



# Identification of Important Citations by Exploiting Research Articles' Metadata



by

Faiza Qayyum

MCS143010

A thesis submitted to the  
Department of Computer Science  
In partial fulfillment of the requirements for the degree of  
MASTER OF SCIENCE IN COMPUTER SCIENCE

Faculty of Computing  
Capital University of Science and Technology  
Islamabad  
April, 2017

Copyright ©2017 by CUST Student

All rights reserved. Reproduction in whole or in part in any form requires the prior written permission of Faiza Qayyum (MCS143010) or designated representative.



**C.U.S.T.**

**CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY  
ISLAMABAD**

**CERTIFICATE OF APPROVAL**

**Identification of Important Citations by Exploiting Research Articles'  
Metadata**

by

Faiza Qayyum

MCS143010

**THESIS EXAMINING COMMITTEE**

<b>S No</b>	<b>Examiner</b>	<b>Name</b>	<b>Organization</b>
(a)	External Examiner	Dr. Zahid Halim	GIKI, KPK
(b)	Internal Examiner	Dr. Nayyer Masood	CUST, Islamabad
(c)	Supervisor	Dr. Muhammad Tanvir Afzal	CUST, Islamabad

---

Dr. Muhammad Tanvir Afzal

**Thesis Supervisor**

April, 2017

---

Dr. Nayyer Masood

Head

Department of Computer Science

Dated : April, 2017

---

Dr. Muhammad Abdul Qadir

Dean

Faculty of Computing

Dated : April, 2017

## **CERTIFICATE**

This is to certify that **Mis Faiza Qayyum (MCS143010)** has incorporated all the observations, suggestions and comments by external examiner as well as internal examiner and thesis supervisor. The title of her thesis is: **Identification of Important Citations by Exploiting Research Articles' Metadata**

Forwarded for necessary action

---

Dr. Muhammad Tanvir Afzal  
(Thesis Supervisor)

## **ACKNOWLEDGEMENT**

I would like to express my gratitude and thanks to ALLAH (S.W.T) for providing me abilities to accomplish this research work. Secondly, I would like to express my sincerest thanks to my supervisor Dr Muhammad Tanvir Afzal for his guidance and encouragement. He has taught me, both consciously and unconsciously, how good experimental work is carried out. Sir, you will always be remembered as an inspirational teacher. Last but not the least; I would like to thank my mother and my husband for their support, encouragement and prayers.

# **DECLARATION**

It is declared that this is an original piece of my own work, except where otherwise acknowledged in text and references. This work has not been submitted in any form for another degree or diploma at any university or other institution for tertiary education and shall not be submitted by me in future for obtaining any degree from this or any other University or Institution.

Faiza Qayyum

MCS143010

April, 2017

## ABSTRACT

Citations play pivotal role in indication of various aspects in scientific literature. The quantitative citation analysis approach has been used over the decades to measure the impact factor of journal, rank the researchers and institutions, making of awards and Nobel prizes policies, allocating research grants, discovering evolving research topics etc. In citation analysis community, researchers have doubted the pure quantities citation analysis approach. They argued that all citations are not of equal importance and the reason of citation should be considered while counting it. Researchers have identified different reasons of citations; some are used to provide background knowledge, to critic the existing work, while some takes an idea from existing schemes or uses the existing work etc. Different approaches have been proposed to classify these reasons automatically. In the recent past, researchers have focused to divide the citation reasons into two categories: (1) Important and (2) Non-important rather than classifying each reason individually. Important citations are those which use or extend the existing work and non-important citations are those which are used just to provide background knowledge. The identification of important and non-important citations can help in quantitative citation analysis approaches via counting only those citations which are important.

We have comprehensively studied more than 40 research articles on this topic and identified research gap. In citation classification community, researches have proposed different techniques relying on the content of the articles. In case of exploiting the content of research articles there should be an open access to articles to have their content. But the content is not freely available most of the time; various journal publishers do not provide open access to their articles. In such scenarios, there is a need of some alternative way to classify citations. To address this issue, we have proposed an approach to classify citations into two categories (1) Important and (2) Non-important by using freely available metadata such as titles, authors, keywords, references etc. We have proposed different formulas to obtain the ratio of similarity between metadata of paper-citation pairs. The score against each formula is calculated and assigned as a feature for supervised machine learning for a binary classification. The classification is performed by using state-of-the-art classifiers which are being used in such research works like: SVM, KLR and Random Forest classifier. Two benchmark datasets have been used for experiments: One of them is taken from recent published paper in Association for the Advancement of Artificial Intelligence (the “AAAI”) and another one is collected from Capital University of Science and Technology (the “CUST”) Computer Science Faculty members. We have compared our results with the content based approach and our system achieved improved precision of 0.73.



## Table of Contents

Chapter 1 .....	14
INTRODUCTION .....	14
1.1. BACKGROUND .....	14
1.2 PROBLEM STATEMENT .....	16
1.3 PURPOSE .....	16
1.4 SCOPE .....	16
1.5 APPLICATIONS OF PROPOSED SOLUTION .....	16
1.6 DEFINITIONS, ACRONYMS, AND ABBREVIATIONS .....	17
Chapter 2.....	18
LITERATURE REVIEW .....	18
2.1 IN-TEXT CITATION FREQUENCY .....	20
2.2 CUE WORDS/PHRASES .....	20
2.3 CRITICAL ANALYSIS .....	28
2.3.1 In-text Citation Limitation .....	30
a. High Frequency but Low Relevance .....	30
b. Low Frequency, High Relevance .....	31
Chapter 3.....	34
PROPOSED METHODOLOGY .....	34
3.1 BENCHMARK DATASETS.....	36
3.1.1 Dataset1.....	36
3.1.2 Dataset2.....	37
3.2 METADATA EXTRACTION.....	37
3.3 TITLES EXTRACTION FROM REFERENCES .....	38
3.4 PRE-PROCESSING .....	39
3.4.1 Stop Words Removal .....	40
3.4.2 Stemming .....	40
3.5 TECHNIQUES .....	40
3.5.1 N-Gram Technique .....	40
3.5.2 Synonyms.....	41
3.5.3 Growbag.....	42
3.6 SCORE CALCULATION FOR SUPERVISED LEARNING .....	43

3.6.1 Title Similarity Score .....	44
3.6.2 Author Overlap Score .....	44
3.6.3 Abstract Similarity .....	45
3.6.4 Keywords Similarity .....	45
3.6.5 Categories Similarity.....	46
3.6.6 Bibliographically Coupled References .....	46
3.7 METDATA PARAMETERS COMBINATIONS .....	46
3.7.1 Single Metadata Parameters.....	47
3.7.2 Double Metadata Parameters .....	47
3.7.3 Triple metadata Parameters.....	47
3.7.4 Quadruple Metadata Parameters .....	48
3.7.5 Quintuple Metadata Parameters .....	49
3.7.6 Hextuple Metadata Parameters Combinations .....	50
3.8 CLASSIFIERS .....	50
3.9 EVALUATION AND COMPARISONS.....	51
Chapter 4.....	52
EXPERIMENTS AND RESULTS .....	52
4.1 DATASET COLLECTION .....	52
4.2 METADATA EXTRACTION.....	53
4.3 PRE-PROCESSING .....	54
4.4 SYNONYMS AND GROWBAG MATCHING .....	54
4.5 TITLES EXTRACTION FROM REFERENCES .....	55
4.6 SCORE CALCULATION .....	55
4.8 EVALUATION.....	58
4.8.1 Single Metadata Parameters.....	58
4.8.2 Double Metadata Parameters .....	59
4.8.3 Triple Metadata Parameters .....	61
4.8.4 Quadruple Metadata Parameters .....	62
4.8.5 Quintuple Metadata Parameters .....	64
4.9 COMPARISONS .....	65
Chapter 5.....	68
CONCLUSION AND FUTURE WORK .....	68

5.1 CONCLUSION ..... 68  
5.2 FUTURE WORK..... 70

## LIST OF FIGURES

Figure 3-1 Context Diagram of Proposed System .....	35
Figure 3-2 Benchmark dataset1 .....	36
Figure 3-3 Extracted Metadata Parameters of <i>d1</i> .....	38
Figure 3-4 Extracted Metadata Parameters of <i>d2</i> .....	38
Figure 3-5 Difference between Same Reference Patterns .....	39
Figure 3-6 Heuristic Approach to Extract Titles from References .....	39
Figure 3-7 Double Metadata Parameters Combinations .....	47
Figure 3-8 Triple Metadata Parameters Combinations .....	48
Figure 3-9 Quadruple Metadata Parameters Combinations .....	49
Figure 3-10 Quintuple Metadata Parameters Combinations .....	50
Figure 3-11 Hextuple Metadata Parameters Combinations .....	50
Figure 4-1 File Loading in WEKA .....	56
Figure 4-2 Balanced Classes using SMOTE Filter .....	56
Figure 4-3 Randomizing Classes .....	57
Figure 4-4 Hybrid Classification Using Random Forest classifier for <i>d2</i> .....	57
Figure 4-5 PRF Bar Chart for <i>d1</i> Single Parameters .....	59
Figure 4-6 PRF Bar Chart for <i>d2</i> Single Parameters .....	59
Figure 4-7 PRF Bar Chart for <i>d1</i> Double Parameters .....	60
Figure 4-8 PRF Bar Chart for <i>d2</i> Double Parameters .....	61
Figure 4-9 PRF Bar Chart for <i>d1</i> Triple Parameters .....	62
Figure 4-10 PRF Bar Chart for <i>d2</i> Triple Parameters .....	62
Figure 4-11 PRF Bar Chart for <i>d1</i> Quadruple Parameters .....	63
Figure 4-12 PRF Bar Chart for <i>d2</i> Quadruple Parameters .....	64
Figure 4-13 PRF Bar Chart for <i>d2</i> Quintuple Parameters.....	64
Figure 4-14 Comparisons between Top Scored Features .....	66
Figure 4-15 Comparisons between Author's Overlap Score .....	66
Figure 4-16 Comparisons between Results .....	67

## LIST OF TABLES

Table 2-1 Critical Analysis of State of the art approaches .....	28
Table 2-2 The paper X cites 10 times paper Y .....	31
Table 2-3 The Paper A cites paper B only one time .....	33
Table 3-1 Unigram Terms .....	41
Table 3-2 Bigram terms .....	41
Table 3-3 Trigram terms .....	41
Table 3-4 Word and its synonyms .....	42
Table 3-5 Semantically Related Terms .....	43
Table 4-1 Successful Extraction of Articles in <i>d1</i> .....	52
Table 4-2 Successful Extraction of Articles in <i>d2</i> .....	53
Table 4-3 Successful Extraction of Metadata Parameters in <i>d1</i> .....	53
Table 4-4 Successful Extraction of Metadata Parameters in <i>d2</i> .....	54

## Chapter 1

### INTRODUCTION

*"If I have seen further than others, it is by standing upon the shoulders of giants"*

*- Issac Newton*

#### 1.1. BACKGROUND

Researchers always conduct research by relying on the legendary work of their eminent predecessors in the field. The statement is justified further by Ziman (Ziman, 1968), indicating that *"a scientific paper does not stand alone; it is embedded in the literature of the subject"*. A reference is the acknowledgement that one document gives to another and citation is the acknowledgement that one document receives from another (Narin, 1976). In the previous century, Ziman (Ziman, 1968) narrated the significance of analyzing citations for various research studies. He narrated that high frequency of citation count determines the significance and popularity of the work. In citation analysis based approaches different authors have correlated citation count with other achievements of researchers, such as: (1) Awards and Nobel Prizes (Inhaber & Przednowek, 1976), (2) Allocation of Research Funds (3) Institutional Ranking (Anderson, Narin, & McAllister, 1978) and (4) Peer Judgments (Smith & Eysenck, 2002) . The analysis of citations has not even subsided in the present century. In a report, Wilsdon et al., (Wilsdon, et al., 2015), examined the role of citations to assess the quality of a research. The most recent study published by Benedictus et al., (Benedictus, Miedema, & Ferguson, 2016), analyzed the role of citation quantity of in measuring the excellence of any individual.

Now the question arises that why do researchers cite a particular work? The one of the founder of bibliometrics, Garfield (Garfield, 1965) discovered 15 reasons of citations, some of them are: (1) providing background knowledge (2) criticizing the work (3) acknowledging the work (4) disclaiming others works as their own work etc. After this study, various authors discovered more facts behind citing a particular article. The identification of these reasons assisted the researchers to critically scrutinize the quantitative citation (citation count) approach.

In 1968, Ziman (Ziman, 1968), have criticized the usage of pure quantitative citation analysis (citation count), they argued that many citations are given where author criticizes cited work and the citations received due to criticism should not be given prime importance (Bonzi, 1982). In

1975, the study of Moravcsik & Murugesan (Moravcsik & Murugesan, 1975), revealed that 40% of the citations are those which are received due to providing background knowledge or general acknowledgement; this increased the doubts on citation count approach. Continuing the branch towards critical analysis of citation count, Teufel et al., (Teufel, Siddharthan, & Tidhar, 2006), argued that all citations are not of equal importance and the reason of citation should be considered while giving importance to it. Benedictus et al., (Benedictus, Miedema, & Ferguson, 2016), argued that quantity prevails quality, when citation count is considered to measure excellence of any individual.

A lot of researchers discovered different reasons of citations but the question arises that how to automatically differentiate between citations? The old citations annotation approaches work manually by interviewing the citer, sometime after publication of article, to recall why he cited the work (Brooks, 1985); or interview the scholars at the time of writing the article that why they are citing the particular work (Case & Higgins, 2000). In 1979, Finney (Finney, 1979), was the first who suggested an idea in her master's thesis that citation classification can be done automatically. Various researchers adopted her idea to classify citations. In 2000, the first automated technique for citation classification was proposed by Garzone & Mercer (Garzone & Mercer, 2000). They categorized citations into 35 categories and built 195 lexical matching rules. This system takes article as input and then produces set of citations along with the corresponding citation category. However in literature, their work has been criticized due to proposing large number of categories that can conflict each other (Radoulov, 2008).

