules which are ahead of current standards of OWL. It checks ontologies for inconsistency errors. Hermit is much better than Fact++ [15], and Pellet [98] as it implements hyper tableau calculus rather than tableau calculus.

ODEVAL is an ontology evaluation tool which has been developed to evaluate ontologies for errors reported by Gomez et al [12]. This tool reads ontologies defined using RDFS [91] or DAML/OIL [99]. Ontology is required to be published publicly if ODEVAL is used for its evaluation.

5.4 User Study Based Evaluation

Another task for evaluation of the ontology is by performing a user study based evaluation of CORES. In this activity, domain experts from the domain of document similarity were selected. They were provided with a questionnaire containing questions about different taxonomical errors as discussed by Gomez et al. [12] and Fahad et al. [16]. The errors reported by these experts are discussed and resolved from the ontology required to be evaluated. This questionnaire is available in Annexure 1. Questions in the questionnaire are related to ontology evaluation errors such as Circulatory, Partition, Incomplete concept classification, and Grammatical Redundancy errors as discussed by Gómez-Pérez et al [16]. Circulatory errors occur in an ontology when a class is defined as a generalization or specialization of itself. Partition errors are related to scenarios for disjoint classes in an ontology. There are two types of partition errors: common classes in disjoint composition and partitions and common instances in disjoint composition and partitions. In case of incomplete concept classification errors concepts are classified without accounting for all of them. It means that concepts in existing domain are overlooked. Grammatical redundancy errors are about redundant knowledge in ontology. Types of the redundancy errors are redundancy of subclass of relationship, redundancy of “InstanceOf” relationship, identification of formal definition of classes and instances.

5.5 Evaluation of CORES

5.5.1 Evaluation metrics and methods used for CORES

The evaluation metrics discussed in Section 5.1 have been considered for evaluation of CORES. Table 5.1 represents the evaluation metrics and evaluation methods discussed in Section 5.1 and 5.2 considered for evaluation of CORES. The grey cells represent the unavailability of a metric or approach for the evaluation purpose. For example for computational efficiency no evaluation tools are available at the moment. While observing ontology evaluation approaches in Section 5.2 we have reached to following conclusions.

CORES cannot be evaluated using any gold standard approach since there is no existing ontology to model the domain of research paper similarity measures. Corpus-based evaluation approaches are also not useful for the evaluation of CORES, due to lack of a text corpus providing information about content based similarity methods, which are conceptually modelled in CORES. Task based approaches can be used for evaluation of CORES, but already published approaches does not represent the task for which CORES is designed. i.e. computing comprehensive research paper similarity measure. For the evaluation of this metric we will use our published approach as discussed in Chapter 7 of this thesis. Since for accuracy and completeness we cannot used gold standard and corpus based approaches, therefore, we need to find other methods for evaluation of CORES for these metrics. For this purpose we have used user study based evaluation methods. Clarity and consistency are evaluated using automated evaluation based tools for CORES.

TABLE 5.1: Evaluation metrics and methods considered for CORES evaluation

Evaluation Metric	Gold standard approaches	Corpus based approaches	Task based approaches	Criteria based approaches
Accuracy	Cannot be applied due to unavailability of existing ontology from the domain of document similarity	Cannot be applied due to unavailability of text corpus from the domain of document similarity		
Completeness	Cannot be applied due to unavailability of existing ontology from the domain of document similarity	Cannot be applied due to unavailability of text corpus from the domain of document similarity		
Conciseness	Cannot be applied due to unavailability of existing ontology from the domain of document similarity	Cannot be applied due to unavailability of text corpus from the domain of document similarity		

Adaptability			Task based approach is useful for CORES for finding comprehensive research paper similarity as discussed in chapter 7	
Clarity				We can use automated tools and user study based evaluation under complex and expert based approach to find clarity
Computational Efficiency	Since no such tools are available at the moment which can use CORES therefore we cannot evaluate CORES for this metric			

Consistency				We can use automated tools and user study based evaluation under complex and expert based approach to find consistency
-------------	--	--	--	--

5.5.2 Evaluation of CORES Using Ontology Evaluation Tools

Ontology evaluation tools which have been used for evaluation of CORES are discussed in this section. OWL Validator is one of these tools, which is used for evaluation of CORES. OWL Validator performed property analysis for conceptual model of the CORES. Hermit Reasoner was used for evaluation of CORES to find errors which it can identify [14]. By using Hermit, no such errors were found in the CORES, which Hermit can identify. We have also used Fact++ reasoner which was unable to find any errors in CORES. Table 5.2 shows the results of different ontology evaluation tools used for evaluation of CORES.

By observing the results from Table 5.2, it is concluded that syntactical and semantic structure of CORES satisfies consistency, and clarity metrics. Figure 5.1 represents the usage of Hermit reasoner on CORES in Protégé 4.3 environment. As shown in the figure, while the reasoner was running but it have not found and displayed any error from CORES. Figure 5.2 also represents another reasoner Fact++ again in Protégé environment. We have evaluated the CORES by using these reasoners but have not found any errors. These reasoners have checked the

TABLE 5.2: Evaluation of CORES using Ontology evaluation tools

Tool Name	Errors which this Tool can identify	Ontology Evaluation Metrics considered	Results /Findings while evaluating CORES
Hermit Reasoner	Object and Data Properties Classifications, Basic circulatory errors	Consistency, Clarity	No errors detected
Fact++ Reasoner	Basic circulatory errors, inconsistency errors	Consistency, Clarity	No errors detected
OWL Validator	Checking for OWL 2 syntax	-	Cannot parse the ontology
ODEVAL	Checking for Circulatory, Partition and Grammatical Redundancy errors	-	Ontology was not readable by ODEVAL

CORES for consistency and clarity metrics. We have checked the proper functioning of these reasoners by inducing errors in CORES and these reasoners were found working properly.

5.5.3 User Study Based Evaluation of CORES

For user study based evaluation of CORES, a questionnaire was designed for experts from the domains of scientific document similarity measures, information retrieval, and digital libraries. This questionnaire is available in the Appendix A of the Appendix section. The questionnaire was given to five evaluators from the domain of research paper similarity measures and related domains, along with document of proposed ontology CORES (link available in Appendix B of Appendix section). The profiles of these evaluators are provided below:

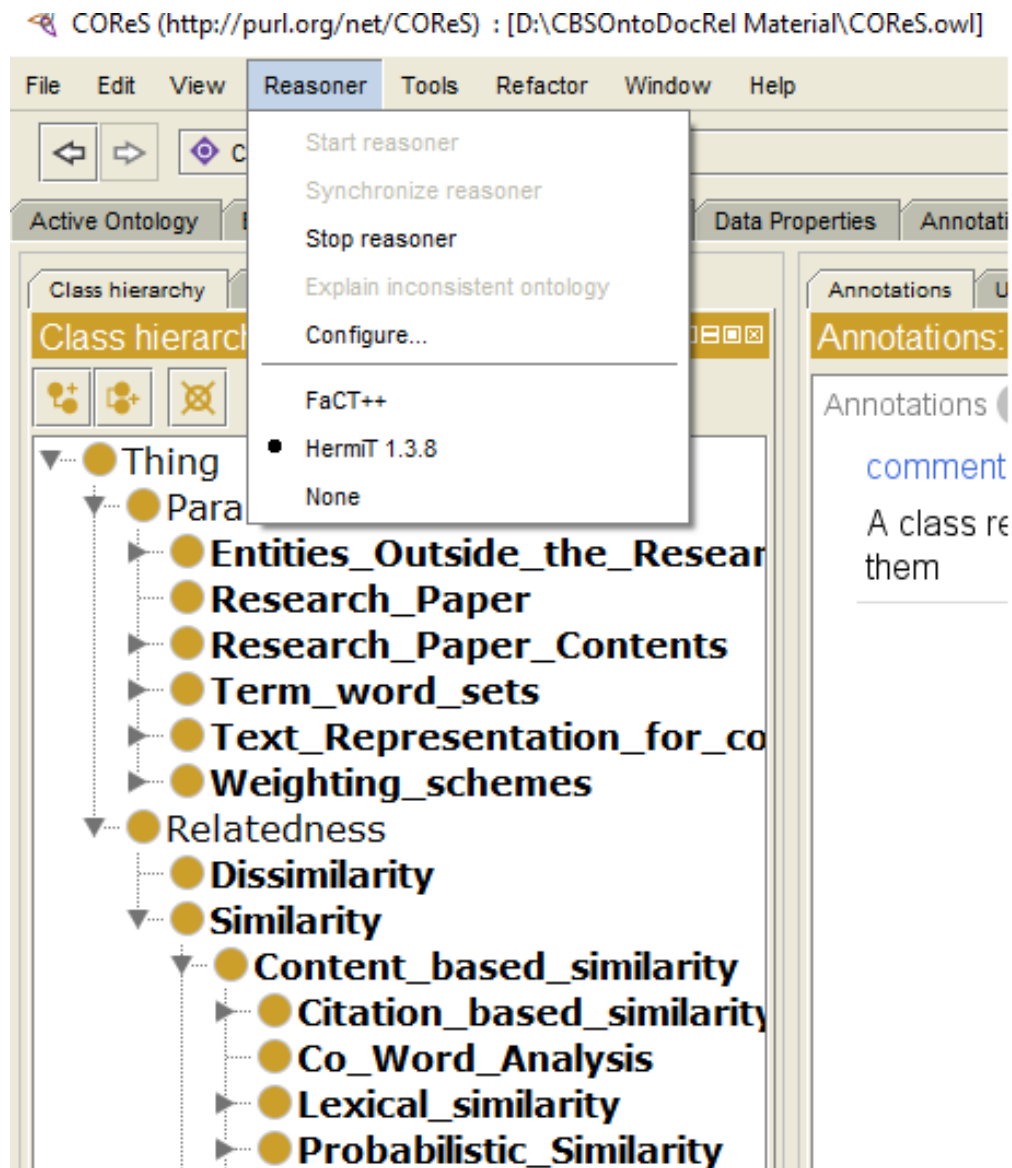


FIGURE 5.1: Hermit reasoner running on CORES without generating any errors

1. Evaluator 1: Associate professor from the domain of Digital Libraries
2. Evaluator 2: Assistant professor from the domain of Information Retrieval
3. Evaluator 3: PhD research scholar from the domain of Section wise content based similarity
4. Evaluator 4: PhD research scholar from the domain of InText citation based similarity
5. Evaluator 5: PhD research scholar from the domain of Ontology Engineering and Citation Reasons

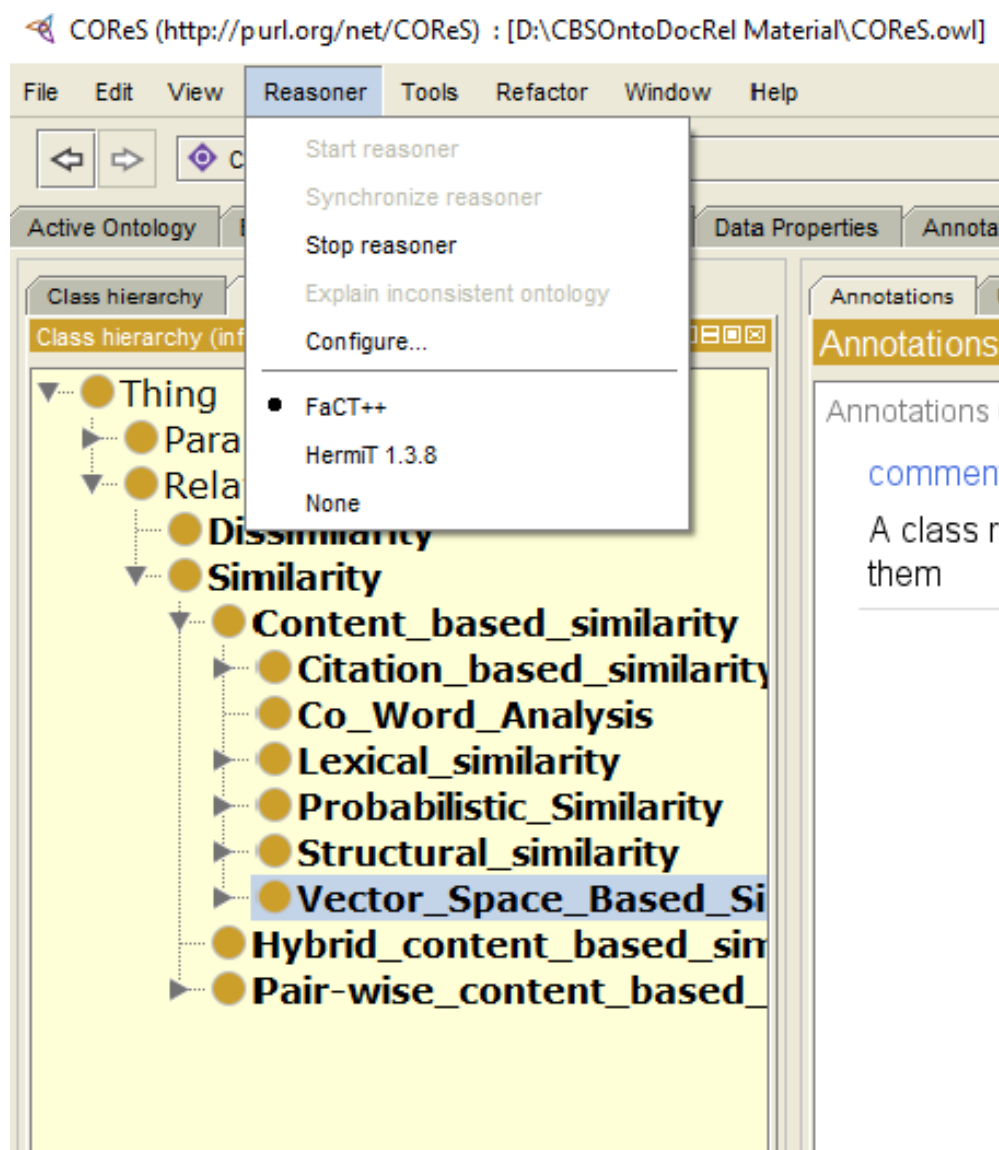


FIGURE 5.2: Fact++ reasoner running on CORES without generating any errors

Table 5.3 provides a mapping of ontology evaluation criteria discussed in section 5.1 with questions from the questionnaire. This table will help in understanding that how many metrics from ontology evaluation criteria were covered in the user study based evaluation of CORES.

Questions in this questionnaire were based on error categories devised for evaluation of ontologies. The questionnaire was prepared with objectivity by adopting the error categories identified and defined by Gómez-Pérez et al. [12] and Fahad et al. [16].

TABLE 5.3: Mapping of ontology evaluation metrics with questions from user study based questionnaire

Evaluation Metric	Mapped questions from user study based Questionnaire
Accuracy	Q.7, Q.8
Completeness	Q.1, Q.2, Q.3, Q.4, Q.5, Q.6, Q.9, Q.10, Q.11, Q.12
Conciseness	
Adaptability	
Clarity	Q.7, Q.8
Computational Efficiency	
Consistency	

Results of user study based evaluation of CORES are presented in figures 5.3 and 5.4. These plots represent the statistics of answers for objective questions from the questionnaire. Following are statements of objective questions from questionnaire:

1. Q1: Do you think that some content based similarity measure category(s) is/are missing in this ontology?
2. Q3: Is there any disjoint relationship between concepts in this ontology missing?
3. Q5: Is a class in this ontology exhaustively decomposed into its subclasses or not?
4. Q7: Is there enough description of each concept of this ontology provided?
5. Q9: Are functional properties properly defined between concepts of ontology?
6. Q11: Are there any inverse functional properties missing between concepts of this ontology?

In Figure 5.3 statistics about answers from evaluators are summarized for above discussed questions, with each of the questions their desirable outcome is mentioned. From Figure 5.3, it is clear that CORES satisfies the completeness metric by this evaluation with an average of 74%.

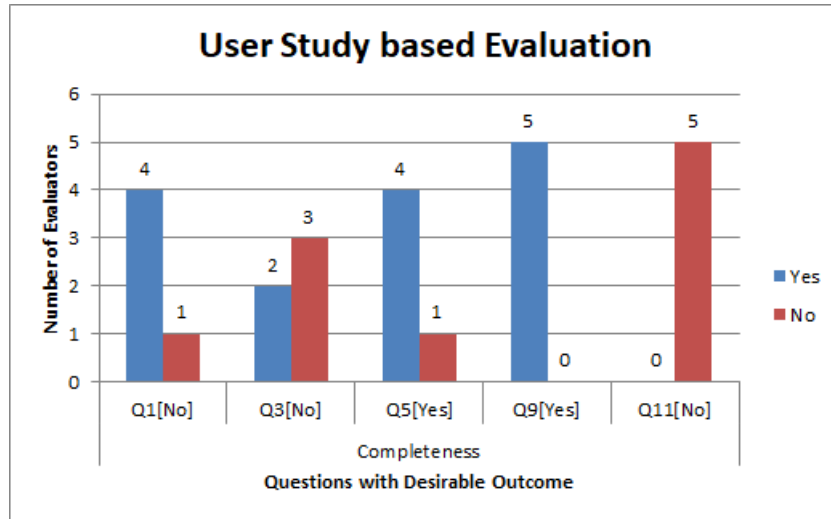


FIGURE 5.3: Plot of User Study based evaluation of COREs for completeness

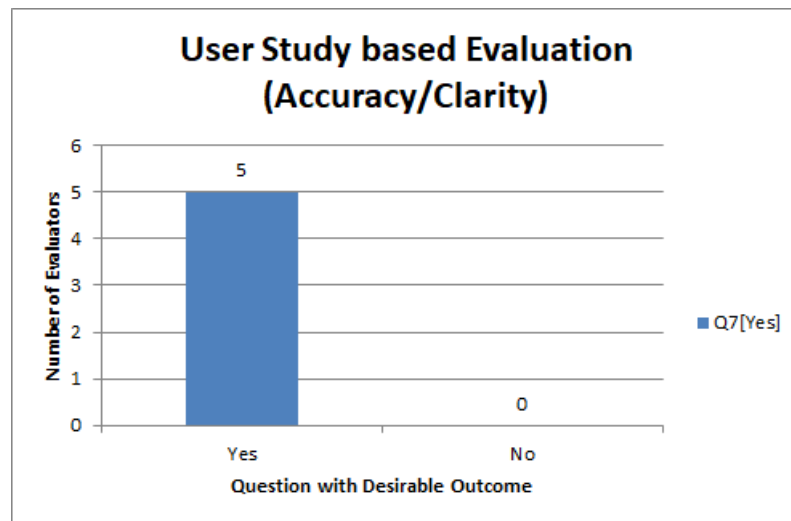


FIGURE 5.4: Plot of User Study based evaluation of COREs for accuracy and clarity

In Figure 5.4 statistics about answers from evaluators are summarized for question Q7 with its desirable outcome is mentioned. From the Figure 5.4 it is clear that COREs satisfies the accuracy and clarity metrics by this evaluation with an average of 100%.

Table 5.4 presents evaluators’ findings after user study based evaluation of COREs.

In the case of “Incomplete Concept Classification” errors according to experts’ evaluation they have recommended the addition of subclasses in the class hierarchy of content-based similarity measures. In the response to their evaluation, it can be

TABLE 5.4: Findings after user study based evaluation of COREs

Category of Error to be evaluated	Missing Concept	Associated Concept/Super Class	Involved evaluation metrics from Section 5.1	Experts' Recommendation
Incomplete Concept Classification	Structure Based Similarity	Content Based Similarity	Completeness Accuracy	The missing concept should be available in the Ontology
Partition Errors: Exhaustive Knowledge Omission	-	-	Completeness Accuracy	The Ontology should be published publicly and from feedback from the public can be used in future to define the concepts of ontology by resolving "Exhaustive Knowledge Decomposition" problem from this ontology.
Sufficient Knowledge Omission	-	-	Completeness Accuracy	According to experts, there are no such concepts in COREs, which are not described properly.

observed in Figure 5.5, that there is a concept "Section Wise" in the gray colored box, which is a subclass of "Content Based Similarity" concept in COREs. This concept is defined for modeling of structure-based similarity measures which focus on sections of research papers for finding similarity between them.

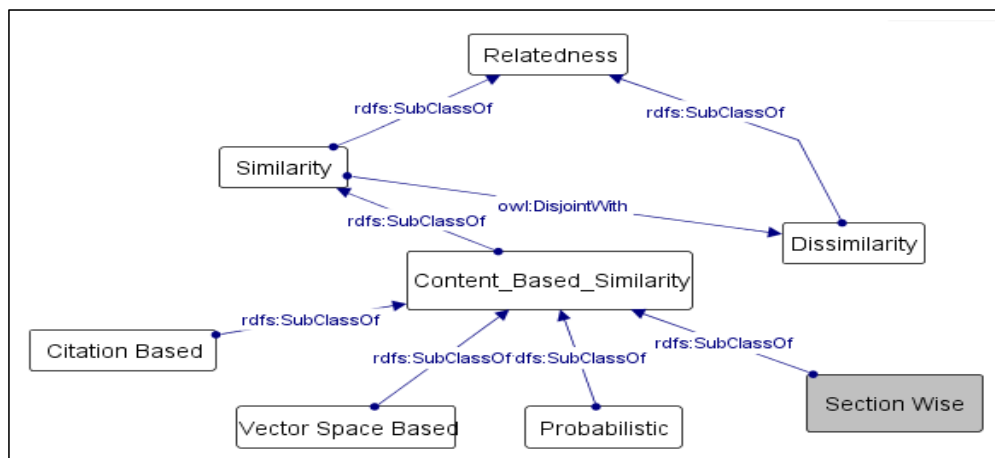


FIGURE 5.5: "Section Wise" similarity concept addition in the COREs

In the case of "Exhaustive Knowledge Omission" error category, experts are of

opinion to publish CORES publicly, so that the public may provide their feedback regarding this error category in this Ontology. This definition of CORES is available under a file name “CORES.owl”, which is uploaded on GitHub on the following link.

<https://github.com/QamarPC103006/CORES>

The errors reported by the public feedback will be rectified in future version of CORES. All the other changes suggested by domain experts during this user study based evaluation were incorporated in the CORES.

5.6 Conclusions

After evaluation of proposed ontology (CORES), we have reached to a number of conclusions. CORES was evaluated for ontology evaluation metrics as reported in literature. We have used two ontology evaluation tools (Hermit and Fact++ reasoners) and user study based evaluation method to evaluate this ontology. CORES was found to be consistent ontology which is defined with clarity. We have also evaluated CORES for accuracy, completeness, and clarity during user study based evaluation. The missing concepts in CORES, which were pointed out by domain experts, were added in the ontology. CORES is provided on Internet for public feedback to improve its completeness. Adaptability for CORES was evaluated by testing it for applications for computation of comprehensive research paper similarity measure. Chapter 6 and Chapter 7 will be evaluating CORES for this metric.

Chapter 6

Comprehensive Research Paper Similarity Measure-Application of COrReS

This chapter is about applications of COrReS and how COrReS can be used to perform different tasks? We will initially discuss different possible application scenarios in which COrReS can be used. Our main focus will be on application, using knowledge from COrReS, to compute research paper similarity comprehensively by combining different content based similarity measuring techniques. This application of COrReS is portrayed with an abstract level algorithm, which is further discussed in concrete form by four use cases. These use cases will represent a specific scenario, in which concepts and their instances from COrReS will be used to compute the comprehensive research paper similarity measure as a weighted sum of different content based similarity measures.

6.1 Different Possible Applications of COrReS

COrReS can be used in different applications to perform different tasks and a number of possible potential applications of COrReS can be:

1. Recommender systems [10] which are helpful in recommending different research papers similar to a research paper provided as a query paper by a researcher. The knowledge about relationships between similarity techniques and weighting schemes of research papers can help in building a recommender system to select research papers in intelligent way by considering profile of researcher.
2. Plagiarism testing systems [100], which may find that whether two research papers are same, and up to how much extent, by using different similarity measuring techniques, not just relying on string based similarity measures. CORES the proposed ontology can provide information about weights of similarity measures, which can be very useful in selection of similarity measuring techniques for plagiarism testing.
3. Identification of similar patents for products [101], so that it can find that which product have copied how many features from any other product.
4. Research paper clustering applications [37], CORES can be used to associate a research paper with a specific cluster/group on the basis of a document similarity measuring category, discussed in that document, and for defining new clusters of research papers. Such clusters of research papers can be useful for researchers, who will be trying to invent new document similarity measuring technique.
5. Development of a framework to compute research paper similarity measures in a comprehensive manner using knowledge from CORES. Semantic relationships between different content based similarity measuring techniques (modeled in CORES) can be used by combining these techniques. This scenarios is discussed in detail with an abstract level algorithm as well as in the form of four use cases in coming sections of this chapter.

6.2 Algorithm for Computation of Comprehensive Similarity Measure

Figure 6.1 represents the abstract view of algorithm for computation of comprehensive similarity measure. A database/repository containing information about 70 plus similarity measuring techniques is used by CORES ontology by modeling (disjoint and overlapping) relationships between these similarity measuring techniques. Pairwise similarity measures among research papers from another data set is computed and stored in the knowledge base of CORES. The similarity measures for overlapping similarity measuring techniques can be calculated using average values or other summation formulas. These measures are represented as $O_Sum_1, O_Sum_2, O_Sum_3, \dots, O_Sum_N$. These measures are further titled as $S_1, S_2, S_3, \dots, S_N$ in next module for comprehensive similarity measure computation. The similarity measures for disjoint similarity measuring techniques are computed as weighted sum of the overlapping similarity measures using weights $W_1, W_2, W_3, \dots, W_N$ (summing up to value 1). A weight tuning system will tune these weights to compute the comprehensive similarity measure with a better accuracy. The weight tuning system can tune the weights using knowledge from people working in domain of research paper similarity measures using techniques like crowd-sourcing. We have developed four use cases, discussed in the coming sec-

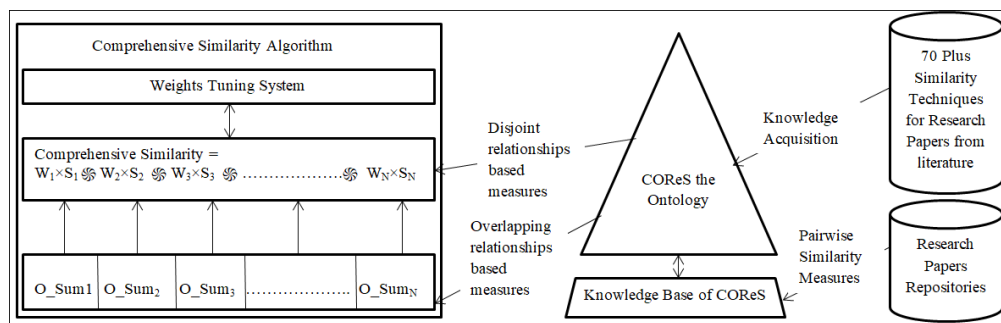


FIGURE 6.1: Algorithm for comprehensive similarity computation, an abstract view

tions, to compute comprehensive similarity measure as a specific case study from this abstract representation of comprehensive similarity measuring algorithm.