Recently, in citation classification community, researchers have focused on categorization of only those reasons into different categories that can assist the reliability of citation count approach. For this purpose, the first approach was proposed by Valenzuela et al., (Valenzuela, Ha, & Etzioni, 2015), in which they classified citations into two reasons: (1) Important and (2) Non-important. Important citations are those which adopt an idea from cited paper or have done a similar work to the cited paper. The non-important citations are those which are used just to provide some theory or background knowledge. They proposed twelve different features relying on the content of the articles. Their dataset is based on 456 annotated paper-citation pairs. In this thesis we will use the same dataset by proposing different features.

## 1.2 PROBLEM STATEMENT

The existing approaches that address the issue of classifying citations are content dependent and most of the time content is not freely available. Major journal publishers like, ACM, Springer, IEEE, Elsevier etc. do not provide open access to their articles. On the other hand various kinds of useful metadata such as titles, authors, keywords etc. are freely available. This has led us to explore the answers of these two questions

- Whether metadata of citations hold the potential in identifying important citations?
- Which metadata parameters or combinations of metadata parameters could achieve the best accuracy?

## 1.3 PURPOSE

The purpose of this thesis is the identification of important and non-important citations by exploiting freely available metadata of citations and references of the source paper. The description of important and non-important citation is described below:

- **Important:** The citations which are using or extending the cited work.
- **Non-important:** The citations done just to provide background knowledge.

## 1.4 SCOPE

The scope of thesis is exploitation of paper-citations/references pairs to determine whether the citation is important or non-important citation of the source paper. The results of this study will be immensely valuable in citation count approaches via counting only those citations which are actually important. It will assist the researchers to have important research articles for their literature survey. Moreover, the authors having relevant interests and current trends in particular areas, can also be discovered

## 1.5 APPLICATIONS OF PROPOSED SOLUTION

This research can assist in various fields such as:

- Authors Ranking
- Impact factor calculation
- Bibliometric studies



## **1.6 DEFINITIONS, ACRONYMS, AND ABBREVIATIONS**

- Association Of Computational Linguistics (ACL)
- Association for the Advancement of Artificial Intelligence (AAAI)
- Capital University of Science and Technology (CUST)
- Support Vector Machine (SVM)
- Random Forest (RF)
- Kernel Logistic Regression (KLR)
- Precision Recall F-measure (PRF)
- Waikato Environment for Knowledge Analysis (WEKA)
- Synthetic Minority Over-sampling Technique (SMOTE)

## Chapter 2

### LITERATURE REVIEW

In scientific literature, citations play paramount role in indication of various factors such as, institutional ranking, peer judgments, authors ranking, impact factor of journal, research grants etc. Generally, citation delineates a relationship between a part or the whole of the cited document and a part or the whole of the citing document. Citation analysis relates to the area of bibliometrics wherein analyses of such relationships are scrutinized (Smith L. C., 1981).

The notion of harnessing citation count was pioneered by Garfield (Garfield, Sher, & Torpie, 1964). In 1964, Garfield et al., (Garfield, Sher, & Torpie, 1964), revealed existence of positive correlation between the Nobel Prize winner authors and the citation count of their articles. Subsequently, different researchers correlated the citation count with the other achievements of researchers such as: (1) Awards and Nobel Prize (Inhaber & Przednowek, 1976) (2) Allocation of research funds (Inhaber & Przednowek, 1976), (3) Global ranking (Anderson, Narin, & McAllister, 1978) and (4) Peer judgments (Smith & Eysenck, 2002). Numerous analyses have been performed with the help of citation analysis. However, the questions pertaining to the purpose of citation remained unanswered. Is it done to appreciate the cited work or to critique the cited work? Such questions emerge as a natural corollary, when one ponders about citation reasons. The discovery of these reasons was initiated by Garfield in 1965 (Garfield, 1965), in which he discovered 15 reasons of citations from which some of them are 1) paying homage to pioneers 2) Giving credit to related work 3) criticizing the work etc.

Until now, the article (Garfield, 1965), “Can citation indexing be automated” has received 296 citations in which some authors have analyzed the aforementioned reasons in depth and further classified these into different other reasons. In 1977, Spiegel-Rosing (Spiegel-Rusing, 1977), discovered thirteen new reasons of citations. The identification of these reasons diverted the attention of researchers towards the reliability of quantitative citation analysis. The researchers started to critically analyze citation count and stated that the reason of citations must be considered to assign weight to the particular citation. In 1968, Ziman (Ziman, 1968), criticized the usage of pure quantitative citation analysis (citation count), he argued that many citations are received where author criticizes the cited work and the citations received as a result of criticism

should not be given significance. In 1975, Moravcsik and Murugesan (Moravcsik & Murugesan, 1975), revealed that 40% of the citations are those which are received due to providing background knowledge or general acknowledgement; this increases the doubt further. In 1979, Garfield and Merton (Garfield & Merton, 1979), critically reviewed the citation count based approaches and concluded that a high citation count could be received by generating low quality work that has received a lot of critiques. The negative citations should not be considered while counting citations for honoring any individual (Bonzi & Snyder, 1991). In 2006, Teufel et al (Teufel, Siddharthan, & Tidhar, 2006), argued that all citations are not of equal importance and while counting citations of an article all citations should not be treated equally. In a report Wilsdon et al., (Wilsdon, et al., 2015), examines the role of citations to assess the quality of research, the results showed that sometime the critiques towards low quality work are large in number that increases the frequency of its citation and it is considered as high quality work because of having high citation count. Analysis of such sort of findings published in a famous journal *nature* examines the role of citation count in measuring the excellence of any individual and concludes that quantity prevails over quality, when pure quantitative citation analysis is performed. Moreover pure quantitative citation analysis shouldn't be utilized and quality should be given more prominence (Benedictus, Miedema, & Ferguson, 2016).

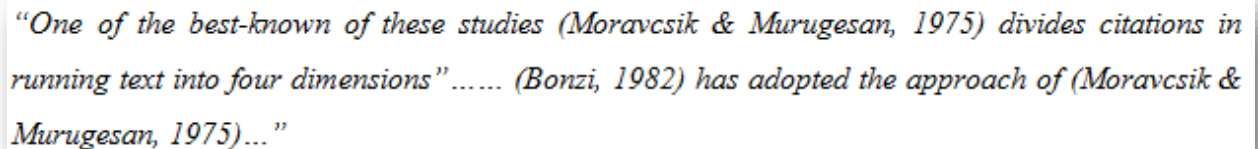
Now the question crops up that how citations classification can be done automatically? The reasons of citation by Garfield (Garfield, 1965), inspired the researchers to discover various other aspects of citing a particular work, but there wasn't any way to classify citations automatically. During that time, citations were manually classified into different reasons by interviewing the citer, sometime after publication of article, to Recall why he cited the work (Brooks, 1985); or interview the scholars at the time of writing the article that why they are citing the particular work (Case & Higgins, 2000).

In 1979, Finney (Finney, 1979), developed an idea in her master's thesis that citation classification can be done automatically. She designed the system in which she associated cue words with citation function and used citation location in the classification algorithm. The critical analysis of this domain revealed that in the year 2000, the first fully automated citation classification technique is proposed (Garzone & Mercer, 2000). After this, various researchers proposed automatic citation classification schemes by using different features. The two major citation classification features are based on:

- 1) In-text citation frequency
- 2) Cue words

## 2.1 IN-TEXT CITATION FREQUENCY

The in-text citation frequency means the count of all citations which appear in body of the paper. The example in figure 2-1 can demonstrate the concept in a better way.



*“One of the best-known of these studies (Moravcsik & Murugesan, 1975) divides citations in running text into four dimensions” ..... (Bonzi, 1982) has adopted the approach of (Moravcsik & Murugesan, 1975)...”*

**Figure 2-1 In-text Citation Example**

In the figure 2-1, the “(Moravcsik and Murugesan, 1975)” appeared twice so its in-text citation count will be counted by counting how many times it will be found in body of the paper.

In 2011, Shahid et al., (Shahid, Afzal, & Qadir, 2011), claimed that if in-text citation frequency is more than 5 then the citing and cited article have strong relevance. Similarly, Hou et al., (Hou, Li, & Niu, 2011), proposed a scheme in which they claimed that if reference found more than 10 times in body of the paper it holds strong relevance between citing and cited article. In citation classification community, many researchers have used in-text citation count of whole article or in-text citation count in a specific location in the paper or by combining both of these (Valenzuela, Ha, & Etzioni, 2015).

## 2.2 CUE WORDS/PHRASES

In 1992, Myers (Myers C. R., 1970), analyzed 50 articles from molecular genetics and reported that some phrases or words can provide cue about belonging to the particular reasons of citations. For example, (Swales, 1990), stated that cue phrases like “to our knowledge” or “as far as we are aware” demonstrate the gap in cited research. Similar Cue phrases are used by (Paice, 1981), to summarize text. In citation classification community, many researchers have used cue phrases that appear in body of the paper (Teufel, Siddharthan, & Tidhar, 2006) or appear in citation context (Valenzuela, Ha, & Etzioni, 2015).

In 1975, Moravcsik and Murugesan (Moravcsik & Murugesan, 1975), argued that all citations are not equal and studied some articles to identify few citation reasons. They divided the citations into four categories. (1) conceptual or operational (i.e., used just to describe a theory or used for technical purpose) (2) organic or perfunctory (i.e., Citing work is based on cited work or citing work is alternative of cited work) (3) evolutionary or juxtaposition (i.e., it is compulsory to read the cited work to understand the work is cited just for giving background knowledge) and (4) confirmative or negational (i.e., the citation is correct or not). In this analysis it is also considered that a citation can belong to more than one category. The dataset used for the study is based on 30 articles having 702 citations, which are selected randomly from Physical Review Spanning and published during the period of 1968 to 1972. The results of the study revealed that 40% citations belong to the perfunctory category. According to them, the results of this study increased suspicion on citation count approach.

Chubin and Moitra's (Chubin & Moitra., 1975), adopted Moravcsik and Muugesan's scheme with the slight amendments. For example, they left "Evolutionary/Juxtapositional" category of Moravcsik and Muugesan's approach due to considering the categories to be mutual exclusive. The dataset which is used for analysis is based on 44 articles having high-energy physics the subject of their research. The articles are taken from four journal Physical Review Letters, Physics Letters, Physical review and nuclear physics published between 1968 and 1969. The upshot of this study revealed that only 5% citations are from perfunctory category.

Ina Spiegel-Rosing (Spiegel-Rusing, 1977), categorized the citations into 13 categories which are sub categories of "cited source is positive or negative". The dataset used for analysis is based on 66 articles belonging from different disciplines. The dataset is selected from different Science Studies volumes. The outcome of the study disclosed that among all the categories, the category "substantiating a statement or an assumption made or pointing to further information" is most popular because 80% of the articles are from this category.

Oppenheim and Renn (Oppenheim & Renn., 1978), technique is slightly different in the terms that they analyzed why old papers are still being cited. For this purpose they analyzed 978 cited articles belonging to physics and chemistry discipline. They categorized the reasons of old papers citations into seven categories (1) Background knowledge (2) elaborating points from results (3) Specific usage of information (4) Comparisons (5) Usage of theoretical equation (6) Usage of practical methods to solve the problem (7) Criticizing the cited work. The study

revealed that 40% of old papers are being cited just to provide background knowledge.

Frost (Frost, 1979), proposed a technique to determine whether the work is cited because of some remarks or the citing work agrees/disagrees to the cited work in the field of humanities. For this, they classified citations into two broad categories (a) Documentation of primary sources and (b) Documentation of secondary sources. In each of the broad categories, there exist many sub categories like a.1. To support an opinion or factual statement on the specific literary author(s) or work(s) discussed in the citing work; a.2. To support an opinion outside the central topic of the citing work; or a.3. To support a factual statement outside the central topic of the citing work or b.1. Independent of approval or disapproval of the citing author.b.2. To acknowledge the pioneering work of other scholars; b.3. To indicate the state of present research, a range of opinions or prevailing views on a topic. The results of the study revealed that most of the citation belongs to b.2 category.

In 1979, Finney (Finney, 1979), initiated an idea that citations classification can be done automatically. For this purpose, she classified citations into 7 categories, (1) Background knowledge (2) Tentative references (3) Methodological references (4) Conformational references (5) Negational references (6) Interpretational references (7) Future research references. She associated cue words and citation location with citation function. According to (Garzone & Mercer, 2000) the (Finney, 1979), approach does not cover all aspects of being cited. In 1982, Bonzi (Bonzi, 1982), explored the parameters that can be promising to find relevance between cited and citing article. Total of 13 parameters are explored, that includes, (1) source of citation (2) date of citation cited and citing works (3) author self-citation (4) journal self-citation (5) type of journal (6) date of publication (7) sex of author (8) type of article (9) length of article (10) Number of citations (11) Number of citations in footnote (12) multiple mention of citations (13) placement of citation in text. For experimentation, they chose 31 articles having 500 citations and published in 19 different journals belonging to the library and information science. The results of the study revealed that source of cited work, source of citing work, number of times a work is cited in text, and type of citing article hold the potential to determine relevance between citing and cited article.

In 1999, Nanba and Okumura (Nanba & Okumura, 1999), classified citations into three classifications (1) Adopting cited work (2) providing background knowledge (3) Other than these two categories. According to citation classification community Nanba and Okumura (Nanba &

Okumura, 1999), stripped citation classification scheme of Garfield (Garfield E. , Can citation indexing be automated, 1965). This classification is done by summarizing the articles based on Cue phrases found around citation context.