6.3 Use Cases for CORES

An approach adopted by Shekarpour et al. [102] was source of inspiration to design four use cases for CORES to demonstrate its application for computing research paper similarity measure in a comprehensive way. The authors have demonstrated the application of their proposed ontology CEVO by use cases for the achievement of text annotation task. Therefore by adopting this idea we have demonstrated the application of CORES to compute research paper similarity measure in a comprehensive way by designing uses cases, which utilize the layer based abstract model of CORES presented in Figure 4.1. One of these layers represents “Content-Based Similarity Measuring Techniques” and other is “Pair-wise Research Paper Similarity Measures”. There is a property named “Generates” shown in Figure 4.1 whose domain is all classes from the layer “Content-Based Similarity Measuring Techniques” and the range is all classes from the layer “Pair-wise Research Paper Similarity Measures”. Each use case is based on conceptual modelling of these two layers and their knowledge bases. These models are discussed in detail for each use case in succeeding sections.

6.3.1 Use Case-1: Vector Space Based Similarity Measures Computation

This use case describes the computation of document similarity measures using Vector Space-based similarity measuring techniques. Figure 6.2 shows this use case, it contains two layers: one represents a conceptual model of Vector Space-based Similarity Measuring techniques while the other one shows the conceptual model of Pairwise Vector Space-based Similarity Measures. These conceptual models show different classes/concepts. Classes from a conceptual model of Vector Space-based similarity measuring techniques, represent different similarity measuring techniques, using Vector Space model of research papers. Table 6.1 represents typical pair-wise Vector Space-based similarity measures which are also shown in Figure 6.2.

As these measures are generated by Vector Space-based Similarity Measuring Algorithms and according to CORES, these techniques have an overlapping relationship with each other. The computations for Vector Space-based similarity measures are represented by the Equation (6.1).

$$PSim_{VSBM} = \sum_{i=1}^{T_{VSBM}} \frac{PVSM_{Li}}{T_{VSBM}} \quad (6.1)$$

In (6.1) $PSim_{VSBM}$ represents Comprehensive Vector Space Based Pair-Wise Similarity Measure while $PVSM_{L_i}$ represents the label of a Pair-wise Vector Space-based Similarity Measure computed using a specific technique. T_{VSBM} represents the total number of Vector Space-based Similarity Measuring techniques used for the computation. It is assumed that each of these similarity measures has even weight while contributing in process of computation. We may calculate the $PSim_{VSBM}$ for typical values from Table 6.1 by following expression using (6.1). The result value represents an average of all these Vector Space-based similarity measures.

$$PSim_{VSBM} = \frac{0.56 + 0.63 + 0.37 + 0.42}{4} = 0.49 \quad (6.2)$$

TABLE 6.1: A table representing Vector Space-based document similarity techniques with gap areas

Label of Similarity	Similarity Name	Value	Description
$PVSM_{L1}$	$PSim_{Cosine}$	0.56	Cosine Similarity Measure
$PVSM_{L2}$	$PSim_{Jaccard}$	0.63	Jaccard Similarity Measure
$PVSM_{L3}$	$PSim_{EuclidDist}$	0.37	Euclidean Distance Similarity Measure
$PVSM_{L4}$	$PSim_{Hellsinger}$	0.42	Hellsinger Distance Similarity Measure

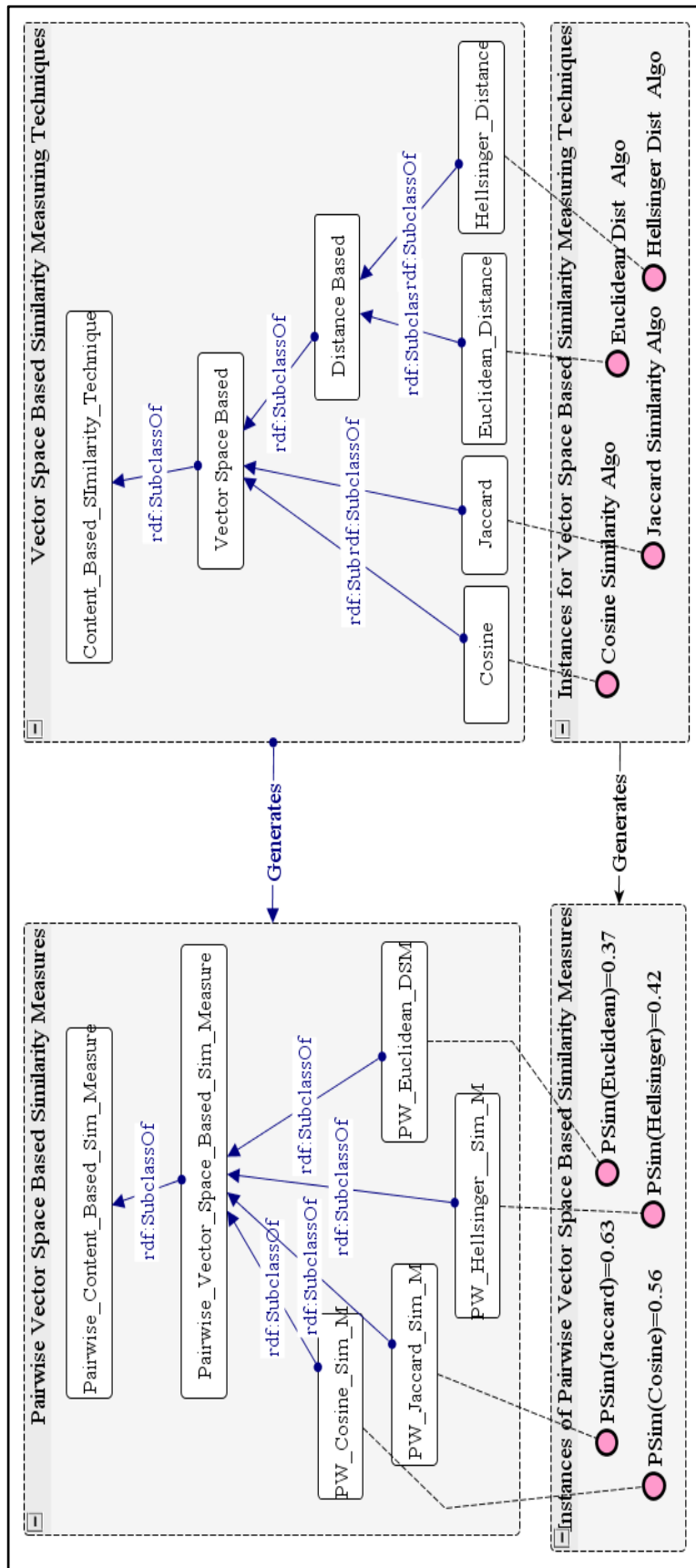


FIGURE 6.2: Use Case-1 representing usage of Vector Space-based Similarities from COREs

6.3.2 Use Case-2: Probabilistic Similarity Measures Computation

This use case deliberates the computation of Probabilistic Similarity measures which use different Probabilistic similarity measuring techniques for computation of similarity measures among research papers. In Figure 6.3 pair-wise Probabilistic Similarity measures are shown beside the conceptualization of Probabilistic similarity measuring techniques through which these measures were generated as shown in Table 6.2.

TABLE 6.2: Pairwise Probabilistic Similarity Measures

Label of Similarity	Similarity Name	Value	Description
P_{P_1}	$PSim_{AvgKLD}$	0.38	Average KL Divergence Similarity Measure
P_{P_2}	$PSim_{IRadius}$	0.33	Information Radius based Similarity Measure
P_{P_3}	$PSim_{ManhattanNormal}$	0.42	Manhattan Normal Form Similarity Measure

According to the conceptual model of CORES, there are overlapping relationships between the subclasses of Probabilistic similarity measuring techniques. Therefore the pair-wise probabilistic similarity measures can be used to compute similarity in a comprehensive way by taking an average of these measures. This computation is formulated in the Equation (6.3)

$$PSim_{Probabilistic} = \sum_{i=1}^{T_{Probabilistic}} \frac{P_{P_i}}{T_{Probabilistic}} \quad (6.3)$$

Equation (6.3), $PSim_{Probabilistic}$ represents the Comprehensive Probabilistic pair-wise Similarity Measure P_{P_i} represents the label of a Pair-wise Probabilistic Similarity Measure computed using a specific technique and $T_{Probabilistic}$ represents the total number of Probabilistic Similarity Measuring techniques used for the computation. It is assumed that each of these similarity measures has even weight in

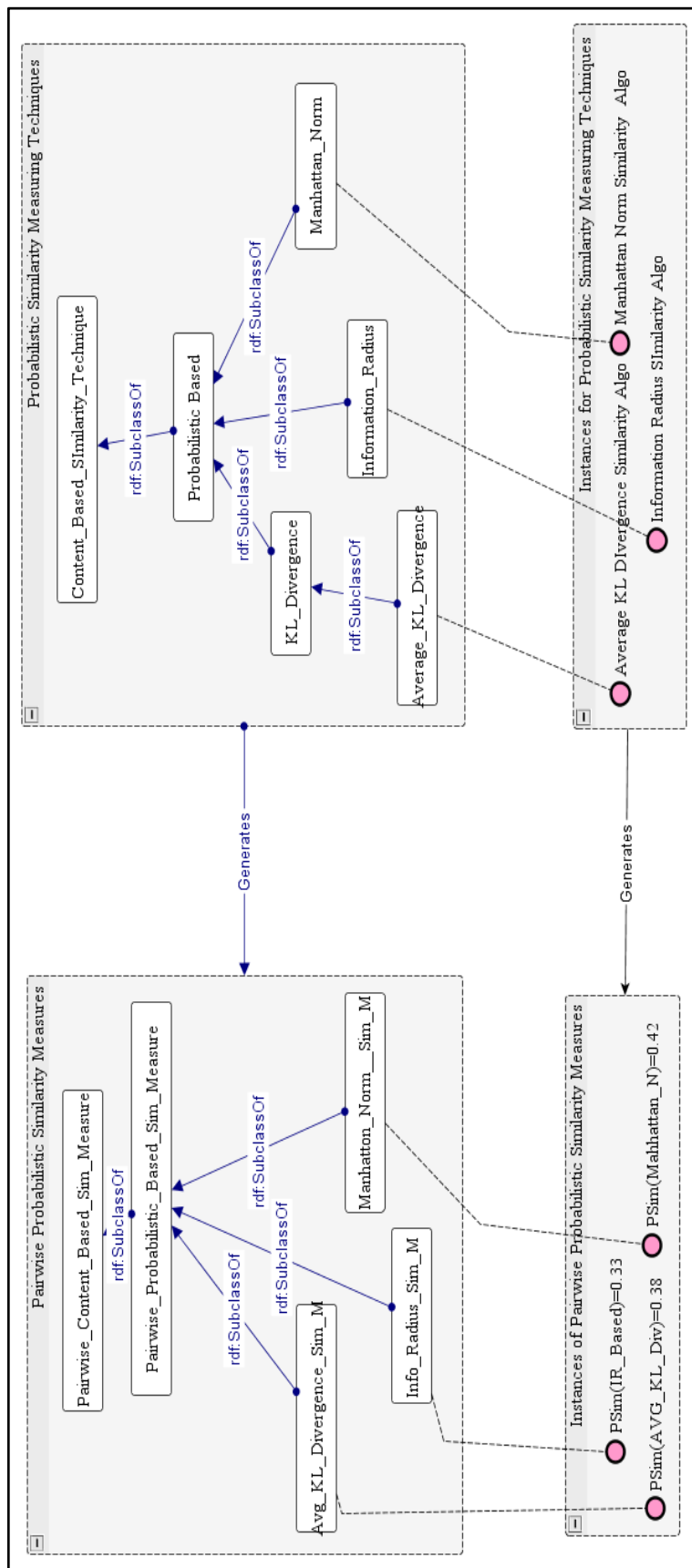


FIGURE 6.3: Use Case-2 representing usage of Probabilistic Similarities from CORES.

process of computation. We may calculate the $PSim_{Probabilistic}$ for typical similarity measures in Table 6.2 by following expression using Equation (6.3).

$$PSim_{Probabilistic} = \frac{0.38 + 0.33 + 0.42}{3} = 0.37 \quad (6.4)$$

6.3.3 Use Case-3: Citation-Based Similarity Measures Computation

In this use case, Citation based similarity measures are used to compute the pair-wise similarity measures between the research papers. Figure 6.4 represents a class hierarchy modeling pair-wise citation based similarity measures and the similarity techniques computing these measures. Table 6.3 describes typical pair-wise Citation based similarity measures computed for a pair of research papers.