The first automatic citation indexing scheme (CiteSeer) is proposed by Giles, Bollacker, and Lawrence (Giles, Bollacker, & Lawrence, 1998). The CiteSeer is a digital library and a search engine that focuses on the articles belong to computer science and information science. It crawls and harvests those documents which are freely available. This system provides the facility to automatically link the documents with their cited documents. The Garzone & Mercer (Garzone & Mercer, 2000) adopted this idea by enhancing the number of categories of (Finney, 1979). In 2000, the pioneer approach towards fully automatic citation classification is proposed by Garzone and Mercer (Garzone & Mercer, 2000). This system takes an article as input along with the set of citations and then produces suitable category to the citations. The citation categories are negational, affirmational, assumptive, tentative, methodological, interpretational/developmental, future research, use of conceptual, contrastive, and reader alert. These categories are further sub divided into 35 categories. The classification of citations is done by building a grammar of 195 lexical matching rules and 14 parsing rules relying on the cue words and section location of the citation. The technique is implemented by using 11 physics and 9 biochemistry articles. Out of which, 8 physics and 3 biochemistry articles are used for designing and 3 physics and 6 biochemistry articles are used for testing. The results are classified into three categories, (1) completely right (2) partially right and (3) completely wrong. The system achieved good results on seen articles and average results on unseen articles. However, in literature the results of this approach are contradictory because of having less number of rules and large number of categories which can conflict each other. Citation classification community focused on proposing different citation classification schemes by enhancing the number of features or parameters and considering only those classes which are important.

Pham and Hoffman (Pham & Hoffmann, 2003), developed a rule-based knowledge system based on cue phrases to classify citations. They classified citations into four categories, 1) basic 2) support 3) limitation 4) comparison. The rule-based knowledge system is made from 482 citation context. This classification is done by making using Ripple Down Rules (the “RDR”) hierarchy by using cue phrases found around citation context. RDR is same as decision trees. Total 482 citation context are used from which 150 are used for testing. They compared their results with

(Nanba & Okumura, 1999), and it is found that their system outperformed. The system achieved 95.2% accuracy. However, similar to (Nanba & Okumura, 1999), citation classification community argues that they stripped citation classification scheme of (Garfield, 1965).

Teufel et al. (Teufel, Siddharthan, & Tidhar, 2006), proposed supervised learning approach for citation classification, in which they differentiated between citation categories on the basis of linguistic rules. Their classification scheme is adoption of (Spiegel-Rusing, 1977), scheme. They categorized citations into four categories: (1) neutral (2) weakness (3) comparisons (4) compatibility. These categories are further divided into 11 categories. They annotated 26 articles having 548 citations. They built 892 linguistic cue phrases and used them to classify citations into the particular category. This system was trained on 90% dataset and tested on 10% dataset. This classification scheme achieved 0.71 F-measure. The results revealed that 65% of the citations belong to the neutral category.

Shahid et al., (Shahid, Afzal, & Qadir, 2011), proposed a technique in which semantic relationship between cited and citing paper is determined with the help of in-text citation frequency. They claimed that if in-text citation frequency is more than 5 five times in the citing paper then cited paper are semantically related to the citing paper. The dataset which is used for experiments is extracted from J.U.C.S. Total 16404 paper-reference pairs are examined. The results of the study revealed that if citation pairs having citation frequency more than 5, they have strong semantic relationship.

Similar to Shahid et al, (Shahid, Afzal, & Qadir, 2011), approach Hou et al., (Hou, Li, & Niu, 2011), introduced an idea to count the frequency of citation within text of the paper. They claimed that high frequency of citation appearing within text of the paper has potential of being influential citation. They analyzed 651 articles published in 2008 in the area of “Biochemistry & Molecular Biology” and “Genetics & Heredity” in the Web of Science. The analysis is done on the basis of Closely Related References (the “CRRs”) and Least Related References (the “LRRs”). CRRs are those which appear 10 or more times in body of the paper and LRRs are those which appear less than 10 times in body of the paper. The results of the study revealed that CRRs are found more frequently in texts of the articles than LRRs.

Agarwal et al., (Agarwal, Choubey, & Yu, 2010), classified citations into eight categories: (1) background/perfunctory (2) contemporary (3) contrast/conflict (4) evaluation (5) explanation of results (6) material/method (7) modality (8) similarity/consistency. They used 43 open access



articles in the field of biomedical science for experiments. These articles are annotated by the authors of papers themselves. The annotation is done on the basis of cue phrases found within citation context and the sentence appears before and after the citation context. The cue-phrases are picked by the annotators. Total 2977 annotations are done from 1710 sentences. The classification is done by using Support Vector Machine (the “SVM”) and Multinomial Naïve Based (the “MNB”) based models. The system achieved average F-measure score of 0.76.

In literature, researchers have proposed different techniques to assign weight to the citations. Citations are assigned on the basis of different factors such as, citing journal’s prestige (Ding, 2011). In 2012, Balaban (Balaban, 2012), presented a technique in which author claimed that citations done by the eminent authors should be given more weight and further claimed the paper belonging to low impact factor journal is cited by some prestigious journal’s article, this shows that the cited article is of high importance.

In 2011, Dong and Schäfer (Dong & Schäfer, 2011), classified citation into three categories positive, Negative and Neutral believing in the fact that large number of categories can conflict each other. For this purpose they expanded the organic/perfunctory category of (Moravcsik & Murugesan, 1975) into four dimensions (1) background (2) fundamental idea (3) technical basis and (4) comparison. They experimented on DFKI dataset of 120 articles having 1768 annotated citations from which 190 are annotated as positive, 57 as negative and 1521 as neutral. The features include Cue-phrases, in-text citation count and syntactical features. The technique achieved F-measure score of 0.66.

In 2012, Jochim and Schütze (Jochim & Schütze, 2012), classified the citations to determine the polarity (negative or positive) of citations. The basic idea is to demonstrate whether citing paper has taken an idea from cited paper, whether it demonstrates the correction or fault of cited paper, whether the cited work is fundamental or is a perfunctory, or whether the citing paper has adopted an idea from cited paper or represent an alternative scheme to the cited paper. They have collected 2008 citations from ACL anthology. From these 2008 citations, 1836 are annotated as positive and 172 are annotated as negative. This classification is done with the help of citation contexts having length of one, two and three sentences. From these sentences the four features are extracted unigrams, sentence location, word-level linguistic features and comparatives. The results of the study revealed that accuracy increases where the context length is greater than one. The best results are achieved for dimension 1 having F-measure score of 68.2.

In 2012, Liakata et al., (Liakata, Saha, Dobnik, Batchelor, & Rebholz-Schuhmann, 2012), implemented a system to automatically classify Core Scientific Concepts (the “CoreSCs”) of an article. The CoreSCs include, (1) hypothesis (2) motivation (3) goal (4) object (5) background (6) method (7) experiment (8) model (9) observation (10) result and (11) conclusion. For experiments 265 full articles belonging to the field of biochemistry and chemistry are examined. The features are based on full article context; these include unigrams, section location, lexicon and syntax of document. The classification is done by using SVM and CRF Classifier. The system achieved highest F-measure score of 76% for experiment CoreSCs.

In 2013, Meyers (Meyers, 2013), classified citation into two categories, Corroborate and Contrast. Corroborate using rate category demonstrates citing work is using the same approach used in cited work, contrast means different approach or opinion. The experiment is performed on 20 PubMed articles. The classification is done by using Random Forest Classifier. The system achieved 67% Recall for Contrast category and 83% for Corroborate category. However, in literature it is found that the results should be proven on large corpus.

In 2013, Li et al., (Li, He, Meyers, & Grishman, 2013) classified citations into three categories: (1) Positive (2) Neutral and (3) Negative. These three categories are further divided into 12 categories. The dataset used for experiment is taken from PubMed based on 91 annotated articles having 6,355 citations instances. The classification is done by using cue n-gram terms in citation context. The system achieved F-measure score of 0.67.

In 2013, Abu-Jbara and Radev (Abu-Jbara & Radev, 2011), classified the citations to determine the polarity of citation. They generated BoW (Bag of Words) by using subjectivity, speculation, and various others similar cue words to determine polarity. For generating BoW, the dataset is taken from ACL having 30 papers having 3,500 citations. The classification is done by using SVM classifier. The system achieved F-measure score of 0.58.

In 2013, Ciancarini et al.,(Ciancarini, Iorio, Nuzzolese, Peroni, & Vitali, 2013), classified citations into 13 categories. The categories include (1) agrees with (2) cites (3) cites as author (4) cites as authority (5) cites a data source (6) cites as evidence (7) cites as metadata document (8) cites as potential solution (9) cites as recommended reading (10) cites as related confirms (11) corrects (12) critiques (13) derides. The citation is done by using cue phrases generated from citation context. However, they did not report results of their experiments as they stated their work was preliminary in nature. In the present century, the citation classification community the

researchers started to focus on merging different reasons into two categories (1) important and (2) non-important, as the importance of citation is more important to make citation count approach a reliable via counting only those citations which are important.

Zhu et al. (Zhu, Turney, Lemire, & Vellino, 2015), classified the citations into two categories

- 1) Influential
- 2) Non-influential

This classification is done by using these five features (1) In-text count based (2) Similarity based (3) Context based (4) Position-based and (5) Miscellaneous. The idea behind the technique is to identify those references which have an academic influence to the citing paper. By the means of influential here is a reference from which the idea, problem, method, experiment is adopted. The term influential has been used first by Narin (Narin, 1976), where they found academic influence of journal. The dataset used for experiments is taken from ACL anthology. They performed experiments on paper-reference pairs. Total 3143 paper-reference pairs are formed from 100 papers. The pairs are annotated by the authors of papers themselves. The classification is done by using SVM classifier. The final results revealed that in-text citation count feature outclassed other features with the Precision 0.35.

Valenzuela et al., (Valenzuela, Ha, & Etzioni, 2015), proposed a novel approach for identification important and non-important citations. According to them, it is the first approach which focused on the problem of important citations identification. They classified citations into two categories: (1) Important and (2) Incidental. Total of 465 paper-citations pairs are taken from ACL anthology. These pairs are annotated as important and non-important. The annotated data is publicly available for experimentation. The pairs are mapped into important and Non-important class by using 12 features. The features include (1) total number of direct citations (2) number of direct citations per section (3) total number of indirect citations and number of indirect citations per section (4) author overlap (5) being helpful (6) citation appears in table or caption (7) number of references (8) number of paper citations / all citations (9) similarity between abstracts (10) page rank (11) number of total citing papers after transitive (12) field of the cited paper. These features are trained on SVM and Random Forest classifier. The achieved F-measure score of the approach is 0.65. Out of all features, the in-text citation count feature outperformed with Precision 0.37.

## 2.3 CRITICAL ANALYSIS

After the comprehensive analysis of state of the art approaches in the field, we found that techniques for citation classification are based on Cue phrases and In-text citation count. The brief overview of these techniques is described in Table 2-1 below along with their results and limitations.

**Table 2-1 Critical Analysis of State of the art approaches**

<b>Authors</b>	<b>Feature</b>	<b>Results</b>	<b>Limitations</b>
<b>(Nanba &amp; Okumura, 1999)</b>	Cue phrases	Precision = 0.76	<ul style="list-style-type: none"> <li>• Cue words need to be defined manually which is time consuming</li> </ul>
<b>(Garzone &amp; Mercer, 2000)</b>	Cue phrases	Good results on seen articles and average results on unseen articles	<ul style="list-style-type: none"> <li>• Set of cue words need to be defined manually which is time consuming</li> <li>• Defining linguistic rules require expert human knowledge</li> <li>• The defined categories are so large in number that they can</li> </ul>

			conflict with each other
<b>(Teufel, Siddharthan, &amp; Tidhar, 2006)</b>	Cue Words	F-measure = 0.68	<ul style="list-style-type: none"> <li>• Those citations are not annotated in which their manually selected word does not appear (e.g; “better”, “used by us)”</li> <li>• Cue phrases are selected manually and the list need to be updated for new dataset</li> </ul>
<b>(Valenzuela, Ha, &amp; Etzioni, 2015)</b>	In-text citation count based features	In-text citation count feature outperformed with Precision = 0.37 Overall Precision=0.65	<ul style="list-style-type: none"> <li>• Ignores the important cue phrases immediately before and after the citation context</li> </ul>
<b>(Zhu, Turney, Lemire, &amp; Vellino, 2015)</b>	In-text citation count	In-text citation count feature outperformed with Precision=0.35	<ul style="list-style-type: none"> <li>• Ignores the important cue phrases immediately</li> </ul>

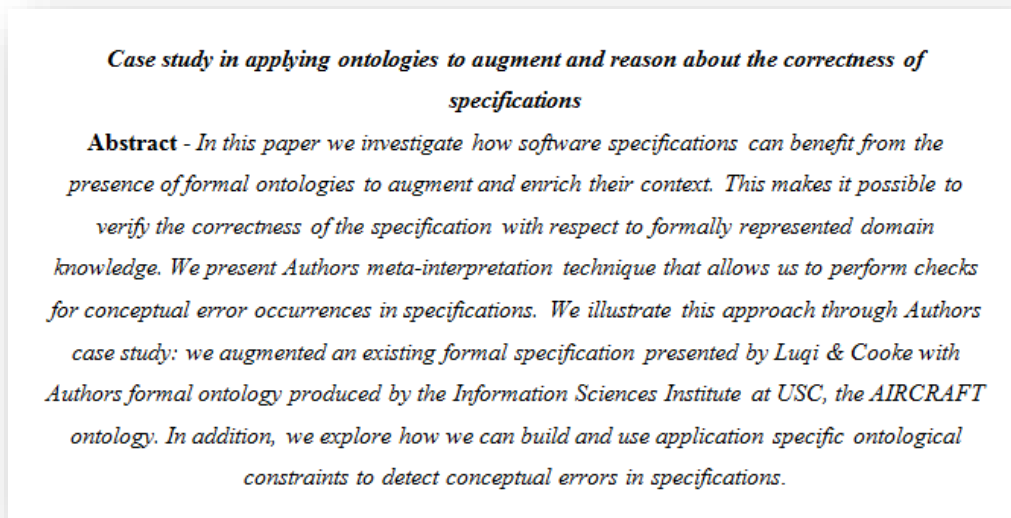
			before and after the citation context
--	--	--	------------------------------------------------

In Table 2-1, it can be seen that all reviewed approaches are content dependent. In a recent citation classification approach by (Valenzuela, Ha, & Etzioni, Identifying Meaningful Citations, 2015) and (Zhu, Turney, Lemire, & Vellino, 2015), the In-text citation frequency based feature performed well as compared to other features. In-text Citation Frequency approach claims that if the frequency of in-text citation in citing paper is 5 or more times, then citing and cited papers are relevant to each other and if the frequency is less than 5 then papers are not relevant to each other (Shahid, Afzal, & Qadir, 2011). But this is not always true as described in example below

### 2.3.1 In-text Citation Limitation

#### a. High Frequency but Low Relevance

Contemplate two papers X and Y. One is “Citing Paper” (Paper X, see figure 2-2) and the other is “Cited Paper” (Paper Y, see figure 2-3).