TABLE 6.3: Pairwise Citation based Similarity Measures

Label of Similarity	Similarity Name	Value	Description
P_{C1}	$PSim_{BibliographicCoupling}$	0.42	Bibliographic Coupling based Similarity Measure
P_{C2}	$PSim_{DirectCitation}$	0.37	Direct Citation based Similarity Measure
P_{C3}	$PSim_{CocitationCount}$	0.53	Cocitation Count based Similarity Measure

There are overlapping relationships between the subclasses of citation based similarity measuring techniques. Therefore pair-wise citation based similarity measures can be used to compute similarity by taking an average of these measures as represented in Equation (6.5).

$$PSim_{Citation} = \sum_{i=1}^{T_{Citation}} \frac{P_{C_i}}{T_{Citation}} \quad (6.5)$$

Equation (6.5), $PSim_{Citation}$ represents Comprehensive Citation based Pairwise Similarity Measure while P_{C_i} represents the label of a Pair-wise Citation based

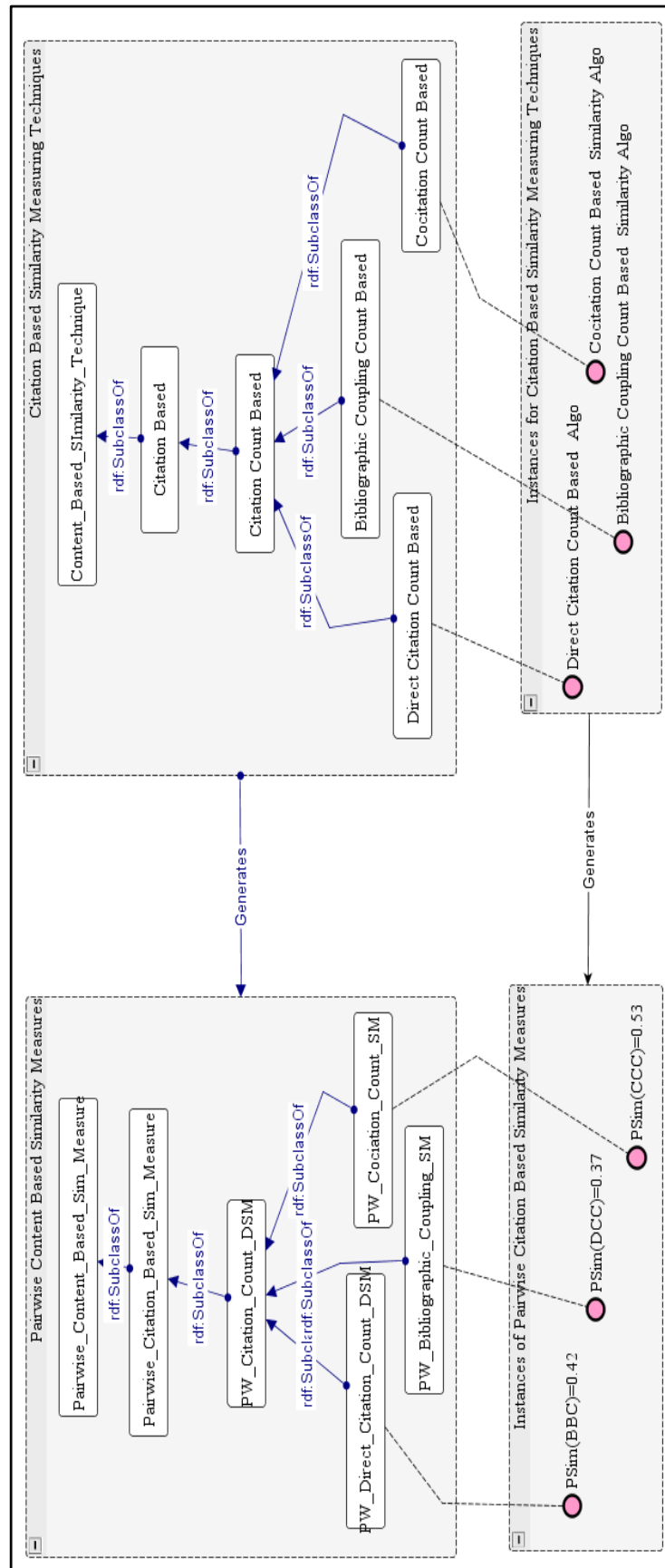


FIGURE 6.4: Use Case-3 representing usage of a conceptual model of Citation-Based Similarities from CORES

similarity measure computed using a specific technique. $T_{Citation}$ represents a total number of Citation based similarity measuring techniques used for this computation. We have calculated the $PSim_{Citation}$ for typical similarity measures from Table 6.3, by the following expression, using (6.5).

$$PSim_{Citation} = \frac{0.42 + 0.37 + 0.53}{3} = 0.44 \quad (6.6)$$

6.3.4 Use Case-4: Comprehensive Similarity Measures Computation

This use case describes the computation of research paper similarity measures for those categories/classes which have a disjoint relationship between each other as conceptually modelled in the CORES. The classes are shown in Figure 6.2, which are immediate subclasses of a superclass “Content_Based_Similarity”, these are labeled as: “Vector Space Based”, “Probabilistic”, “Citation-Based”, and “Lexical”.

Typical comprehensive research paper similarity measures for these classes have been computed for a pair of research papers in use cases 1, 2, and 3 respectively. These measures are represented by labels: $PSim_{VSBM}$, $PSim_{Probabilistic}$, and $PSim_{Citation}$. The typical values computed for these measures are 0.49, 0.37, and 0.44. Following equation represents a comprehensive research paper similarity measure $PSim_{Comprehensive}$, computed using the above-described measures in equation (6.7).

$$PSim_{Comprehensive} = \alpha \times PSIM_{VSBM} + \beta \times PSIM_{Probabilistic} + \gamma \times PSIM_{Citation} \quad (6.7)$$

where α , β and γ are weights, which will be assigned values manually according to significance of different research paper similarity measures in literature. Equation (6.7) represents a normalization function which will confirm that the similarity measure comprehensively computed will have a value in the range of 0 to 1. According to Beel et al. [10] Vector Space-based similarity techniques

are most commonly used techniques to find content based similarity measures in the research papers as compared to other content-based similarity measuring techniques, therefore, a value of 0.3 is assigned to α weight which is used with $PSim_{VSBM}$. Citation-based similarity measuring techniques are also considered more reliable for finding similarity measures between the research papers [6] because citation between two research papers depicts a cognition link by author(s) of citing paper. Therefore we will use a value of 0.5 for γ weight which will be used with $PSim_{Citation}$. Probabilistic similarity measures are less commonly used in finding relatedness between research papers as compared to Vector Space based and Citation based techniques; therefore, β which is a weight for these measures is assigned a value of 0.2. By using equation (6.7) we have calculated the typical value for a pair of research papers which were used in use cases 1, 2, and 3 in previous sections by the following expression.

$$PSim_{Comprehensive} = 0.5 \times 0.49 + 0.2 \times 0.37 + 0.3 \times 0.44 = 0.45 \quad (6.8)$$

Therefore, from these use cases, it is concluded that CORES can be useful for computing comprehensive research paper similarity measure. This computation uses relationships between different content based research paper similarity techniques (modeled by CORES) in a comprehensive way.

6.4 Conclusions

We conclude this chapter after presenting use cases for an application of CORES. Knowledge of disjoint and overlapping relationships between different content based similarity measuring techniques and similarity measures from CORES was useful in computing the comprehensive research paper similarity measure. Different weights associated with vector space based, probabilistic, and citation based similarity measures as discussed Equation(6.7) have impact on computation of

comprehensive similarity measure. It is further concluded that CORES is successfully evaluated for adaptability metric (discussed in Section [5.1](#)) for use of comprehensive research paper similarity measure computation.

Chapter 7

Analysis of Comprehensive Research Paper Similarity Measure

In this chapter an experiment is presented, which was used to compute comprehensive research paper similarity measure among research papers from a gold standard data set. This similarity measure computation used the knowledge about relationships between content based similarity measuring techniques (modeled in CORES). We have devised a case study for these techniques that lays foundation for the presented experiment. We have also discussed the experimental setup and data set used in this experiment. A java based application was developed to perform the computation of different research paper similarity measures including comprehensive similarity measure as their weighted sum. The results of experiment were discussed by comparing performance of different similarity measures with user study based similarity measure (considered as a benchmark). Performance analysis of results was done by using Fractional Regression Coefficient (Percentage Difference) coefficient [18], from which it was concluded that comprehensive similarity measure was performing better than other similarity measures in experiment and their combinations.

7.1 Case Study for Experiment

An approach adopted by Shekarpour et al. [102] helped us to devise a case study for CORES, as the authors have verified the application of their ontology CEVO with the help of case studies for text annotation. An experiment using this case study as a foundation is further presented in the coming sections. The two layer model of CORES shown in Figure 4.1 is used to describe the layers of CORES adopted in this case study.

Figure 7.1 signifies a portion of CORES with the knowledge base containing information about instances of the described concepts. The concepts to represent the similarity measuring techniques and the measures used to calculate the pairwise similarity between two research papers are shown in the Figure 7.1. The concept “Vector Space-Based” comprises subclasses: “Cosine”, “Jaccard”, and “Distance Based”. There is an overlapping relationship between these three subclasses. The concept “Distance-Based” has a subclass “Euclidean_Distance”. The concept “Content_Based_Similarity_Technique” contains two sub classes “Vector Space Based” and “Citation Based” classes. These classes obligate a disjoint relationship with each other.

In the layer representing “Pairwise Content based Similarity Measures”, the different content based similarity measures between a pair of research papers are represented. These measures were produced from the content based similarity measuring techniques, which are characterized in the layer named as “Content based Similarity Measuring Techniques”. A concept “Pairwise_Content_Based_Similarity_Measure” from the “Pairwise Content based Similarity Measures” layer represents a super class. There are sub classes named as “PW_Vector_Space_Based_Sim_Measure” and “PW_Citation_Based” for this super class. There is a disjoint relationship between these sub classes. The concept “PW_Vector_Space_Based_Sim_Measure” further contains sub classes “PW_Cosine_Sim_M”, “PW_Jaccard_Sim_M”, and “PW_Euclidean_DSM”. There is an overlapping relationship between these sub classes.

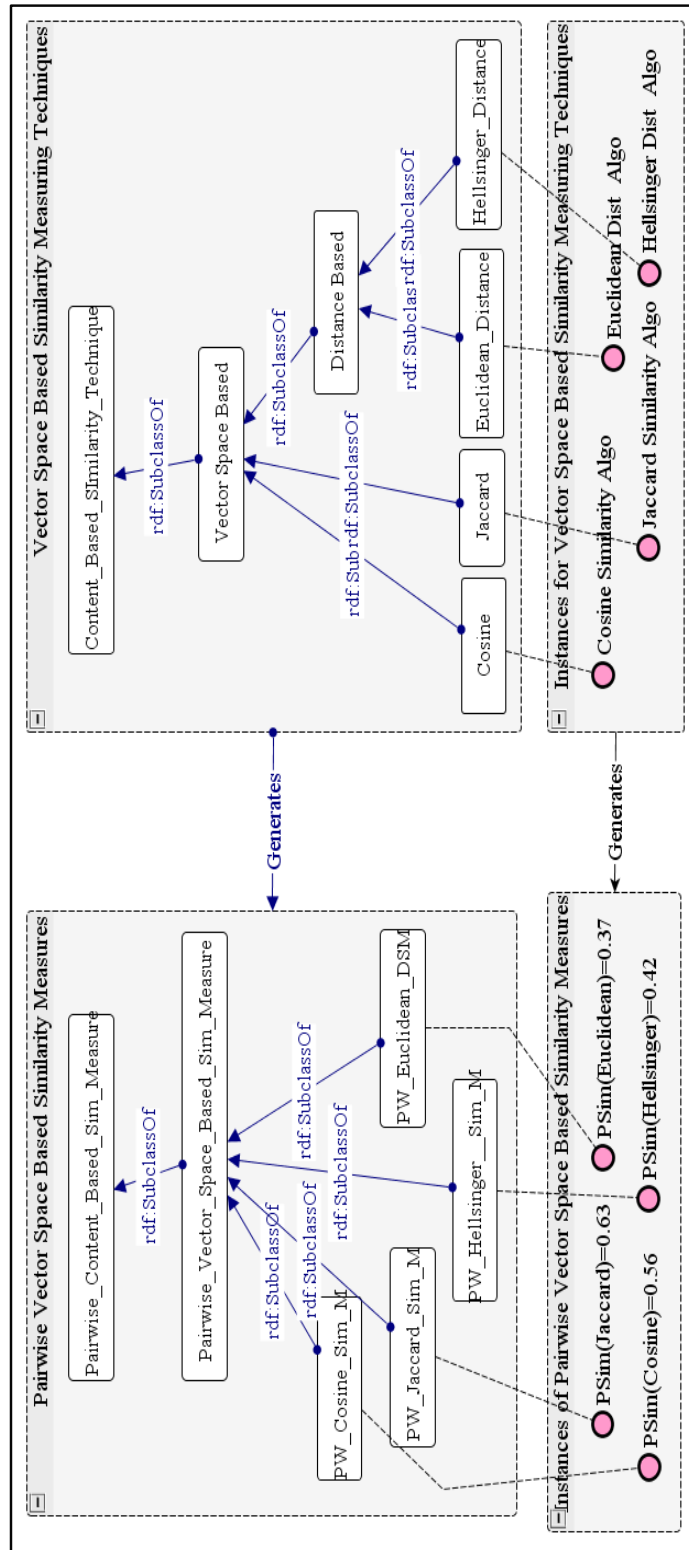


FIGURE 7.1: Concepts and Knowledge based on content based similarities from CORES to be used for Case Study

The knowledge base for the layer “Content Based Similarity Measuring Techniques” denotes different similarity measuring algorithms as its instances. The knowledge base for layer “Pair Wise Content based Similarity Measures” characterizes numeric values for these measures as its instances. These values are used for calculation of comprehensive similarity measures.

Different similarity measures like Cosine, Jaccard, Euclidean, and InText Citation based similarity measures between abstracts of two selected research papers were calculated in this case study. The abstracts of these research papers were harnessed to find Vector Space based similarity measures between them. For this purpose, keywords from these abstract were mined and document vectors were constructed from these keywords. TF/IDF measures for these document vectors were calculated and used for computation of Vector Space based similarity measures.