**Figure 2-2 PAPER X**

**How to combine nonmonotonic logic and rapid prototyping to help maintain software**

**Abstract** -In this paper explores the possibility of automated support for detecting inconsistencies in software systems and requirements. The inconsistencies are introduced when the environment of the software system changes. We refer to the software environment as its context. We review the recent research progress on nonmonotonic logics, pointing out the significance of these results to software maintenance. We explain how Authors practical implementation of such logics can be obtained via Authors simple extension to logic programming in the form of an answer procedure that realizes the Extended Logic Semantics [7] for nonmonotonic logic programs that have Authors unique answer set (which is Authors large and useful class of logic programs).We augment the existing automated capabilities of the Computer-Aided Prototyping System (CAPS) for rapid prototyping via the extension to logic programming to provide an improved automated capability for detecting certain kinds of inconsistencies created by implicit requirements changes. We illustrate the significance of this capability via an example prototype for Authors problem originally suggested by Lehman.

**Figure 2-3 PAPER Y**

The information about both papers is presented in the table 2-2 below. The paper X cites 10 times paper Y. Having gone through the content of both articles, it is analyzed that the papers are not related to each other. The claim of in-text citation frequency fails here which postulates that the citing paper and cited paper are related if the frequency of in-text citation in citing paper is 5 or higher (Shahid, Afzal, & Qadir, 2011). On the contrary, if we scrutinize the metadata of both papers, it can be seen that paper X and Y do not share any similarity between titles and authors.

**Table 2-2 The paper X cites 10 times paper Y**

<b>Paper X</b>	<b>Paper Y</b>
<i>Title: case study in applying ontologies to augment and reason about the correctness of specifications</i>	<i>Title: How to Combine Nonmonotonic Logic and Rapid Prototyping to Help Maintain Software</i>
<i>Authors: Yannis Kalfoglou, David Robertson</i>	<i>Authors: Luqi, Daniel Cooke</i>
<i>Keywords: Not found</i>	<i>Keywords: Not found</i>

**b. Low Frequency, High Relevance**

Contemplate two papers A and B. One is “Citing Paper” (Paper A, see figure 2-4) and the other is “Cited Paper” (Paper B, see figure 2-5).

***PAPER A: Measuring Semantic Similarity between Biomedical Concepts within Multiple Ontologies***

***Abstract***—Most of the intelligent knowledge-based applications contain components for measuring semantic similarity between terms. Many of the existing semantic similarity measures that use ontology structure as their primary source cannot measure semantic similarity between terms and concepts using multiple ontologies. This research explores Authors new way to measure semantic similarity between biomedical concepts using multiple ontologies. We propose Authors new ontology-structure-based technique for measuring semantic similarity in single ontology and across multiple ontologies in the biomedical domain within the framework of Unified Medical Language System (UMLS). The proposed measure is based on three features: (1) cross-modified path length between two concepts; (2) Authors new feature of common specificity of concepts in the ontology; and (3) local granularity of ontology clusters. The proposed technique was evaluated relative to human similarity scores and compared with other existing measures using two terminologies within UMLS framework: Medical Subject Headings and Systemized Nomenclature of Medicine Clinical Term. The experimental results validate the efficiency of the proposed technique in single and multiple ontologies, and demonstrate that our proposed measure achieves the best results of correlation with human scores in all experiments

**Figure 2-4 PAPER A**

***PAPER B: An approach for measuring semantic similarity between words using multiple information sources***

***Abstract*** -Semantic similarity between words is becoming Authors generic problem for many applications of computational linguistics and artificial intelligence. This paper explores the determination of semantic similarity by Authors number of information sources, which consist of structural semantic information from Authors lexical taxonomy and information content from Authors corpus. To investigate how information sources could be used effectively, variety of strategies for using various possible information sources are implemented. A new measure is then proposed which combines information sources nonlinearly. Experimental evaluation against Authors benchmark set of human similarity ratings demonstrates that the proposed measure significantly outperforms traditional similarity measures

**Figure 2-5 PAPER B**

The information about both papers is presented in the Table 2-3 below. The Paper A cites Paper B only once. After scrutinizing content of both articles, it is analyzed that the papers are strongly related to each other. The claim of in-text citation frequency approach fails here which envisions that the Citing Paper and Cited Paper are not related if the frequency of in-text citation in Citing Paper is less than 5 (Shahid, Afzal, & Qadir, 2011). Conversely, if we analyze the metadata of both papers, it can be seen that Paper A and B share the similarity between titles, authors and



keywords.

**Table 2-3 The Paper A cites paper B only one time**

<b>Paper A</b>	<b>Paper B</b>
<b>Title:</b> <b>Measuring Semantic Similarity Between</b> Biomedical Concepts Within Multiple Ontologies	<b>Title:</b> An Approach for <b>Measuring Semantic Similarity between</b> Words Using Multiple Information Sources
<b>Author:</b> Hisham Al-Mubaid	<b>Authors:</b> Yuhua Li, Zuhair A. Bandar, and David McLean
<b>Keywords:</b> Biomedical information retrieval, biomedical ontology, biomedical terminology, <b>Semantic similarity</b> , Unified Medical Language System (UMLS).	<b>Keywords:</b> <b>Semantic similarity</b> , lexical database, information content, corpus statistics

The illustrations above narrate that how that in-text citation frequency does not perform well all the time. Moreover, to get the in-text citation count, it is paramount to go through content the paper and most of the time content is not freely available. Journals of all major publishers like IEEE, ACM, Springer, Elsevier and IOS do not provide open access to their articles. There are financial, legal and technical barriers hampering access to content of the paper. Alternatively, various kinds of useful metadata associated with research papers such as title, keywords, authors, categories, references etc. are freely available.

## **Chapter 3**

### **PROPOSED METHODOLOGY**

The comprehensive analysis of state-of-art approaches in previous chapter depicts that in citation classification community researchers have proposed useful techniques to classify citations. As per our knowledge, no classification scheme exists that relies fully on freely available metadata. Our technique focuses on binary classification via supervised machine learning: Given an article, classify its citations as either important or non-important by exploiting their metadata. In this chapter the detailed methodology to tackle the problem of important citations identification is described. The figure 3.1 is a graphical representation of whole proposed system. Each chunk of figure 3.1 is described in detail.

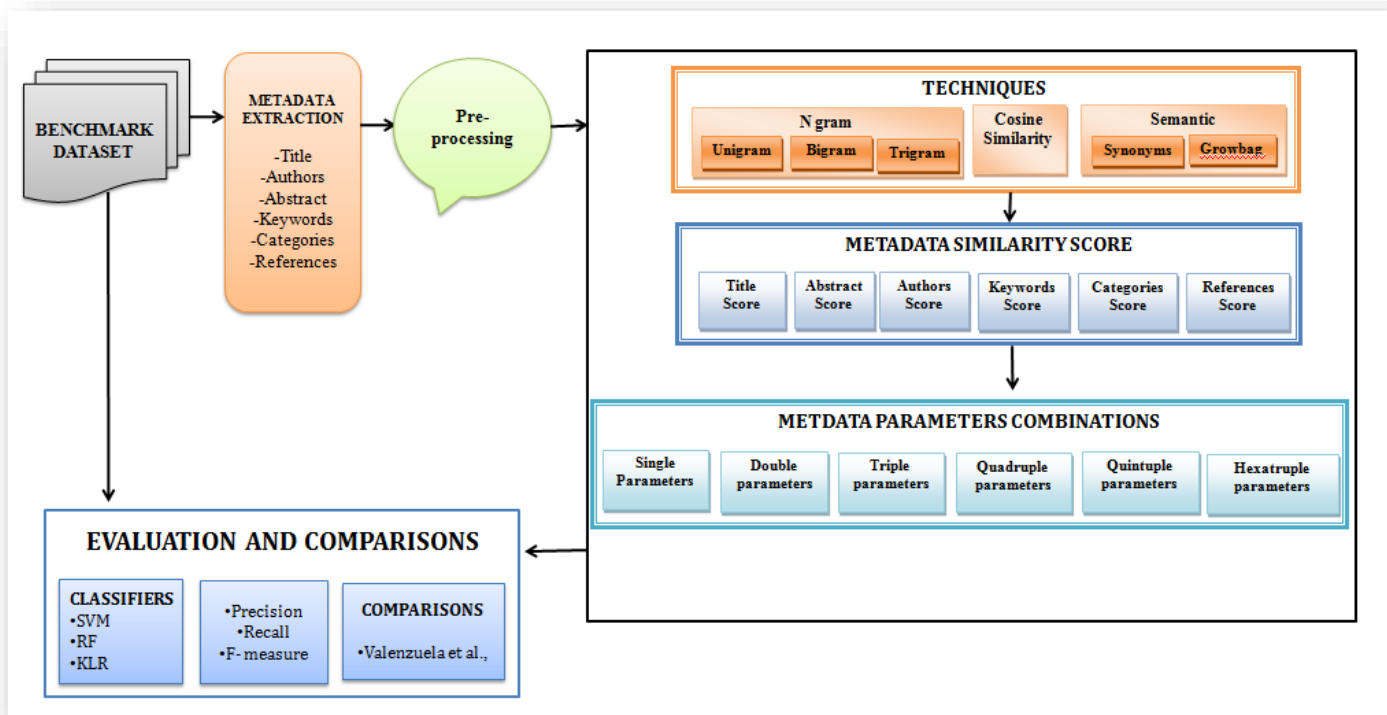


Figure 3-1 Context Diagram of Proposed System

### 3.1 BENCHMARK DATASETS

To classify citations into important and non-important categories, there is a need of some standard dataset. We preferred to use dataset collected by Valenzuela et al., (Valenzuela, Ha, & Etzioni, 2015), by considering different factors: (1) This is the benchmark dataset and is available online for experiments (2) Using same dataset; the comparison of outcomes with their approach would be more justified. The second dataset is collected and annotated from Capital University of Science and Technology (CUST) Computer Science faculty members. Being part of this institution, it would be convenient for us to annotate the citations and references from actual authors of the papers, because we think authors are in the best position to label their citing and cited work. Let's discuss these two datasets in more detail.

#### 3.1.1 Dataset1

This is the benchmark dataset taken by Valenzuela et al., (Valenzuela, Ha, & Etzioni, 2015), available online for experiments. There are total 465 annotated paper-citation pairs collected from Association of Computational and Linguistics (ACL) anthology belonging to the field of Information Systems. This dataset will be referred as *d1* from hereafter. ACL anthology is a digital archive of research papers in computer linguistics and a citation network which contains only those papers and their citations which are published in ACL anthology. The figure 3-2 demonstrates the description of dataset in a better way. The first column represents the annotator who annotated these pairs. The dataset is annotated by the two domain experts. The second column contains the source paper ID of ACL anthology. The third column contains the IDs of citation paper of source paper. The fourth column "Follow Up" contains the score assigned by the annotators (i.e. the score of 0 for *Non-important* and 1 for *Important* paper-citations pairs).

	A	B	C	D
1	<b>Annotator</b>	<b>Paper</b>	<b>Cited-by</b>	<b>Follow-up</b>
2	A	A00-1043	C00-2140	0
3	A	A00-1043	P02-1057	0
4	A	A97-1011	W09-1118	1
5	A	A97-1011	A00-2017	1
6	A	A97-1011	C00-2099	0
7	A	A97-1011	W04-1505	0
8	A	A97-1011	P99-1033	0

**Figure 3-2 Benchmark dataset1**

### 3.1.2 Dataset2

Another dataset having 500 paper-reference/citation pairs is collected from CUST Computer Science faculty members designated on the positions of associate and assistant professors. We went to them and asked: *kindly provide us any of your research paper that has the maximum no. of citations and references.* We have collected all the required information of citations and references from Google scholar. We provided them source paper and list of its all references and citations with the abstract and author's information, and asked them to kindly label those references and citations as important, from which source paper has adopted an idea or idea has been adopted from source paper, have done a similar work,. We are thankful to them for their co-operation in generating this gold standard dataset. This dataset will be referred as *d2* from hereafter.

### 3.2 METADATA EXTRACTION

After getting the information about citations and references of *d1* and *d2*, the next step is the extraction of important metadata parameters and their score calculation for supervised machine learning. For *d1*, the metadata parameters are extracted from ACL anthology by using paper ID provided in benchmark dataset (see figure 3.2). For *d2* we have the complete list of citations and references articles as described above (see section 3.1.2). From all the articles, these metadata parameters based on their free availability, are manually extracted.

- a) Title
- b) Authors
- c) Abstract
- d) Keywords
- e) Categories
- f) References

Since, the papers published in ACL anthology do not contain “Keywords” and “Categories”. Therefore, the metadata parameters of *d1* are based on title, authors, abstract and references. The overview of extracted metadata parameters of *d1* and *d2* is described in figure 3-3 and 3-4 respectively. However, we were unable to extract few parameters as some papers do not contain abstract in both *d1* and *d2*. In *d2*, some authors have not assigned keywords or categories to their papers and so on. The detailed stats of availability and successful extraction of metadata

parameters is reported in chapter 4.

	A	B	C	D	E
1	<b>Paper ID</b>	<b>Title</b>	<b>Authors</b>	<b>Abstract</b>	<b>References</b>
2	A00-1043	Sentence Reduction for Ar	Hongyan Jing	Detecting the linguistic	E. Apostolova and N. Tomuro. 20
3	C00-2140	DiaSumm: Flexible Summar	Klaus Zechner ,Alex W	In this paper, we preser	Arto Anttila. 1995. How to recogn
4	A97-1011	A non-projective depende	Pasi Tapanainen , Timo	We describe a practical	A. Berger and H. Printz. 1998. Rec
5	A00-2017	A Classification Approach	Yair Even-Zohar , Dan	The eventual goal of a	Hiyan Alshawi. 1996. Head autom
6	C00-2099	A Statistical Theory of De	Christer Samuelsson	A generative statistical	R. Artstein and M. Poesio. 2008.
7	E12-1072	Elliphant: Improved Auton	Luz Rello,Ricardo Baez	In pro-drop languages,	Chinatsu Aone and Scot W. Ben
8	P01-1006	Evaluation tool for rule-ba	Catalina Barbu,Ruslan	In this paper we argue	Steven Abney. 1996. Partial pars
9	P99-1033	Dependency Parsing with	Kemal Oflazer	This paper presents a c	Steven Abney. 1995. Chunks and
10	W04-1505	Fast, Deep-Linguistic Stati	Gerold Schneider,Fabi	We present and evalua	Razvan Bunescu and Raymond M

Figure 3-3 Extracted Metadata Parameters of *d1*

	A	B	C	D	E	F	G
1	<b>Sr #</b>	<b>Title</b>	<b>Authors</b>	<b>Abstract</b>	<b>Keywords</b>	<b>Categories</b>	<b>References</b>
2	1	JavaSymphony	Muhamma	Today, softwa	Not found	Not found	Denis Caromel; M.L.: P
3	2	Scheduling Jav	Muhamma	JavaSymphon	Not found	Not found	Aleem; M.; Prodan; R.
4	3	On the Evalua	Muhamma	Programming	Not found	Not found	Aleem; M.; Prodan; R.
5	4	Parallelism as	Cristian M	We are facing	Java,parallel so	Not found	W. Kim and M. Voss; "
6	5	The JavaSym	Muhamma	Today, the us	Parallel softwa	Not found	NVIDIA GTX490 Specif
7	6	A semi-autom	Hirsch, Ma	Because of th	Parallel softwa	Not found	Doug Lea. The java.uti
8	7	Feedback-Dire	Fengguang	This paper de	Distributed sha	D.3.4 [Softw	T. Davis. University of
9	8	ProActive Par	Denis Car	The Proactive	Not found	Not found	Marco Aldinucci; Soni

Figure 3-4 Extracted Metadata Parameters of *d2*

### 3.3 TITLES EXTRACTION FROM REFERENCES

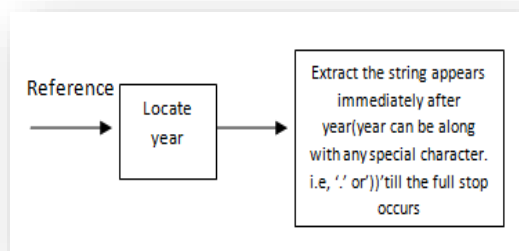
Most of the time, the articles that do the similar work are more likely to cite the same articles in their bibliography. Based on this assumption we have matched the *n* references of the “source paper” with the *n-1* references of “cited by” paper. The *n-1* references is due to the fact that the reference of source paper will appear in bibliography of the “Cited by” paper and of course it wouldn’t be present in the source paper bibliography, therefore, it needs to be excluded from bibliography of “Cited By” paper. For this purpose, the reference of source paper from

bibliography section of “Cited By” articles is manually removed. Since, it was found during references matching that reference of a same article is written in a different way in citing and cited article. Consider the example in figure 3-5, the same article is cited in a different way in terms of writing author’s name.

- ✦ Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. *In European Association for Machine Translation.*
- ✦ Hildebrand, A. S., Eck, M., Vogel, S., & Waibel A. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. *In European Association for Machine Translation.*

**Figure 3-5 Difference between Same Reference Patterns**

We preferred to match only titles of all references because we believe every article holds a unique title. The titles of references are extracted by applying heuristic approach described in figure 3-6. To ensure the correction, the extracted titles are verified manually. This heuristic helped us to extract 89% of the titles. The extraction of remaining 11% titles is done manually. The method of titles extraction is described in example 3-6.



**Figure 3-6 Heuristic Approach to Extract Titles from References**

### 3.4 PRE-PROCESSING

There are few parameters that needed to be cleaned (i.e., titles) and stemmed (i.e., titles, keywords) for experimentation. The stop words removal and stemming is done on different parameters. Let’s discuss it step by step.

### **3.4.1 Stop Words Removal**

In English language the words like, is, the, a, which, at, in etc is almost found in multiple sentences. Therefore, their removal is necessary to get the unique terms from titles. To remove stop words from titles of papers, the widely used Onix Text Retrieval Toolkit Stop Words List<sup>1</sup> is utilized.

### **3.4.2 Stemming**

In our experiments, the terms of titles and keywords are converted into their root terms via stemming. For example, a source paper title has the term “parallel” and its cited paper had used the term “Parallelizing”, they wouldn’t be matched if we apply our approach without stemming. The stemming is done by using porter stemmer algorithm (Porter, 1980), which converts all the terms of titles into their root terms. For example, “parallel”, “parallelizing” and “parallelism” will convert into their root term “parallel”. The stemming algorithm is applied on (1) Titles (2) Keywords

## **3.5 TECHNIQUES**

### **3.5.1 N-Gram Technique**

The idea of using N-gram was proposed by Locke (Locke, 1956), where he drew an analogy between machine translation and cryptography. Till now numerous researchers have used N-grams techniques in different machine learning problems like text summarization, classifying citations etc (Zhu, Turney, Lemire, & Vellino, 2015). In our experiments, the title similarity score is calculated by considering unigram, bigram and trigrams terms. After pre-processing, the terms of titles are divided into N-gram chunks. The examples below provide a better overview of title terms conversion into unigram, bigram and trigram. Consider the title “Sentence Reduction for Automatic Text Summarization”. The stemmed title will become “Sentenc Reduct Automat Text Summar”. The table 3-1 below shows the total 5 unigram terms.

---

<sup>1</sup> <http://www.lextek.com/onix/>



**Table 3-1 Unigram Terms**

Sr #	Unigrams
1	Sentenc
2	Reduct
3	Automat
4	Text
5	Summar

Similarly, the bigram terms of stemmed title are shown in table 3-2, and trigrams terms are shown in table 3-3. The terms are split into unigram, bigram and trigram by using “n-gram” library is R tool.

**Table 3-2 Bigram terms**

Sr #	Bigrams
1	Sentenc Reduct
2	Reduct Automat
3	Automat Text
4	Text Summar

**Table 3-3 Trigram terms**

Sr #	Trigrams
1	Sentenc Reduct Automat
2	Reduct Automat Text
3	Automat Text Summar

### 3.5.2 Synonyms

Usually it is seen that two person use different words to present the same thing, as everyone is not aware of every word in English vocabulary. Therefore, synonyms dataset is used to get maximum matching between terms of titles and keywords. In order to enrich the results, terms of titles are replaced with their synonyms for best results. The synonyms are matching by using

WordNet<sup>2</sup> library. The table 3-4 shows the example of word and its synonyms.

**Table 3-4 Word and its synonyms**

<b>Word</b>	<b>Synonyms</b>
<b>Distributed</b>	Dispersed
	Spread
	Disseminated
	Circulated
	Scattered

### **3.5.3 Growbag**

While doing experiments it is seen that some terms are strongly related to each other but they are not synonyms of each other. For example, semantic web and RDF are strongly related to each other but they are not synonyms of each other. Such relationship can be found by using Growbag algorithm by Diederich and Balke (Diederich & Balke, 2007), which divides the words into first order co-occurrence and second order co-occurrence of more than two million research papers indexed in DBLP<sup>7</sup>. This algorithm has produced 0.3 million such strong semantic relationship between words. In this thesis, the first order co-occurrence (strongly related terms) are used to enrich metadata similarity. Table 3.4 shows the word and its strongly related terms. In order to enrich the results, the unigram, bigram and trigram terms of titles and terms of keywords are replaced and matched with their Growbag terms. The following combinations are applied on titles and keywords to get maximum matching through Growbag.

- i. Title - Title
- ii. Title - Growbag
- iii. Growbag - Title
- iv. Growbag - Growbag

The final score of matched title terms is obtained by taking average of all these four combinations. Since we have stemmed the terms of our title, therefore, all terms in Growbag dataset are also stemmed for accurate matching. The example of word and its strongly related terms is given in table 3-5.

---

<sup>2</sup> <http://wordnet.princeton.edu/>

**Table 3-5 Semantically Related Terms**

Word	Strongly Related Terms
<b>Semantic web</b>	Resource Description Framework or RDF
	Web Ontology Language or OWL
	Extensible Markup Language or XML
	SPARQL

### 3.6 SCORE CALCULATION FOR SUPERVISED LEARNING

We want to investigate what ratio of similarity can produce useful results; therefore, the similarity between metadata of all pairs is calculated.

Let  $\langle S_i, c_{ij} \rangle$  be a paper-reference or paper-citation pair,

where  $S_i$  is the  $i^{\text{th}}$  source paper and  $c_{ij}$  is the  $j^{\text{th}}$  reference or citation of  $S_i$ .

Let  $p_n$  be the  $n^{\text{th}}$  parameter in our parameters set and let  $v(S_i, C_{ij}, p_n)$  be the value of parameter  $p_n$  in paper-citation/reference pair  $\langle S_i, c_{ij} \rangle$ .

Suppose that  $S_i$  contains  $q$  citations and references  $(S_i, c_{i1}), \dots, (S_i, c_{iq})$ , resulting in  $m$  values for  $p_n, v(S_i, c_{i1}, p_n), \dots, v(S_i, c_{in}, p_n)$ .

Let  $P$  be a set of parameters.  $P = \{p_u, p_b, p_t, p_a, p_{ab}, p_k, p_c, p_r\}$

$p_u = \{ \text{List of unigram terms present in titles of } S_i, \text{ and } c_{ij} \}$

$p_b = \{ \text{List of bigram terms present in titles of } S_i \text{ and } c_{ij} \}$

$p_t = \{ \text{List of trigram terms present in titles of } S_i \text{ and } c_{ij} \}$

$p_a = \{ \text{List of authors present in } S_i \text{ and } c_{ij} \}$

$p_k = \{ \text{List of keywords present in } S_i \text{ and } c_{ij} \}$

$p_c = \{ \text{List of categories present in } S_i \text{ and } c_{ij} \}$

$p_r = \{ \text{List of titles of references present in } S_i \text{ and } c_{ij} \}$

$m = \text{Total no. of citation papers}$

### 3.6.1 Title Similarity Score

The formulas in equations below calculate the score of matched N-gram (i.e., unigram, bigram and trigram) terms between titles of citing and cited papers. The equation 3.1 calculates the score between unigram terms of titles for each paper-citation pair. The equation 3.2 calculates the score between bigram terms of titles and equation 3.3 calculates the score between trigram terms. The detailed results of all formulas are described in chapter 4.

$$P1_{ij} = \frac{|(S_i(p_u)) \cap \sum_{j=1}^m (C_{ij}(p_u))|}{|(S_i(p_u)) \cup \sum_{j=1}^m (C_{ij}(p_u))|} \quad (3.1)$$

$$P2_{ij} = \frac{|(S_i(p_b)) \cap \sum_{j=1}^m (C_{ij}(p_b))|}{|(S_i(p_b)) \cup \sum_{j=1}^m (C_{ij}(p_b))|} \quad (3.2)$$

$$P3_{ij} = \frac{|(S_i(p_t)) \cap \sum_{j=1}^m (C_{ij}(p_t))|}{|(S_i(p_t)) \cup \sum_{j=1}^m (C_{ij}(p_t))|} \quad (3.3)$$

For instance, consider two titles “Semantic Similarity Computation” and “Semantic Similarity in Biomedical Ontologies”. The P1 score between both titles would be  $|(\text{Semantic}), (\text{Similarity})| / |(\text{Semantic}), (\text{Similarity}), (\text{Computation}), (\text{Biomedical}), (\text{Ontologies})| = 2/5$ . P2 score would be  $|(\text{Semantic Similarity})| / |(\text{Semantic Similarity}), (\text{Similarity Computation}), (\text{Similarity Biomedical}), (\text{Biomedical Ontologies})| = 1/5$ . Similarly, P3 score would be 0.

### 3.6.2 Author Overlap Score

It is seen that most of the time author of citing paper extends or adopt an idea from his previously done work. Based on this assumption, in this thesis the author’s similarity score is calculated to find out what ratio of similarity can provide better accuracy in tracing *important* citations. The author similarity score is calculated by using formula in equation 3.4. The detailed results are described in chapter 4.

$$P4_{ij} = \frac{|(S_i(p_a)) \cap \sum_{j=1}^m (C_{ij}(p_a))|}{|(S_i(p_a)) \cup \sum_{j=1}^m (C_{ij}(p_a))|} \quad (3.4)$$

For instance consider authors of source paper “M. Mujtaba, Z. Hasham, W. David” and citation paper “M. Mujtaba, S. Zhu”. The P4 score would be 1/4.