Using this case study as a starting point, we have performed an experiment on a data set of research papers to compute InText citation based and vector space based similarity measures, their different binary combinations, and comprehensive similarity measure.

7.2 Experimental Setup

For this experiment we have used a core-i5 machine. The speed of its microprocessor is 2.40 GHz and it has four cores with 64 bit bus. This machine has 4GB RAM installed. The platform used for the experiment is Microsoft Windows 10 Education edition with 64 bit version. The experiment uses a Java Virtual Machine (JVM) for this platform. The data set being used in experiment is hosted on Microsoft SQL Server DBMS 2016. Java language was used to develop software for experiment with a library Lucene 4.10.0 (for implementation of different similarity methods). COrES was developed using Protégé 4.3 and was stored as an .owl file which was uploaded on GitHub.

7.3 Data Set Description

We have used a gold standard data set [103] for this experiment. The reasons for using this data set are: it contains InText citation based similarity measure and a user study based similarity measures for pairs of research papers. The data set also covers ranking scores of research papers based on these similarity measures. This data set comprises of 72 query papers and for each query paper, there are reference papers, which are cited by query paper (about 3 to 8). The complete set of query and reference papers were circulated among 124 human raters for user study based ranking. These raters have assigned ranks to reference papers according to their valuation of similarity of reference papers with the query paper. The similarity measures between research papers were assessed by raters on the basis of three categories: High, Medium, and Weak similarity relationships. The ranking is done on the basis of inter-rater agreements. For research papers from the experimental data set, InText citation based similarity and the ranking scores have been calculated in a previous research study [103, 104]. This data set covers 368 pair-wise combinations of research papers for calculating similarity measures among these pairs.

7.4 Application Built for Experiment

A JAVA application using Lucene 4.10.2 for calculation of similarity measures was built for this experiment. The algorithm used by this application to compute comprehensive research paper similarity measure is provided in Appendix B of the Appendix section. For the research papers of this data set, we have further calculated different types of similarity measures between query and reference papers. The types of these similarity measures are Cosine, Jaccard, Euclidean distance based similarity. Different binary combinations of similarity measures were calculated from this data set, listed below:

1. InText Citation based similarity and Cosine similarity.

2. InText Citation based similarity and Jaccard similarity.
3. InText Citation based similarity and Euclidean similarity.
4. InText Citation based similarity and average Vector Space Model based similarity.
5. Jaccard and Cosine similarity.
6. Jaccard and Euclidean similarity.
7. Cosine and Euclidean similarity.
8. Comprehensive Similarity Measure.

In CORES; Cosine, Jaccard, and Euclidean similarity measures are classified under vector space based similarity class. According to the CORES these similarity measures have an overlapping relationship with each other. Therefore, to calculate vector space based similarity measures, we need to take an average of these three measures (by assigning even weight to each of them). Whereas, InText citation and vector space model based similarity measures are classified under the disjoint relationships in CORES, due to the difference of their similarity computation techniques and adopted document models. Therefore, to calculate the similarity measures between the research papers in a comprehensive way using these measures, we need a number of weights. For this purpose, we have defined two weights α and β , to be used with InText citation based and vector space based similarity measures respectively.

7.5 Results and Discussions

7.5.1 Comparison of Different Similarity Measures

The results of comprehensive similarity measures defend our argument of exploiting relationships between different similarity measures (defined in CORES). We

have selected user study based similarity measure from gold standard data set as a benchmark to compare the performance of different similarity measures. Because this measure involves human judgement for finding similar papers so it is a good cognitive measure to be used as a benchmark. Figure 7.2 represents the values of different similarity measures between a query paper and six reference papers from the data set used in this experiment. As shown in the Figure 7.2, the InText citation based similarity measure is close to user study based similarity measure than the vector space based similarity measures. Euclidean distance is also higher as compared to Cosine and Jaccard similarity measures. We have calculated average vector space based similarity measure by taking an average of Cosine, Jaccard, and Euclidean similarity measures.

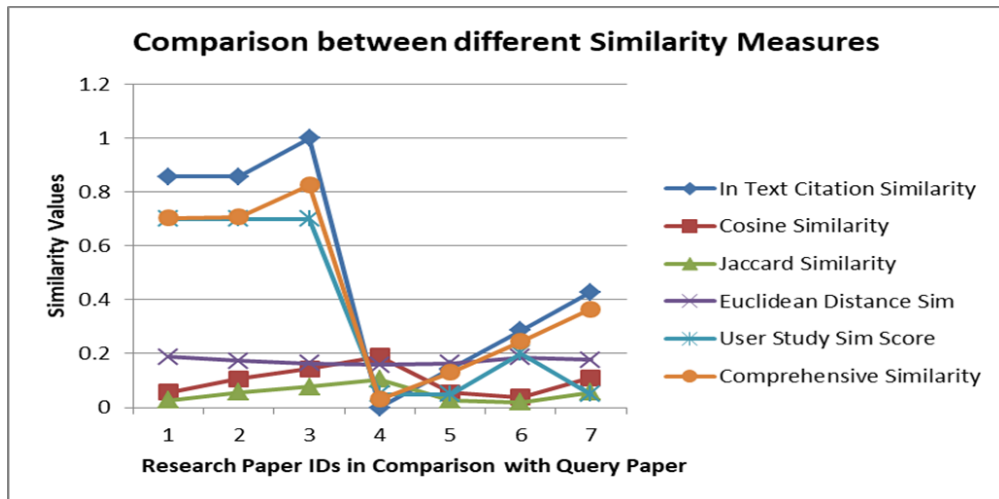


FIGURE 7.2: Different similarity measures compared with User Study based similarity

Figure 7.3 represents a judgment between user study based similarity measures and combinations of InText citation based similarity with different vector space based similarity measures. The comprehensive similarity measure is calculated by combining InText citation based similarity measure with average vector space based similarity measure by using the equation (7.1).

$$\text{ComprehensiveSimilarity} = \alpha \times \text{InTextCitationBasedSimilarity} + \beta \times \text{AverageVectorSpaceBasedSimilarity} \quad (7.1)$$

where α and β are parameters to calculate comprehensive similarity measures. In CORES Ontology, there is a disjoint relationship between InText citation based similarity and vector space based similarity. Therefore, for calculation of comprehensive similarity, different weights are needed with these similarity measures. For this purpose parameters α and β are used. Moreover, Equation (7.2) computes Average Vector Space based Similarity.

$$\begin{aligned} & \text{AverageVectorSpaceBasedSimilarity} && (7.2) \\ = & \frac{\text{CosineSimilarity} + \text{JaccardSimilarity} + \text{EuclideanSimilarity}}{3} \end{aligned}$$

From Figure 7.3, it is clear that comprehensive similarity measure is more close to the user study based similarity measure as compared to InText citation based similarity and vector space based similarity measures.

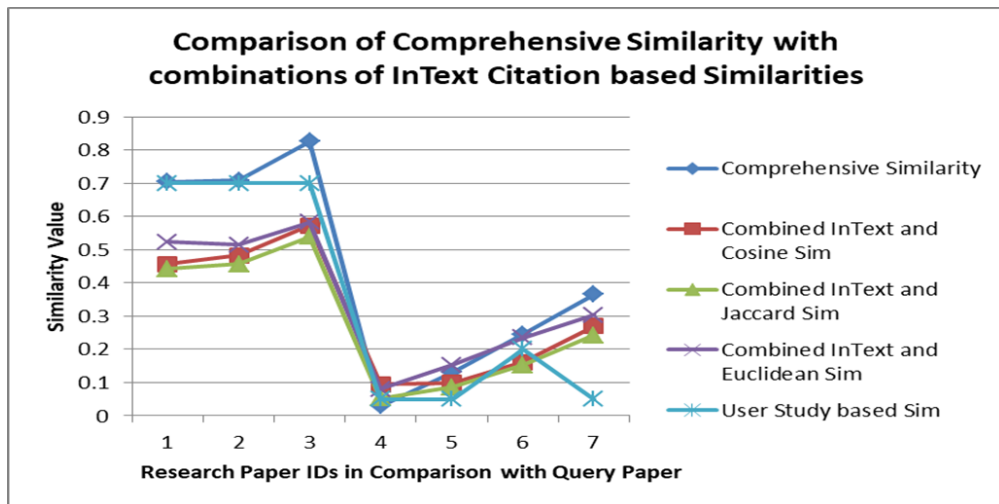


FIGURE 7.3: Comparison of user study based similarity with different combinations involving InText Citation based similarity

Figure 7.3 displays a comparison of user study based similarity measure with a different combination of vector space based similarity measures along with InText citation based similarity measure. Comprehensive similarity measure is more close to the user study based similarity measure than the different combinations of similarity measures calculated in this experiment. The results of comprehensive similarity measure were calculated by using the values of $\alpha = 0.8$ and $\beta = 0.2$. These values were picked because citation based similarity measures are thought

more reliable as compared to vector space based similarity measures in the current state-of-the-art [10]. Figure 7.4 represents the comparison of the user study based

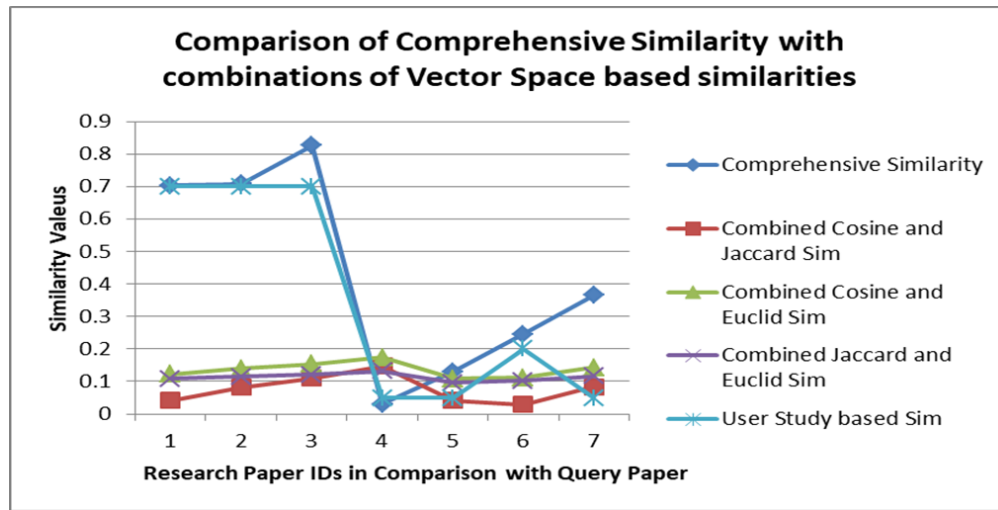


FIGURE 7.4: Comparison of user study based similarity with a combination of vector space based similarity measures

similarity measure with the combinations of vector space based similarity measures. The comprehensive similarity measure is more close to the user study based similarity measure than all other combinations of InText citation based and vector space based similarity measures. Therefore, it was concluded that comprehensive similarity measures performed better than InText citation based and vector space based similarity measures and their combinations used in this experiment, when compared with benchmark of user study based similarity measure. Detailed results for different similarity measures computed in this experiment for a selected set of research papers, and their comparisons are available in Appendix C of the Appendix section.

7.5.2 Performance Analysis Using Percentage Difference Coefficient

We have used Fractional Regression Coefficient (Percentage Difference) [18] for performance analysis of comprehensive similarity measure. User study based similarity measure from gold standard data set has been selected as a benchmark. The

values of Percentage Difference coefficient were computed to compare performance of InText citation based, vector space based similarity measures, their binary combinations, and comprehensive similarity measure with the selected benchmark. A similarity measure with minimum percentage difference value with user study based similarity measure shows a high correlation between the two similarity measures.

Table 7.1 represents the comparison of different similarity measures with user study based similarity using Percentage Difference coefficient. The value of this coefficient is minimum in case of comprehensive similarity measure, which shows that comprehensive similarity measure has better performance than InText citation based and vector space based similarity measures.

TABLE 7.1: Performance comparison of different similarity measures with user study based similarity

Percentage Difference with User Study based Similarity	Difference Value in (%)
InText Citation Similarity	80.780065638499
Cosine Similarity	112.125753460461
Jaccard Similarity	117.405265802158
Euclidean Similarity	98.5663200812536
Comprehensive Similarity	46.9628522892433

Figure 7.5 represents a bar chart based comparison of percentage difference values for InText citation based, vector space based, and comprehensive similarity measure when compared to user study based similarity.

Table 7.2 represents the comparison of different binary combinations of InText citation based and vector space based similarity measures with user study based similarity measure using Percentage Difference coefficient. The value of this coefficient is minimum in case of comprehensive similarity measure when compared to different binary combinations, which shows a better performance for comprehensive similarity measure than the different combinations of InText citation based

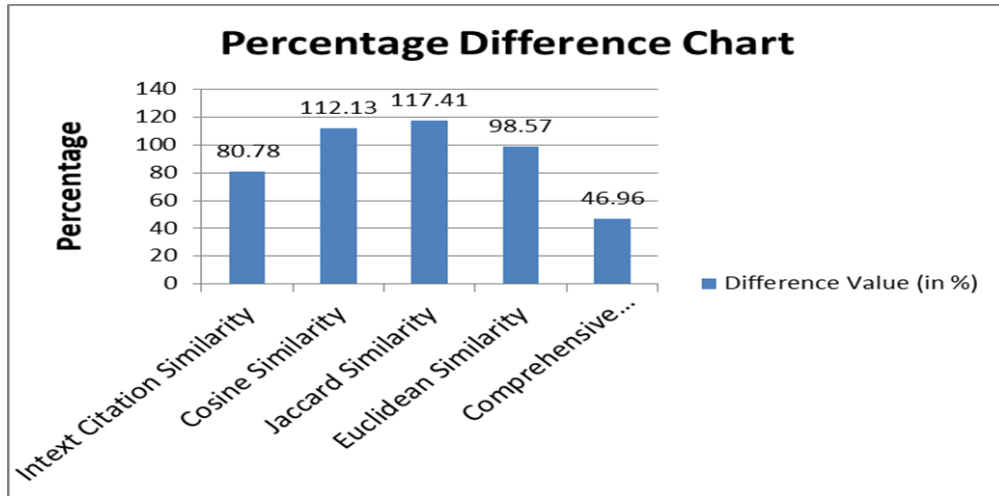


FIGURE 7.5: Comparison of percentage difference values for different similarity measures

and vector space based similarity measures. Figure 7.6 represents a bar chart based comparison of percentage difference values for combinations of InText citation based, vector space based similarity measures, and comprehensive similarity measure with user study based similarity measure.