### 3.6.3 Abstract Similarity

The abstract of research article describes the purpose, hints that idea is adopted from someone's work and briefly demonstrates overall outcome of the article. If high similarity exists between abstract of research articles, this increases the chances that current work extends the previous work. Based on this assumption, the abstract similarity between paper-citation pairs is calculated. The similarity is computed by using cosine similarity of *tf-idf* scores. The cosine similarity between two terms or documents on the vector space is a measure that calculates the cosine of the angle between them. In machine learning, cosine similarity between two documents is calculated to examine how much the content in two documents is similar. In this thesis, the similarity is computed by using cosine similarity of *tf-idf* scores of abstract of citing and cited papers. We are applying cosine similarity because it is preferred over other similarity measures in literature (Valenzuela, Ha, & Etzioni, 2015). The formula to calculate cosine similarity is given in equation 3.5.

$$\cos(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| |\vec{d}|} = \frac{\vec{q}}{|\vec{q}|} \cdot \frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|\mathcal{V}|} q_i d_i}{\sqrt{\sum_{i=1}^{|\mathcal{V}|} q_i^2} \sqrt{\sum_{i=1}^{|\mathcal{V}|} d_i^2}} \quad (3.5)$$

### 3.6.4 Keywords Similarity

In any research article, keywords depict the domain and description of the paper. The authors of research articles choose these keywords in the way which becomes easy for readers to get to know about the domain and flow of the research work. In this thesis, the freely available author's assigned keywords are exploited believing in the fact that similar keywords of citing and cited articles increase the chances of being *important* paper-citation pair. Similar to title similarity scheme the synonyms and semantic approach is applied here as well to get maximum matching. The score of keywords is calculated by using formula in equation 3.6. The detailed results are described in chapter 4.

$$P5_{ij} = \frac{|(S_i(p_k)) \cap \sum_{j=1}^m (C_{ij}(p_k))|}{|(S_i(p_k)) \cup \sum_{j=1}^m (C_{ij}(p_k))|} \quad (3.6)$$

For instance, consider a keywords of source paper “web mining, machine learning, content similarity” and keywords of citation paper “machine learning, supervised learning, classification”. The P5 score between both keywords would be 1/5.

### 3.6.5 Categories Similarity

Similar to keywords, categories of research paper depicts the category of research article from where it belongs to, which eases to get the idea of research flow or domain. ACM classification system is one such system that has defined 13 top level categories in the domain of computer science. ACM classification technique is adopted globally. Most of research articles publishing conferences and journals use this categorization system. It is sub-divided into various other categories, as research article can belong to more than one category. The similarity score of categories is calculated by using formula in equation 3.7. The detailed results are described in chapter 4. The example of P6 score computation is same as of P5.

$$P6_{ij} = \frac{|(S_i(p_c)) \cap \sum_{j=1}^m (C_{ij}(p_c))|}{|(S_i(p_c)) \cup \sum_{j=1}^m (C_{ij}(p_c))|} \quad (3.7)$$

### 3.6.6 Bibliographically Coupled References

Most of the time, most relevant papers cite same work in their bibliography. So frequently the references of citing and cited papers are matched, the chance of being its *important* paper-citation pair increases. The references title similarity score between pairs calculated by using the formula in equation 3.8. The detailed results against this formula are explained in chapter 4.

$$P7_{ij} = \frac{|(S_i(p_r)) \cap \sum_{j=1}^m (C_{ij}(p_r))|}{|(S_i(p_r)) \cup \sum_{j=1}^m (C_{ij}(p_r))|} \quad (3.8)$$

For instance, consider there are four same titles in source and citation papers, and there are total 10 and 13 references in source and citation papers respectively, the P7 score would be 4/19.

## 3.7 METDATA PARAMETERS COMBINATIONS

After score calculation of all metadata parameters, these are combined into different levels to explore which combinations provide best accuracy. Total  ${}^n C_r$  combinations of metadata parameters are explored, where n is the total number of parameters and r is the size of combination (i.e, single, double, triple etc). In the case of titles, the unigram, bigram and trigrams are combined individually with other parameters. Let's discuss these levels one by one.

### 3.7.1 Single Metadata Parameters

In single metadata parameters, it is analyzed that out of title, authors, abstract, keywords, categories and references, which metadata parameter has produced the best result.

### 3.7.2 Double Metadata Parameters

In double metadata parameters, every possible combination of two metadata parameters is examined to analyze which produces the best results. There are total 25 double metadata parameters combinations are analyzed as described in figure 3-7.

- |                                |                                |
|--------------------------------|--------------------------------|
| 1. Title_Unigram + Author      | 14. Title_Bigram + References  |
| 2. Title_Bigram + Author       | 15. Title_Trigram + References |
| 3. Title_Trigram + Author      | 16. Authors + Abstract         |
| 4. Title_Unigram + Abstract    | 17. Authors + Keywords         |
| 5. Title_Bigram + Abstract     | 18. Authors + Categories       |
| 6. Title_Trigram + Abstract    | 19. Authors + References       |
| 7. Title_Unigram + Keywords    | 20. Abstract + Keywords        |
| 8. Title_Bigram + Keywords     | 21. Abstract + Categories      |
| 9. Title_Trigram + Keywords    | 22. Abstract + References      |
| 10. Title_Unigram + Categories | 23. Keywords + Categories      |
| 11. Title_Bigram + Categories  | 24. Keywords + References      |
| 12. Title_Trigram + Categories | 25. Categories + References    |
| 13. Title_Unigram + References |                                |

Figure 3-7 Double Metadata Parameters Combinations

### 3.7.3 Triple metadata Parameters

In triple metadata parameters, every possible combination of three metadata parameters is analyzed to determine which produces the best results. There are total 40 triple metadata parameters combinations as described in figure 3-8

- |                                           |                                            |
|-------------------------------------------|--------------------------------------------|
| 1. Title_Unigram + Authors + Abstract     | 21. Title_Trigram + Abstract + References  |
| 2. Title_Bigram + Authors + Abstract      | 22. Title_Unigram + Keywords +Categories   |
| 3. Title_Trigram + Authors + Abstract     | 23. Title_Bigram + Keywords +Categories    |
| 4. Title_Unigram + Authors + Keywords     | 24. Title_Trigram + Keywords +Categories   |
| 5. Title_Bigram + Authors + Keywords      | 25. Title_Unigram+Keywords + References    |
| 6. Title_Trigram + Authors + Keywords     | 26. Title_Bigram+ Keywords + References    |
| 7. Title_Unigram + Authors + Categories   | 27. Title_Trigram+ Keywords + References   |
| 8. Title_Bigram + Authors + Categories    | 28. Title_Unigram+Categories+ References   |
| 9. Title_Trigram + Authors + Categories   | 29. Title_Bigram + Categories + References |
| 10. Title_Unigram + Authors + References  | 30. Title_Trigram+Categories + References  |
| 11. Title_Bigram + Authors + References   | 31. Authors + Abstract + Keywords          |
| 12. Title_Trigram + Authors + References  | 32. Authors + Abstract + Categories        |
| 13. Title_Unigram + Abstract + Keywords   | 33. Authors + Abstract + References        |
| 14. Title_Bigram + Abstract + Keywords    | 34. Abstract + Keywords + Categories       |
| 15. Title_Trigram + Abstract + Keywords   | 35. Abstract + Keywords + References       |
| 16. Title_Unigram + Abstract + Categories | 36. Abstract + Categories + References     |
| 17. Title_Bigram + Abstract + Categories  | 37. Keywords + Categories + References     |
| 18. Title_Trigram + Abstract + Categories | 38. Categories + References + Authors      |
| 19. Title_Unigram + Abstract + References | 39. Categories + References + Abstract     |
| 20. Title_Bigram + Abstract + References  | 40. Categories + References + Keywords     |

**Figure 3-8 Triple Metadata Parameters Combinations**

### **3.7.4 Quadruple Metadata Parameters**

In quadruple metadata parameters, every possible combination of four metadata parameters is analyzed to determine which produces the best results. There are total 35 quadruple parameters combinations as described in figure 3-9.



1. Title\_Unigram+ Authors + Abstract + Keywords
2. Title\_Bigram + Authors + Abstract + Keywords
3. Title\_Trigram + Authors + Abstract + Keywords
4. Title\_Unigram + Authors + Abstract + Categories
5. Title\_Bigram + Authors + Abstract + Categories
6. Title\_Trigram + Authors + Abstract + Categories
7. Title\_Unigram + Authors + Abstract + References
8. Title\_Bigram + Authors + Abstract + References
9. Title\_Trigram + Authors + Abstract + References
10. Title\_Unigram + Abstract + Keywords + Categories
11. Title\_Bigram + Abstract + Keywords + Categories
12. Title\_Trigram + Abstract + Keywords + Categories
13. Title\_Unigram + Abstract + Keywords + References
14. Title\_Bigram + Abstract + Keywords + References
15. Title\_Trigram + Abstract + Keywords + References
16. Title\_Unigram + Authors + Keywords + Categories
17. Title\_Bigram + Authors + Keywords + Categories
18. Title\_Trigram + Authors + Keywords + Categories
19. Title\_Unigram + Authors + Keywords + References
20. Title\_Bigram + Authors + Keywords + References
21. Title\_Trigram + Authors + Keywords + References
22. Title\_Unigram + Authors + Categories + References
23. Title\_Bigram + Authors + Categories + References
24. Title\_Trigram + Authors + Categories + References
25. Title\_Unigram + Keywords + Categories + References
26. Title\_Bigram + Keywords + Categories + References
27. Title\_Trigram + Keywords + Categories + References
28. Categories + References + Title\_Unigram + Abstract
29. Categories + References + Title\_Bigram + Abstract
30. Categories + References + Title\_Trigram + Abstract
31. Authors + Abstract + Keywords + Categories
32. Authors + Abstract + Keywords + References
33. Authors + Keywords + Categories + References
34. Abstract + Keywords + Categories + References
35. Categories + References + Authors + Abstract

**Figure 3-9 Quadruple Metadata Parameters Combinations**

### 3.7.5 Quintuple Metadata Parameters

In quintuple parameters, every possible combination of five metadata parameters is analyzed to obtain which produces best results. There are total 13 quintuple parameters combinations as described in figure 3-10.

- |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ol style="list-style-type: none"> <li>1. Title_Unigram + Authors + Abstract + Keywords + Categories</li> <li>2. Title_Bigram + Authors + Abstract + Keywords + Categories</li> <li>3. Title_Trigram + Authors + Abstract + Keywords + Categories</li> <li>4. Title_Unigram + Authors + Abstract + Keywords + References</li> <li>5. Title_Bigram + Authors + Abstract + Keywords + References</li> <li>6. Title_Trigram + Authors + Abstract + Keywords + References</li> <li>7. Authors + Abstract + Keywords + Categories + References</li> </ol> | <ol style="list-style-type: none"> <li>8. Abstract + Keywords + Categories + References + Title_Unigram</li> <li>9. Abstract + Keywords + Categories + References + Title_Bigram</li> <li>10. Abstract + Keywords + Categories + References + Title_Trigram</li> <li>11. Keywords + Categories + References + Title_Unigram + Authors</li> <li>12. Keywords + Categories + References + Title_Bigram + Authors</li> <li>13. Keywords + Categories + References + Title_Trigram + Authors</li> </ol> |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

**Figure 3-10 Quintuple Metadata Parameters Combinations**

### 3.7.6 Hextuple Metadata Parameters Combinations

In hextuple parameters, every possible combination of six metadata parameters is analyzed to obtain which produces best result. There are total 3 hextuple parameters combinations as described in figure 3-11.

1. Title\_Unigram + Authors + Abstract + Keywords + Categories + References
2. Title\_Bigram + Authors + Abstract + Keywords + Categories + References
3. Title\_Trigram + Authors + Abstract + Keywords + Categories + References

**Figure 3-11 Hextuple Metadata Parameters Combinations**

## 3.8 CLASSIFIERS

In citation classification community, researchers have classified the citations into different categories by using different classifiers. Every classifier has its own importance, to classify

citation into *Important* and *Non-important* classes, we have utilized (1) The Random Forest (RF) (2) Support Vector Machine (SVM) and (3) Kernel Logistic Regression (KLR) Machine learning classifiers. The reason of using these classifiers is due to their high use in literature where citations are classified into important and non-important classes. The detailed results against each classifier are explained in Chapter 4.

### 3.9 EVALUATION AND COMPARISONS

To evaluate the results of our proposed technique, the standard formula of Precision, Recall and F-measure is calculated. The formula of Precision Recall and F-measure is demonstrated in equation 3.9, 3.10 and 3.11 respectively.

$$\mathbf{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (3.9)$$

$$\mathbf{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (3.10)$$

$$\mathbf{F - measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.11)$$

The results of our proposed technique will be compared with the results of (Valenzuela, Ha, & Etzioni, 2015), as we have used the same dataset with different proposed parameters.

## Chapter 4

### EXPERIMENTS AND RESULTS

In the previous chapter, the comprehensive methodology to solve the existing gap is explained in detail. This chapter focuses on the results achieved by applying that methodology.

#### 4.1 DATASET COLLECTION

Our experiments are based on two datasets as discussed in chapter 3. The dataset1 *d1*, is a based on Valenzuela et al., (Valenzuela, Ha, & Etzioni, 2015), they have collected and annotated 465 *paper-citation* pairs, from which 14.6% pairs are annotated as *Important* and remaining 85.4% are annotated as *Non-important*. While downloading all the pairs, the articles of 33 pairs are not found on ACL anthology. The experimentation is done on remaining 432 *paper-citation* pairs. Out of 33 pairs, 11 were *Important* and 22 were *Non-important*. The amount of remaining 432 pairs is described in Table 4-1.

**Table 4-1 Successful Extraction of Articles in *d1***

<b>CLASS</b>	<b>CITATION</b>
<b>Annotated pairs in <i>d1</i></b>	465
<b>Whose metadata was available in ACL</b>	432
<b><i>Non-important</i></b>	375
<b><i>Important</i></b>	57

From dataset *d2*, all the citations and references articles of the source papers are collected from Google Scholar. However, those references and citations which are other than research articles for examples, link of websites, link of some tool or book etc are excluded from *d2*, because these citations and references do not contain those metadata which is required for our experiments. The amount of successful extraction of all *paper-reference* and *paper-citation* pairs is described in table 4-2. The experiments are performed on remaining 324 pairs.