TABLE 7.2: Performance comparison of combinations of different similarity measures with user study based similarity

Percentage Difference with User Study based Similarity	Difference Value in (%)
InText Citation & Cosine Similarity	54.9835673992705
InText Citation & Jaccard Similarity	51.8750180495239
InText Citation & Euclidean Similarity	54.97238471511
Jaccard & Cosine Similarity	114.323542971291
Cosine & Euclidean Similarity	105.704049423067
Jaccard & Euclidean Similarity	104.141705403067
Comprehensive Similarity	46.9628522892433

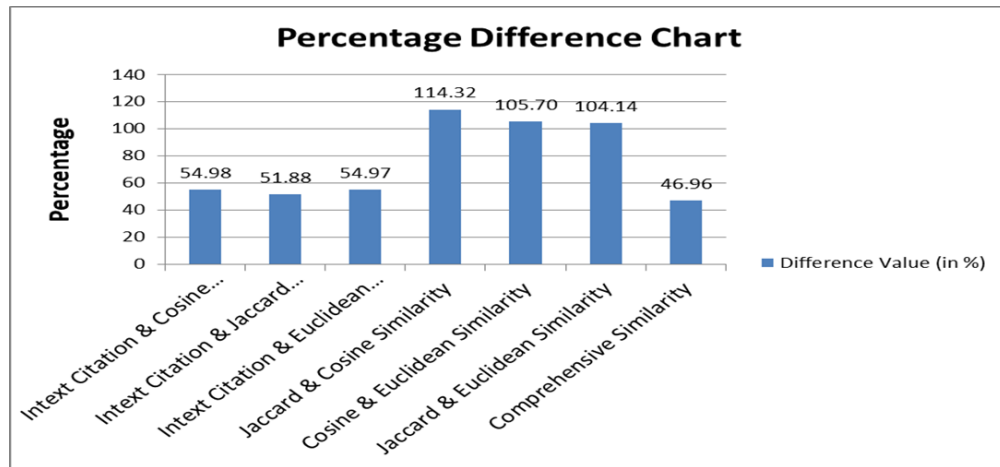


FIGURE 7.6: Comparison of percentage difference values with different combinations of similarity measures

By observing results in Figure 7.5, 7.6, Table 7.1, and 7.2, we conclude that performance of comprehensive similarity measure is better than InText citation based, vector space based similarity measures and their combinations. Therefore knowledge of (overlapping and disjoint) relationships between similarity measuring techniques from CORES have contributed in computation of comprehensive similarity measure which improved the performance by combining different similarity measures in a comprehensive way.

7.6 Conclusions

We have performed an experiment to compute comprehensive research paper similarity measure using knowledge from CORES. For this experiment we have used the user study based similarity measure as a benchmark during the performance comparison of different similarity measures. For performance based comparisons we have used Fractional Regression Coefficient (Percentage Difference) coefficient to compare the different similarity measures and their binary combinations with user study based similarity measure from the gold standard data set. By observing the performance of results it was concluded that comprehensive similarity measure has better performance as compared to other content based similarity measuring techniques and their combinations. Because comprehensive similarity

measure uses knowledge (about relationships among similarity measures) from CORES, therefore CORES was used effectively for the application of computing research paper similarity measures in a comprehensive way. This experiment also evaluates CORES for the adaptability metric (discussed in section [5.1](#) of chapter 5).

Chapter 8

Conclusions and Future Tasks

In this chapter we have concluded the thesis completely by keeping the research contributions of thesis (Section 1.6 of Chapter 1) in mind. The major conclusions from the thesis are listed below.

1. Domain modelling is helpful for machines and algorithms to understand and process knowledge from a specific domain. It was found that domain of research paper similarity is not yet modelled by research community.
2. Different researchers have perceived different classifications of similarity measures but no comprehensive classification for these measures were found.
3. Ontology is a way to model a specific domain; we were unable to find such an ontology which models the domain of research paper similarity measures. A number of ontologies (modelling digital libraries and semantic publishing tasks) were found, but those were not much relevant to this domain.
4. Scope of this domain modelling was restricted to content based similarity measuring techniques due to broad nature of domain. Content based similarity measuring technique was found to be most dominant category from this domain [10].
5. An ontology named CORES was proposed and built to model the domain of research paper similarity measures. This ontology was evaluated by using

state of the art ontology evaluation techniques (tool based evaluation and user study based evaluation) and was found consistent and defined with clarity.

6. An important application of CORES was identified to compute research paper similarity measures in a comprehensive way. This similarity measure was computed by using different similarity measuring techniques comprehensively using the knowledge about relationships between similarity measuring techniques (disjoint and overlapping) from CORES.
7. It was found that knowledge from CORES was helpful in computing comprehensive research paper similarity measure. The performance of comprehensive similarity measure was found to be better than InText citation based and vector space based similarity measuring techniques. It was evident from the results of Fractional Regression Coefficient (Percentage Difference) from an experiment presented in the Chapter 7.

We are looking forward to accomplish following future tasks related to the research conducted in this thesis.

1. The equation (6.7) have used weights/parameters α , β , and γ which were assigned values manually. A future direction of this research work could be devising an algorithm for updating values for these weights, getting knowledge from researchers community using crowd sourcing.
2. The conceptual model of CORES can be enriched so that it can provide more knowledge about different research paper similarity measuring techniques to improve the accuracy of the comprehensive similarity measure computation process. A future research activity will be focusing on exploring relationships between similarity measuring techniques and weighting schemes of research papers.
3. A possibility of importing ontologies related to Digital Libraries and semantic publishing in CORES will be explored for computing comprehensive

research paper similarity measures for annotated documents and research paper repositories.

Bibliography

- [1] Donald Metzler, Susan Dumais, and Christopher Meek. Similarity measures for short segments of text. In *European Conference on Information Retrieval*, pages 16–27. Springer, 2007.
- [2] Wen-Tau Yih and Christopher Meek. Improving similarity measures for short segments of text. In *AAAI*, volume 7, pages 1489–1494, 2007.
- [3] Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 435–440. Association for Computational Linguistics, 2012.
- [4] C Lee Giles, Kurt D Bollacker, and Steve Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98. ACM, 1998.
- [5] Anne-Wil K Harzing and Ron Van der Wal. Google scholar as a new source for citation analysis. *Ethics in science and environmental politics*, 8(1):61–73, 2008.
- [6] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40, 2009.

-
- [7] Kevin W Boyack and Richard Klavans. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the Association for Information Science and Technology*, 61(12):2389–2404, 2010.
- [8] Christian K Shin and David Scott Doermann. Classification of document page images based on visual similarity of layout structures. In *Document Recognition and Retrieval VII*, volume 3967, pages 182–191. International Society for Optics and Photonics, 1999.
- [9] Andrew Nierman and HV Jagadish. Evaluating structural similarity in xml documents. In *webdb*, volume 2, pages 61–66, 2002.
- [10] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitingner. Paper recommender systems: A literature survey. *International Journal on Digital Libraries*, 17(4):305–338, 2016.
- [11] Natalya F Noy, Deborah L McGuinness, et al. *Ontology development 101: A guide to creating your first ontology*, 2001.
- [12] Asuncion Gomez-Perez, Mariano Fernández-López, and Oscar Corcho. *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Science & Business Media, 2006.
- [13] Óscar Corcho, Asunción Gómez-Pérez, Rafael González-Cabero, and M Carmen Suárez-Figueroa. ODEval: a tool for evaluating RDF (S), DAML+OIL, and OWL concept taxonomies. In *Artificial Intelligence Applications and Innovations*, pages 369–382. Springer, 2004.
- [14] Birte Glimm, Ian Horrocks, Boris Motik, Giorgos Stoilos, and Zhe Wang. Hermit: an owl 2 reasoner. *Journal of Automated Reasoning*, 53(3):245–269, 2014.
- [15] Dmitry Tsarkov and Ian Horrocks. FaCT++ description logic reasoner: System description. *Automated reasoning*, pages 292–297, 2006.

- [16] Muhammad Fahad, Muhammad Abdul Qadir, and Muhammad Wajahaat Noshairwan. Ontological errors-inconsistency, incompleteness and redundancy. In *ICEIS (3-2)*, pages 253–285, 2008.
- [17] Joe Raad and Christophe Cruz. A survey on ontology evaluation methods. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 2015.
- [18] Jeffrey O Bennett, William L Briggs, and Anthony Badalamenti. *Using and understanding mathematics: A quantitative reasoning approach*. Pearson Addison Wesley Reading, MA, 2008.
- [19] Boanerges Aleman-Meza, Farshad Hakimpour, I Budak Arpinar, and Amit P Sheth. SwetoDblp ontology of Computer Science publications. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(3):151–155, 2007.
- [20] Silvio Peroni. *Semantic Publishing: issues, solutions and new trends in scholarly publishing within the Semantic Web era*. PhD thesis, alma, 2012.
- [21] Silvio Peroni and David Shotton. FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17:33–43, 2012.
- [22] Michael Ley. DBLP: some lessons learned. *Proceedings of the VLDB Endowment*, 2(2):1493–1500, 2009.
- [23] Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The semantic web revisited. *IEEE intelligent systems*, 21(3):96–101, 2006.
- [24] Dublin Core Metadata Initiative et al. Dublin Core Metadata Element Set, version 1.1. <http://dublincore.org/documents/dcmi-terms/>, 2012.
- [25] International Digital Enterprise Alliance. <http://www.idealliance.org>, 2009.

- [26] Alistair Miles and Sean Bechhofer. SKOS simple knowledge organization system reference. 2009.
- [27] Ye Zhang and Zhan-lin Yu. Research on ontology-based semantic similarity computation. In *Machine Vision and Human-Machine Interface (MVHI), 2010 International Conference on*, pages 472–475. IEEE, 2010.
- [28] Ahmad Fayez S Althobaiti. Comparison of ontology-based semantic-similarity measures in the biomedical text. *Journal of Computer and Communications*, 5(02):17, 2017.
- [29] Vladimir Oleshchuk and Asle Pedersen. Ontology based semantic similarity comparison of documents. In *null*, page 735. IEEE, 2003.
- [30] David Sánchez, Montserrat Batet, David Isern, and Aida Valls. Ontology-based semantic similarity: A new feature-based approach. *Expert systems with applications*, 39(9):7718–7728, 2012.
- [31] Rouzbeh Meymandpour and Joseph G Davis. A semantic similarity measure for linked data: An information content-based approach. *Knowledge-Based Systems*, 109:276–293, 2016.
- [32] Juan J Lastra-Díaz, Ana García-Serrano, Montserrat Batet, Miriam Fernández, and Fernando Chirigati. Hesml: A scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. *Information Systems*, 66:97–118, 2017.
- [33] Sébastien Harispe, David Sánchez, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *Journal of biomedical informatics*, 48:38–53, 2014.
- [34] Mehran Sahami and Timothy D Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International conference on World Wide Web*, pages 377–386. AcM, 2006.

-
- [35] Dan Gusfield. *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge university press, 1997.
- [36] Michael J Wise. YAP3: Improved detection of similarities in computer program and other texts. *ACM SIGCSE Bulletin*, 28(1):130–134, 1996.
- [37] Anna Huang. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56, 2008.
- [38] Svetlozar Todorov Rachev. *Probability metrics and the stability of stochastic models*, volume 269. John Wiley & Son Ltd, 1991.
- [39] Cristiano Nascimento, Alberto HF Laender, Altigran S da Silva, and Marcos André Gonçalves. A source independent framework for research paper recommendation. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 297–306. ACM, 2011.
- [40] Zhi Ping Zhang, Lin Na Li, and Hai Yan Yu. A hybrid document recommender algorithm based on random walk. In *Applied Mechanics and Materials*, volume 336, pages 2270–2276. Trans Tech Publ, 2013.
- [41] Michael D Ekstrand, Praveen Kannan, James A Stemper, John T Butler, Joseph A Konstan, and John T Riedl. Automatically building research reading lists. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 159–166. ACM, 2010.
- [42] Yichen Jiang, Aixia Jia, Yansong Feng, and Dongyan Zhao. Recommending academic papers via users’ reading purposes. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 241–244. ACM, 2012.
- [43] Zan Huang, Wingyan Chung, Thian-Huat Ong, and Hsinchun Chen. A graph-based recommender system for digital library. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 65–73. ACM, 2002.