**Table 4-2 Successful Extraction of Articles in *d2***

<b>DATA</b>	<b>NO. OF INSTANCES</b>
<i>Paper-reference pairs</i>	298
<i>Paper-citation pairs</i>	202
<b>References and Citations other than research articles</b>	158
<b>Not found</b>	18

#### **4.2 METADATA EXTRACTION**

The next step is the extraction of metadata parameters from collected source papers, citations and references. There are two ways to extract these parameters, (1) manual and (2) machine oriented. We preferred manual extraction method in order to have maximum accurate extraction. For all the pairs, we have collected titles, authors, abstract, keywords, categories and references as discussed in previous chapter. However, the *d1* does not contain keywords and categories as all articles found on ACL anthology do not have these metadata parameters. Therefore, keywords and categories are not present in *d1*. The amount of successful extraction of other metadata parameters is described in table 4-3.

**Table 4-3 Successful Extraction of Metadata Parameters in *d1***

<b>METADATA PARAMETER</b>	<b>SUCCESSFUL EXTRACTION PERCENTAGE</b>
<b>Titles</b>	100%
<b>Authors</b>	100%
<b>Abstract</b>	99.7%
<b>References</b>	100%

Similarly, in table 4-4 the percentage of extracted metadata parameters from *d2* is described.

**Table 4-4 Successful Extraction of Metadata Parameters in *d2***

<b>METADATA PARAMETER</b>	<b>SUCCESSFUL EXTRACTION PERCENTAGE</b>
<b>Titles</b>	100%
<b>Authors</b>	100%
<b>Abstract</b>	98.7%
<b>Keywords</b>	58.3%
<b>Categories</b>	4.3%
<b>References</b>	93.2%

In *d2*, all the titles and authors of pairs are extracted successfully, abstract of 3 papers are not found, only 189 pairs contain the keywords and 322 pairs contain references. In case of categories, only 14 pairs contain the categories, the conclusion therefore could not be made on the basis of such small amount of categories. Hence, the categories parameter has been skipped from our experiments. However, 3 out of 5 categories between *important paper-citation* pairs are matched and no category matched between *Non-important paper-citation* pairs, which hint that categories can be a useful metadata parameter to identify *important* pairs but still we cannot demonstrate a generic conclusion based on this small amount.

### **4.3 PRE-PROCESSING**

After all metadata extraction, there are some parameters that needed to be cleaned such as titles, and stemmed, such as titles, keywords and Growbag dataset. These two steps are involved in preprocessing step.

- (1) Removal of stop words from titles using Onix Stop Words Toolkit<sup>3</sup>.
- (2) Conversion of titles, keywords, synonyms and Growbag terms into their root terms by using porter stemmer algorithm (Porter, 1980).

### **4.4 SYNONYMS AND GROWBAG MATCHING**

To get the maximum matching of titles terms (for *D1* and *D2*), keywords matching (for *D2*) between pairs, the synonyms and Growbag technique is applied as discussed in chapter 3.

---

<sup>3</sup> <http://www.lextek.com/onix>

Unfortunately, only 3 synonym terms of *D1* titles are acquired from WordNet library, therefore the synonym matching scheme has been skipped from our experiments. In case of Growbag matching scheme, 43% semantically related terms of titles of *D1*, 56% semantically related terms of titles of *D2* and 61% semantically related terms of keywords of *D2* are found. Hence, the Growbag scheme has played vital role in achieving good results.

#### **4.5 TITLES EXTRACTION FROM REFERENCES**

In extracted references, 92% references follow the same structure as of figure 3-5 (see chapter 3). The titles of such references are retrieved by extracting the string appears immediately after the year appears, till the full stop appears. The remaining 8% references varies in structure, therefore the titles of those references are manually extracted.

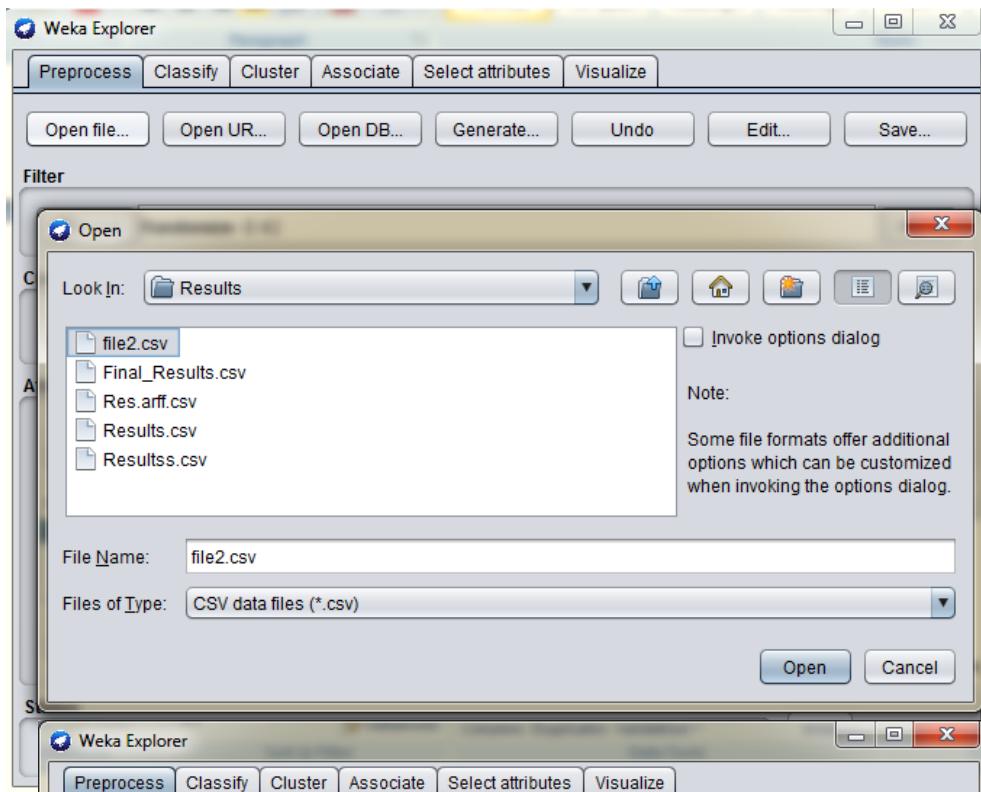
#### **4.6 SCORE CALCULATION**

After the preprocessing and splitting terms of titles into unigram, bigram and trigram, all the extracted metadata parameters (1) Titles (2) Authors (3) Abstract (4) Keywords and (5) References are ready for experiments. All the proposed formulas described in chapter 3 (see section 3.6) are applied on these parameters. The resulting score of each formula lies between 0 and 1.

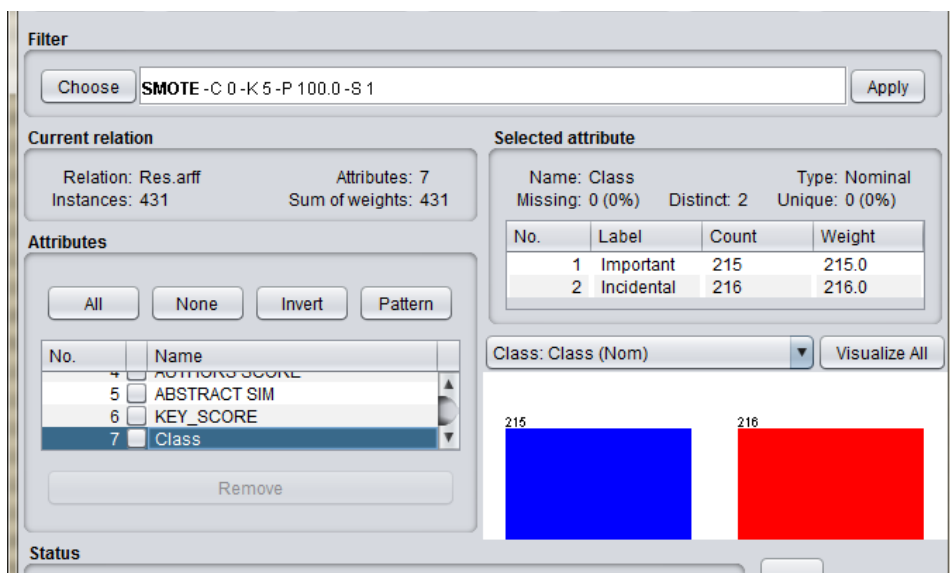
#### **4.7 CLASSIFICATION OF PAIRS**

The classification of each pair is done on the basis of scores obtained by applying all the formulas described in chapter 3. The popular suite of machine learning WEKA (Waikato Environment for Knowledge Analysis) is utilized for classification (Garner, 1995). Our features are the scores achieved against each metadata parameters (see section 3.6). These features and their combinations are manually selected and different machine learning algorithms are applied in WEKA. In the figures 4-2 to 4-5, the classification is done by combining all features of *d2* and applying Random Forest Classifier to give an idea that how classification is done in WEKA. Since, we have class imbalanced problem as number of *non-important* or say a negative classes are greater than positive classes (i.e. 57 vs 375 for *d1* and 92 vs 216 for dataset 2). To solve this problem, the SMOTE (Synthetic Minority Oversampling Technique) filter is applied (see figure 4-3). The SMOTE equalizes the number of positive and negative instances for better

classification (see figure 4-4). After applying SMOTE, the scores in features are randomly shuffled by using Randomize filter in WEKA preprocessing panel (see figure 4-4), as these both techniques help in better classification (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). In figure 4-5, classification using all parameters of *d2* and Random Forest classifier is presented. The same method of classification is applied to evaluate metadata parameters and their combinations. The detailed evaluation of each feature and their combinations against different classifiers is discussed in evaluation step.



**Figure 4-1 File Loading in WEKA**



**Figure 4-2 Balanced Classes using SMOTE Filter**



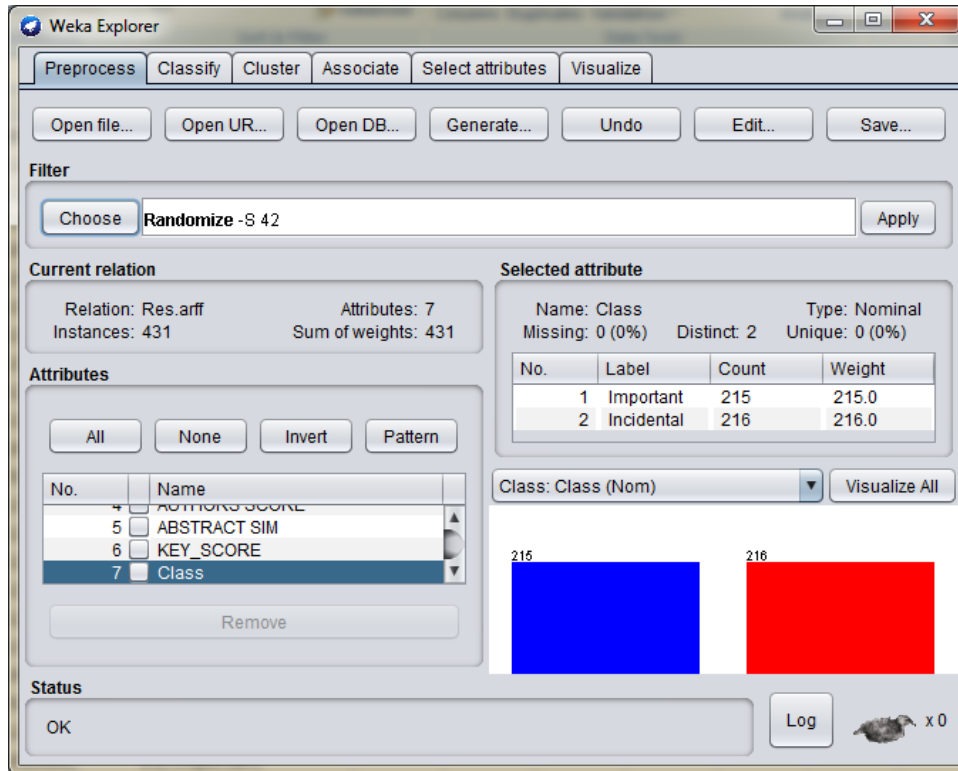


Figure 4-3 Randomizing Classes

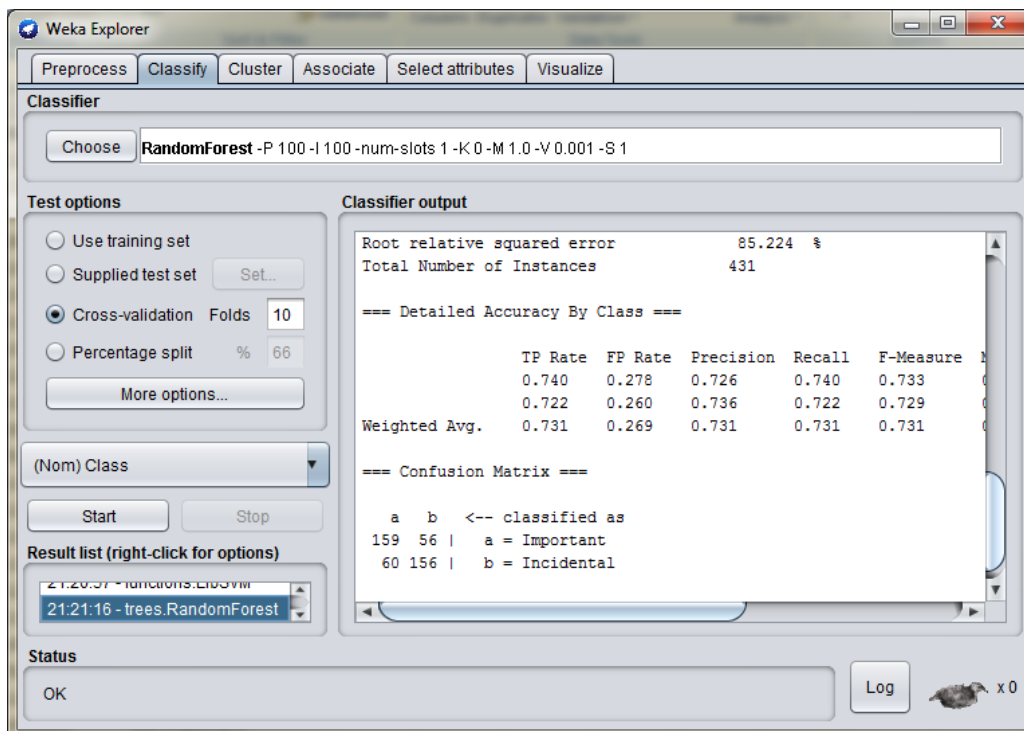


Figure 4-4 Hybrid Classification Using Random Forest classifier for  $d_2$

## 4.8 EVALUATION

The standard formula of Precision, Recall and F-measure is applied for evaluation. The Random Forest, SVM and Kernel Logistic Regression Machine learning algorithms using 10-fold cross validation are applied for classification. The reason of using these classifiers is due to their high usage in literature where citations are classified into *Important* and *non-important* classes (Valenzuela, Ha, & Etzioni, 2015; Zhu, Turney, Lemire, & Vellino, 2015). To analyze the contribution of each feature individually and by making their different combinations, we performed a post-hoc analysis, where we evaluated variants of our model containing single to multiple parameters group. While building all possible combinations, we have considered only metadata of only those pairs whose all parameters in the combinations are available to avoid biasness. The results of top 3 metadata combinations are reported in this section.