-
- [44] K Jack. Mendeley: recommendation systems for academic literature. *Presentation at Technical University of Graz (TUG)*, 2012.
- [45] Elena Erosheva, Stephen Fienberg, and John Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5220–5227, 2004.
- [46] Saurabh Kataria, Prasenjit Mitra, and Sumit Bhatia. Utilizing context in generative bayesian models for linked corpus. In *AAAI*, volume 10, page 1, 2010.
- [47] Felice Ferrara, Nirmala Pudota, and Carlo Tasso. A keyphrase-based paper recommender system. In *IRCDL*, pages 14–25. Springer, 2011.
- [48] Kazunari Sugiyama and Min-Yen Kan. Scholarly paper recommendation via user’s recent research interests. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 29–38. ACM, 2010.
- [49] Kannan Chandrasekaran, Susan Gauch, Praveen Lakkaraju, and Hiep Luong. Concept-based document recommendations for citeseer authors. In *Adaptive hypermedia and adaptive web-based systems*, pages 83–92. Springer, 2008.
- [50] Steven Bethard and Dan Jurafsky. Who should i cite: learning literature search models from citation behavior. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 609–618. ACM, 2010.
- [51] Xiaoyu Tang and Qingtian Zeng. Keyword clustering for user interest profiling refinement within paper recommender systems. *Journal of Systems and Software*, 85(1):87–101, 2012.
- [52] Kurt D Bollacker, Steve Lawrence, and C Lee Giles. Citeseer: An autonomous web agent for automatic retrieval and identification of interesting publications. In *Proceedings of the second international conference on Autonomous agents*, pages 116–123. ACM, 1998.

- [53] Jyoti1, Sanjeev Dhawan, and Kulvinder Singh. Comparison of various similarity measure techniques for generating recommendations for e-commerce sites and social websites. *American International Journal of Research in Science, Technology, Engineering & Mathematics*, pages 219–221, 2015.
- [54] Christian Sternitzke and Isumo Bergmann. Similarity measures for document mapping: A comparative study on the level of an individual scientist. *Scientometrics*, 78(1):113–130, 2009.
- [55] Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3):491–504, 2014.
- [56] Hung Chim and Xiaotie Deng. A new suffix tree similarity measure for document clustering. In *Proceedings of the 16th international conference on World Wide Web*, pages 121–130. ACM, 2007.
- [57] Sarah Kohail and Chris Biemann. Matching, reranking and scoring: Learning textual similarity by incorporating dependency graph alignment and coverage features. In *18th International Conference on Computational Linguistics and Intelligent Text Processing. Budapest, Hungary*, 2017.
- [58] Joeran Beel, Stefan Langer, Marcel Genzmehr, and Andreas Nürnberger. Introducing docear’s research paper recommender system. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 459–460. ACM, 2013.
- [59] Christopher D Manning, Hinrich Schütze, et al. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.
- [60] Shen Huang, Gui-Rong Xue, Ben-Yu Zhang, Zheng Chen, Yong Yu, and Wei-Ying Ma. Tssp: A reinforcement algorithm to find related papers. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 117–123. IEEE Computer Society, 2004.

- [61] Shuming Shi, Fei Xing, Mingjie Zhu, Zaiqing Nie, and Ji-Rong Wen. Anchor text extraction for academic search. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 10–18. Association for Computational Linguistics, 2009.
- [62] Marcos Baez, Daniil Mirylenka, and Cristhian Parra. Understanding and supporting search for scholarly knowledge. *Proceeding of the 7th European Computer Science Summit*, pages 1–8, 2011.
- [63] Onur Küçüktunç, Kamer Kaya, Erik Saule, and Ümit V Çatalyürek. Fast recommendation on bibliographic networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 480–487. IEEE, 2012.
- [64] Yicong Liang, Qing Li, and Tiejun Qian. Finding relevant papers based on citation relations. In *International Conference on Web-Age Information Management*, pages 403–414. Springer, 2011.
- [65] Allison Woodruff, Rich Gossweiler, James Pitkow, Ed H Chi, and Stuart K Card. Enhancing a digital book with a reading recommender. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 153–160. ACM, 2000.
- [66] Tadachika Ozono, Shoji Goto, Nobuhiro Fujimaki, and Toramatsu Shintani. P2p based knowledge source discovery on research support system papits. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, pages 49–50. ACM, 2002.
- [67] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Forum*, volume 51, pages 268–276. ACM, 2017.
- [68] Jimmy Lin and W John Wilbur. Pubmed related articles: a probabilistic topic-based model for content similarity. *BMC bioinformatics*, 8(1):423, 2007.

- [69] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 421–430. ACM, 2010.
- [70] Michal Jacovi, Vladimir Soroka, Gail Gilboa-Freedman, Sigalit Ur, Elad Shahr, and Natalia Marmasse. The chasms of cscw: a citation graph analysis of the cscw conference. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 289–298. ACM, 2006.
- [71] Levent Bolelli, Seyda Ertekin, and C Lee Giles. Clustering scientific literature using sparse citation graph analysis. In *PKDD*, volume 6, pages 30–41. Springer, 2006.
- [72] Wangzhong Lu, J Janssen, E Milios, Nathalie Japkowicz, and Yongzheng Zhang. Node similarity in the citation graph. *Knowledge and Information Systems*, 11(1):105–129, 2007.
- [73] Lior Rokach, Prasenjit Mitra, Saurabh Kataria, Wenyi Huang, and Lee Giles. A supervised learning method for context-aware citation recommendation in a large corpus. *INVITED SPEAKER: Analyzing the Performance of Top-K Retrieval Algorithms*, page 1978, 1978.
- [74] David Buttler. A short survey of document structure similarity algorithms. Technical report, Lawrence Livermore National Laboratory (LLNL), Livermore, CA, 2004.
- [75] Zongda Wu, Hui Zhu, Guiling Li, Zongmin Cui, Hui Huang, Jun Li, Enhong Chen, and Guandong Xu. An efficient wikipedia semantic matching approach to text document classification. *Information Sciences*, 393:15–28, 2017.
- [76] Xiaohua Hu, Xiaodan Zhang, Caimei Lu, Eun K Park, and Xiaohua Zhou. Exploiting wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 389–396. ACM, 2009.

- [77] Palakorn Achananuparp, Xiaohua Hu, and Xiajiong Shen. The evaluation of sentence similarity measures. *Data warehousing and knowledge discovery*, pages 305–316, 2008.
- [78] Lan Huang, David Milne, Eibe Frank, and Ian H Witten. Learning a concept-based document similarity measure. *Journal of the Association for Information Science and Technology*, 63(8):1593–1608, 2012.
- [79] LM Vilches-Blázquez, JA Ramos, Francisco J López-Pellicer, Oscar Corcho, and Javier Nogueras-Iso. An approach to comparing different ontologies in the context of hydrographical information. In *Information fusion and geographic information systems*, pages 193–207. Springer, 2009.
- [80] Leo Obrst, Werner Ceusters, Inderjeet Mani, Steve Ray, and Barry Smith. The evaluation of ontologies. In *Semantic web*, pages 139–158. Springer, 2007.
- [81] Thomas R Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5-6):907–928, 1995.
- [82] Aldo Gangemi, Carola Catenacci, Massimiliano Ciaramita, and Jens Lehmann. Ontology evaluation and validation: an integrated formal model for the quality diagnostic task. *On-line: http://www.loa-cnr.it/Files/OntoEval4OntoDev_Final.pdf*, 2005.
- [83] Alexander Ulanov, Georgy Shevlyakov, Nikolay Lyubomishchenko, Pankaj Mehra, and Vladimir Polutin. Monte carlo study of taxonomy evaluation. In *Database and Expert Systems Applications (DEXA), 2010 Workshop on*, pages 164–168. IEEE, 2010.
- [84] Alexander Maedche and Steffen Staab. Measuring similarity between ontologies. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 251–263. Springer, 2002.

- [85] Chintan Patel, Kaustubh Supekar, Yugyung Lee, and EK Park. Ontokhoj: a semantic web portal for ontology searching, ranking and classification. In *Proceedings of the 5th ACM international workshop on Web information and data management*, pages 58–61. ACM, 2003.
- [86] Matthew Jones and Harith Alani. Content-based ontology ranking. 2006.
- [87] Christopher Welty, Ruchi Mahindru, and Jennifer Chu-Carroll. Evaluating ontology cleaning. In *AAAI*, pages 311–316, 2004.
- [88] Peter Haase and York Sure. Usage tracking for ontology evolution. *SEKT deliverable*, 3(1).
- [89] Miriam Fernández, Chwhynny Overbeeke, Marta Sabou, and Enrico Motta. What makes a good ontology? a case-study in fine-grained knowledge reuse. In *Asian Semantic Web Conference*, pages 61–75. Springer, 2009.
- [90] Harith Alani and Christopher Brewster. Metrics for ranking ontologies. 2006.
- [91] Dan Brickley, Ramanathan V Guha, and Brian McBride. Rdf schema 1.1. *W3C recommendation*, 25:2004–2014, 2014.
- [92] Owl semantic web standards. <https://www.w3.org/OWL/>, December 2012.
- [93] Matthew Horridge. CO-ODE Project. <http://owl.cs.manchester.ac.uk/research/co-ode/>, Last Accessed, January, 2015.
- [94] Masahiro Hori, Jérôme Euzenat, and Peter Patel-Schneider. Owl web ontology language xml presentation syntax. November 2009.
- [95] Boris Motik, Peter F Patel-Schneider, Bijan Parsia, Conrad Bock, Achille Fokoue, Peter Haase, Rinke Hoekstra, Ian Horrocks, Alan Ruttenberg, Uli Sattler, et al. OWL 2 web ontology language: Structural specification and functional-style syntax. *W3C recommendation*, 27(65):159, December 2012.
- [96] Matthew Horridge and Peter F Patel-Schneider. OWL 2 web ontology language Manchester syntax. *W3C Working Group Note*, December 2012.

-
- [97] Ian Horrocks, Peter F Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosf, Mike Dean, et al. SWRL: A semantic web rule language combining OWL and RuleML. *W3C Member submission*, 21:79, May 2004.
- [98] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet: A practical owl-dl reasoner. *Web Semantics: science, services and agents on the World Wide Web*, 5(2):51–53, 2007.
- [99] Dan Connolly, Frank Van Harmelen, Ian Horrocks, Deborah L McGuinness, Peter F Patel-Schneider, and Lynn Andrea Stein. Daml+ oil (march 2001) reference description. 2001.
- [100] Alberto Barrón-Cedeno, Paolo Rosso, Eneko Agirre, and Gorka Labaka. Plagiarism detection across distant language pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 37–45. Association for Computational Linguistics, 2010.
- [101] Jean O Lanjouw and Mark Schankerman. Characteristics of patent litigation: a window on competition. *RAND journal of economics*, pages 129–151, 2001.
- [102] Saeedeh Shekarpour, Valerie Shalin, Krishnaprasad Thirunarayan, and Amit P Sheth. Cevo: Comprehensive event ontology enhancing cognitive annotation. *arXiv preprint arXiv:1701.05625*, 2017.
- [103] Abdul Shahid, Muhammad Tanvir Afzal, and Muhammad Abdul Qadir. Lessons learned: The complexity of accurate identification of in-text citations. *Int. Arab J. Inf. Technol.*, 12(5):481–488, 2015.
- [104] Abdul Shahid and Muhammad Tanvir Afzal. Section-wise indexing and retrieval of research articles. *Cluster Computing*, pages 1–12, 2017.

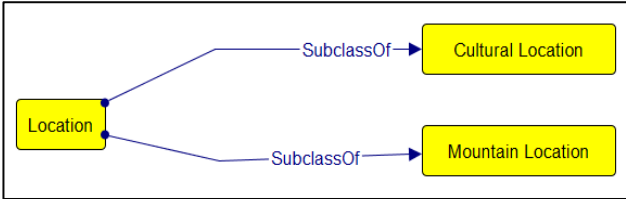
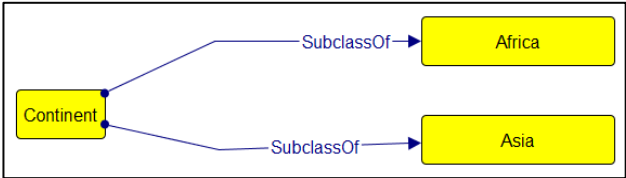
Appendix A

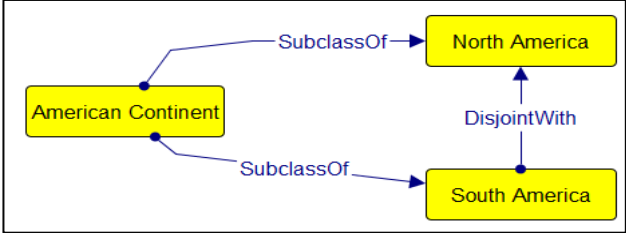
Questionnaire for User-based evaluation of CORES


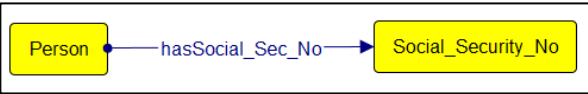
Following Table represents a questionnaire which was given to different experts from the domain of document similarity. This questionnaire contains questions related to three major Ontology error categories identified by Gómez et al. [12]. Each question contains an example scenario, designed to help an expert in understanding the question statement. Some questions need a descriptive and subjective answer, whereas others can be answered just as Yes/No. There are also some questions about errors related to functional and inverse functional properties as reported by Fahad et al. [16]. After getting feedback from experts on these questionnaires, statistical measures regarding results for evaluation of CORES were computed.

A sample questionnaire for user study based evaluation of CORES

Question	Type of Question	Answer
----------	------------------	--------

<p>Incomplete Concept Classification</p> <p>1. Do you think that some content based similarity measure category(s) is/are missing in this ontology?</p> <p>Example Scenario:</p> <p>In following diagram a class “Location” is categorized as sub-classes “Cultural Location” and “Mountain Location”, but other categories of Location (such as “Beach Location” and “Skiing Location”) are missing in this hierarchy.</p>  <pre> graph LR Location[Location] -- SubclassOf --> Cultural[Cultural Location] Location -- SubclassOf --> Mountain[Mountain Location] </pre>	<p>Binary Yes/No</p>	
<p>Incomplete Concept Classification</p> <p>2. If the answer to Question 1 is YES then please list such missing categories!</p>	<p>Subjective</p>	
<p>Partition Error: Disjoint Knowledge Omission</p> <p>Is there any disjoint relationship between concepts in this ontology missing?</p> <p>Example Scenario:</p> <p>For example, “Continent” class is subdivided into subclasses “Africa” and “Asia”. But disjoint constraint between these subclasses is not defined.</p>  <pre> graph LR Continent[Continent] -- SubclassOf --> Africa[Africa] Continent -- SubclassOf --> Asia[Asia] </pre>	<p>Binary Yes/No</p>	
<p>Partition Error: Disjoint Knowledge Omission</p> <p>4. In answer to Question 3 is YES then please provide such concept pairs which have missing disjoint relationships.</p>	<p>Subjective</p>	

<p>Partition Error: Exhaustive Knowledge Omission</p> <p>5. Is a class in this ontology exhaustively decomposed into its subclasses or not?</p> <p>Example Scenario:</p> <p>For example if “American Continent” class is decomposed into “North America” and “South America” with the disjoint relationship between them. This decomposition exhaustively decomposes “American Continent”. But this knowledge is missing from this classification.</p>  <pre> classDiagram class AmericanContinent[American Continent] class NorthAmerica[North America] class SouthAmerica[South America] AmericanContinent -- > NorthAmerica AmericanContinent -- > SouthAmerica NorthAmerica .. SouthAmerica : DisjointWith </pre>	<p>Binary Yes/No</p>	
<p>Partition Error: Exhaustive Knowledge Omission</p> <p>6. If the answer to Question 5 is NO then please point out those concepts for which sub-concepts (subclasses) have not decomposed exhaustively.</p>	<p>Subjective</p>	
<p>Scientific Knowledge Omission Error</p> <p>7. Is there enough description of each concept of this ontology provided, by reading which one can understand that for which purpose concept was defined?</p> <p>Note: Please refer to Document of ontology provided with this Questionnaire.</p>	<p>Binary Yes/No</p>	
<p>Scientific Knowledge Omission Error</p> <p>8. If the answer to Question 7 is NO, then please list those concepts whose description are missing, or does not provide enough information about that concept.</p>	<p>Subjective</p>	

<p>Functional Property Omission for Single-Valued Attributes</p> <p>9. Are functional properties properly defined between concepts of ontology?</p> <p>Example Scenario:</p> <p>In below diagram, “HasBloodGroup” property will have a single value for “Person” class and “Blood Group” class. Such properties are also called functional properties.</p>  <pre> graph LR Person[Person] -- HasBloodGroup --> BloodGroup[Blood Group] </pre>	<p>Binary Yes/No</p>	
<p>Functional Property Omission for Single-Valued Attributes</p> <p>10. If the answer to Question 9 is NO, please list down those concepts for which functional properties are not defined/missing.</p>	<p>Subjective</p>	
<p>Inverse Functional Property Omission</p> <p>11. Are there any inverse functional properties missing between concepts of this ontology?</p> <p>Example Scenario:</p> <p>In below diagram there is an Inverse Functional property “hasSocialSec_No”, according to which value of “Social_Security_No” uniquely determines a “Person”.</p>  <pre> graph LR Person[Person] -- hasSocial_Sec_No --> SocialSecurityNo[Social_Security_No] </pre>	<p>Binary Yes/No</p>	
<p>Inverse Functional Property Omission</p> <p>12. If the answer to Question 11 is YES please identify such missing properties and concept pairs for them.</p>	<p>Subjective</p>	

Appendix B

Algorithm for Computation of Comprehensive Similarity Measure

An algorithm for computation of comprehensive similarity measure is discussed in this section. This algorithm computes pair-wise similarity measure between the research papers. A Java application was built to implement this algorithm using Lucene 4.10.2 library in an experiment performed on a Gold Standard dataset as discussed in Chapter 7. Following are some assumptions used by this algorithm.

Assumptions: P_1, P_2, \dots, P_n represents a set of research papers for which comprehensive similarity measure is required to be calculated. P_i and P_j represents a pair of research papers between which different similarity measures will be computed. Where $i \neq j$ and $i, j \in \{1, 2, \dots, n\}$

Algorithm 1 Computation of comprehensive similarity measure

1: Inputs:

Set of Research Papers P_1, P_2, \dots, P_n

2: Outputs:

Results of Comprehensive Similarity values for pairs of papers

$(P_1, P_2), (P_1, P_3), \dots, (P_1, P_n)$

-
- 3: ComprehensiveSimilarityComputation (P_i, P_j)
 - {
 - 4: CosineSimilarityComputation (P_i, P_j)
 - ▷ compute Cosine similarity between pair of research papers
 - (P_i, P_j)
 - 5: JaccardSimilarityComputation (P_i, P_j)
 - ▷ compute Jaccard similarity between pair of research papers
 - (P_i, P_j)
 - 6: EuclideanSimilarityComputation (P_i, P_j)
 - ▷ compute Euclidean similarity between pair of research
 - papers (P_i, P_j)
 - 7: Average_VSM_SimilarityComputation (P_i, P_j)
 - ▷ compute average vector space based similarity measures
 - between pair of research papers (P_i, P_j)
 - 8: InTextCitationBasedSimilarity(P_i, P_j)
 - ▷ InText Citation based similarity between pair of research
 - papers (P_i, P_j) is already Available in Gold Standard dataset
 - 9: ComprehensiveSimilarity = $\alpha \times$
 InTextCitationBasedSimilarity(P_i, P_j) + $\beta \times$
 Average_VSM_SimilarityComputation(P_i, P_j)
 - ▷ α and β are parameters defined on basis of knowledge from
 - COReS Ontology
 - }
-

Algorithm 2 Driver Application Algorithm

1: Inputs:

 Driver Application P_1, P_2, \dots, P_n

- 2: var i, j ; ▷ i and j are local variables
 - 3: $i = 1$;
 - 4: **for** $j = 2$ to n **do** ComprehensiveSimilarityComputation (P_i, P_j)
 - ▷ Compute pair-wise comprehensive similarity between P_1, P_2, \dots, P_n
 - 5: **end for**
-

Appendix C

Different Similarity Measure values computed for a set of research papers from Gold Standard Dataset

In this section we have represented different similarity measure values which were computed for a set research papers from the Gold Standard Data set selected for experiment as discussed in Chapter 7. Following tables will show these measures computed for a set of research papers which were used to compute similarity measures and whose ranking scores with Spearman Correlation Coefficients were discussed in Chapter 7. The value of each measure is between 0 and 1.

TABLE C.1: Comparison of different similarity measures with Comprehensive Similarity measure

Query Paper	Papers in Comparison	InText Citation Similarity	Cosine Similarity	Jaccard Similarity	Euclidean Distance Similarity	Average Vector Space Sim	Comprehensive Similarity
367054	86	0.8571429	0.05575217	0.0278044	0.18817051	0.09057569	0.703829424
367054	87	0.8571429	0.10697505	0.055885	0.17227435	0.11171147	0.708056579
367054	88	1	0.14386244	0.0771009	0.16429009	0.12841779	0.825683559
367054	2665	0	0.18990754	0.1044344	0.159174	0.151172	0.030234399
367054	2666	0.1428571	0.05405642	0.0277786	0.1634526	0.08176254	0.130638222
367054	2667	0.2857143	0.03800743	0.0189394	0.18583146	0.0809261	0.244756648
367054	2668	0.4285714	0.11006522	0.0571493	0.17653035	0.11458161	0.365773465

TABLE C.2: Combinations of InText Citation based similarity measure with Vector Space based similarity measures.

Query Paper	Papers in Comparison	Combined InText and Cosine Sim	Combined InText and Jaccard Sim	Combined InText and Euclidean Sim	Combined InText and VSM Sim
367054	86	0.4564475	0.44247363	0.5226567	0.47385927
367054	87	0.482059	0.45651393	0.5147086	0.48442716
367054	88	0.5719312	0.53855043	0.582145	0.5642089
367054	2665	0.0949538	0.05221722	0.079587	0.075586
367054	2666	0.0984568	0.08531787	0.1531549	0.11230984
367054	2667	0.1618609	0.15232684	0.2357729	0.18332019
367054	2668	0.2693183	0.24286034	0.3025509	0.27157652

TABLE C.3: Combinations of different Vector Space based Similarity Measures

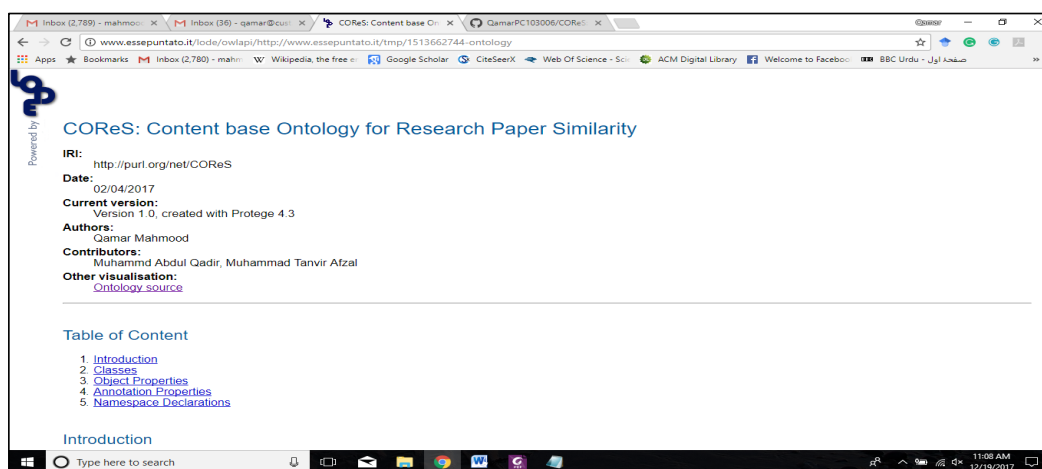
Query Paper	Papers in Comparison	Combined Cosine and Jaccard Sim	Combined Cosine and Euclid Sim	Combined Jaccard and Euclid Sim
367054	86	0.0417783	0.12196134	0.1079875
367054	87	0.08143	0.1396247	0.1140797
367054	88	0.1104816	0.15407626	0.1206955
367054	2665	0.147171	0.17454077	0.1318042
367054	2666	0.0409175	0.10875451	0.0956156
367054	2667	0.0284734	0.11191945	0.1023854
367054	2668	0.0836072	0.14329778	0.1168398

Appendix D

Documentation of CORES

In this section CORES Ontology's documentation is provided in terms of OWL syntax. This document contains definition of classes, properties, and literals. There are two types of relationships between classes which are provided in this definition: Disjoint and Overlapping. These relationships are used for computation of Comprehensive similarity in this thesis. In this annexure only some portion of documentation is presented. Complete document is available as a PDF document at the following link. The name of document is "CORES_Content base Ontology for Research Paper Similarity.pdf".

<https://github.com/QamarPC103006/COREs/>



The screenshot shows a web browser window with the following content:

- Page Title:** CORES: Content base Ontology
- URL:** www.essepuntato.it/ode/owlapi/http://www.essepuntato.it/tmp/1513662744-ontology
- Navigation:** 4 Annotation Properties, 5 Namespace Declarations
- Introduction:**

CORES stands for Content based Ontology for Research paper Similarity. This Ontology provides classifications for content based similarity techniques. There are two layers in the CORES.

One of the layers represent a a class hierarchy for different content based similarity techniques.

Second layer represents the class hierarchy of pair-wise content based similarity measures computed using techniques of first layer.
- Classes:**

average k l divergence, bibliographic coupling sim, citation based similarity, citation context based sim, citation count based sim, citation graph based sim, cocitation analysis sim, content based similarity, cosine similarity, dice coefficient, direct citation count sim, dissimilarity, distance based similarity, edit distance, euclidean similarity, hellsinger similarity, hybrid content based similarity, information radius, jaccard similarity, k l divergence, l cosine similarity, latent dirichel allocation, lexical similarity, manhattan norm, matching coefficient, non content based similarity, normalized pointwise mutual info, overlap coefficient, p w average k l divergence sim measure, p w bibliographic coupling based sim measure, p w citation based sim measure, p w citation context based sim measure, p w citation count based sim measure, p w citation graph based sim measure, p w cocitation analysis based sim measure, p w cosine sim measure, p w dice coefficient s im measure, p w direct citation count based sim measure, p w edit distance sim measure, p w information radius based sim measure, p w jaccard sim measure, p w l cosine sim measure, p w latent dirichel allocation sim measure, p w lexical based sim measure, p w manhattan norm sim measure, p w matching coefficient sim measure, p w normalized p m i sim measure, p w overlapping coefficient sim measure, p w pearson correlation coefficient sim measure, p w pointwise mutual information sim measure, p w probabilistic sim measure, p w structural sim measure, p w free edit distance sim measure, p w vector space based sim measure, p w visual sim measure, p w x m l based sim measure, pair wise content based similarity measures, pearson corelation coefficient similarity, pointwise mutual information, probabilistic similarity, relatedness, similarity, structural similarity, tree edit distance, vector space based similarity, x m l based similarity
- Footer:** average k l divergence^C, [back to ToC](#) or [Class ToC](#)