### 4.8.1 Single Metadata Parameters

The classification based on every metadata parameter alone is helpful to draw a conclusion about which parameter has contributed more in achieving the best results. The Precision, Recall and F-measure score based on above mentioned three classifiers is calculated and the average Precision, Recall and F-measure score is obtained by calculating the arithmetic mean of all three classifiers score. Out of all metadata parameters, the *Title\_Bigram* has achieved the highest average Precision of 0.42, Recall of 0.59 and F-measure of 0.49, then *Title\_Unigram*, *Bibliographically Coupled References*, *Title-Trigram*, *Authors*, and *Abstract\_Similarity* respectively achieved good results as shown in figure 4-5. In case of *d2*, the similar results are achieved in case of *Title\_Bigram*, as *Title\_Bigram* parameter in *d2* outperformed other parameters with average Precision of 0.54, Recall of 0.53 and F-measure of 0.53, then *Authors*, *Title\_Unigram*, *Abstract\_SIM*, *keywords*, *Title-Trigrams* and *Bibliographically Coupled References* respectively achieved best scores as shown in figure 4-6. Similar behavior of *Title\_Bigram* parameter in both datasets shows that *Bigram* holds a strong potential in identification of *important paper-citation* pairs. For both *d1* and *d2*, the Random Forest classifier has achieved the best PRF scores among other classifiers.

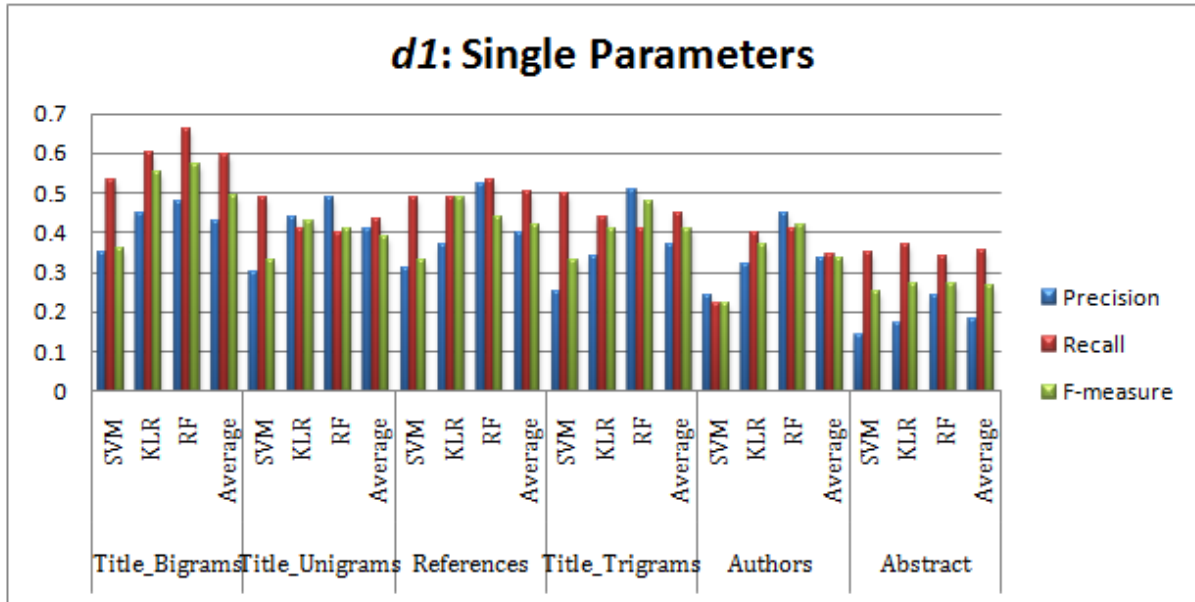


Figure 4-5 PRF Bar Chart for *d1* Single Parameters

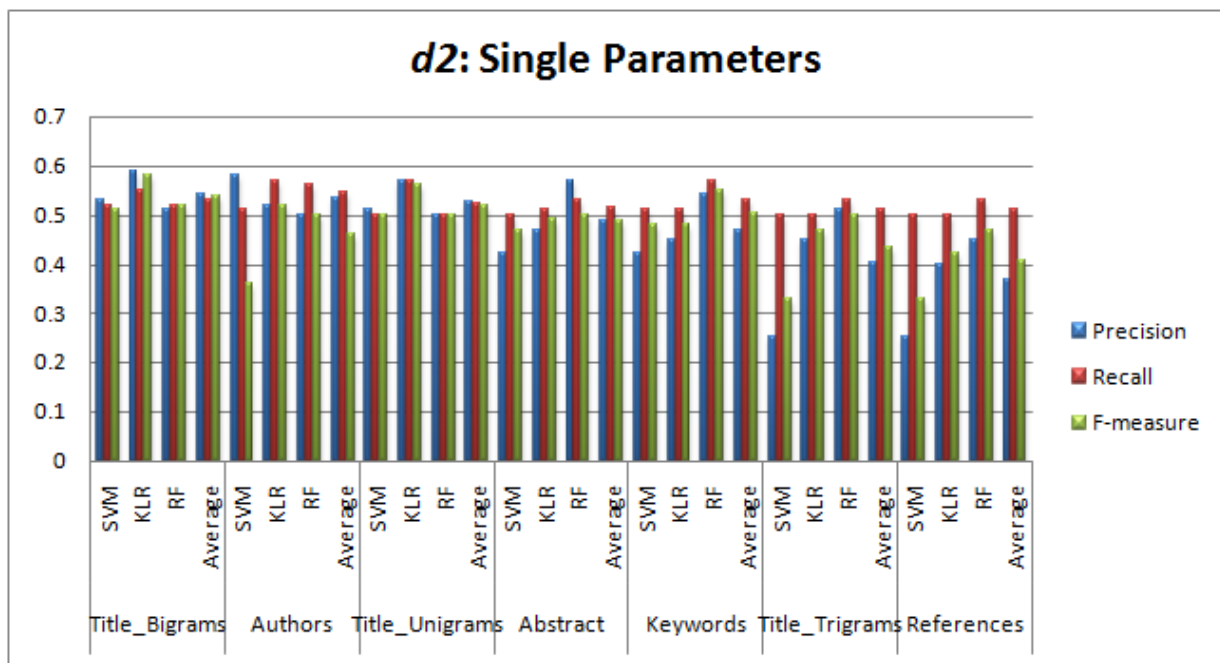
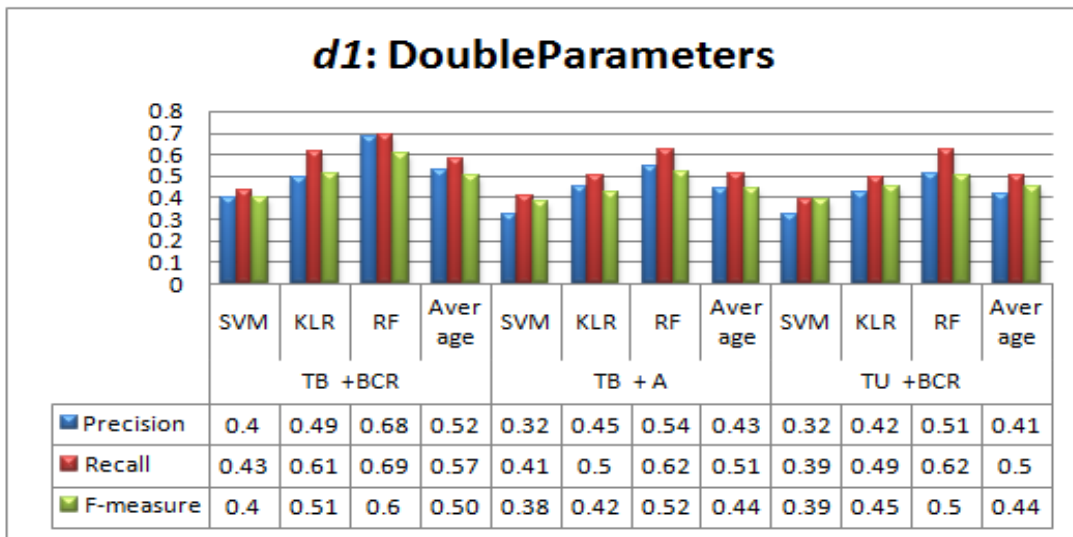


Figure 4-6 PRF Bar Chart for *d2* Single Parameters

### 4.8.2 Double Metadata Parameters

In double metadata parameters every possible combination of two metadata parameters is exploited to obtain Precision, Recall and F-measure scores against three classifiers. In the case of

*d1*, the “*Title\_Bigram + Bibliographically Coupled References*” combination outperformed other combinations with average Precision of 0.52, Recall of 0.57 and F-measure of 0.50. The second top scored combination is “*Title\_Bigram + Authors*” and the third one is “*Title\_Unigram + Bibliographically Coupled References*”. The scores obtained against each classifier are demonstrated in figure 4-7. For *d2*, the combination “*Title\_Bigram + Authors*” outperformed other combinations with the average Precision of 0.61, Recall of 0.64 and F-measure of 0.62. The second top scored combination is “*Title\_Bigram + Abstract*” and the third one is “*Title\_Unigram + Abstract*” shown in figure 4-8. Same as the results of single metadata parameters, the Random Forest classifier has achieved the best PRF scores among other classifiers. The abbreviation of metadata parameters presented in all figures contains Precision Recall, F-measure and average score are as follows: (1) TU: Title\_Unigram (2) TB: Title\_Bigram (3) TT: Title\_Bigram(4) A: Authors (5) Ab: Abstract (5) K: Keywords and (6) BCR: Bibliographically Coupled References



**Figure 4-7 PRF Bar Chart for *d1* Double Parameters**

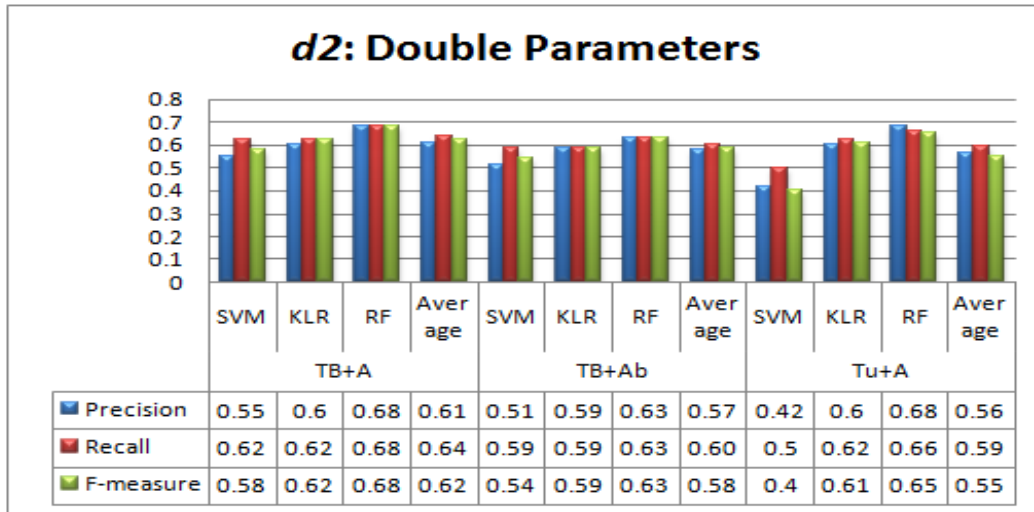


Figure 4-8 PRF Bar Chart for *d2* Double Parameters

### 4.8.3 Triple Metadata Parameters

In triple metadata parameters every possible combination of three metadata parameters is exploited to obtain Precision, Recall and F-measure scores against three classifiers. For *d1*, the best scored combination is “*Title\_Bigram + Authors + Bibliographically Coupled References*” with the average Precision of 0.59, Recall of 0.61 and F-measure of 0.59. The second top scored triple combination is “*Title\_Unigram + Abstract + Bibliographically Coupled References*” and the third one is “*Authors + Bibliographically Coupled References + Title\_Trigram*”. For *d1*, the KLR model has provided best PRF scores. The scores obtained against each classifier are envisioned in figure 4-9. For *d2*, “*Title\_Bigram + Authors + Abstract*” outperformed other parameters with the average Precision of 0.61, Recall of 0.65 and F-measure of 0.64. The second top scored triple combination is “*Title\_Bigram + Abstract + Bibliographically Coupled References*” and the third one is “*Title\_Bigram + Abstract + Keywords*” as it can be seen in figure 4-10. For *d1* and *d2* the Random Forest classifier has achieved best scores.

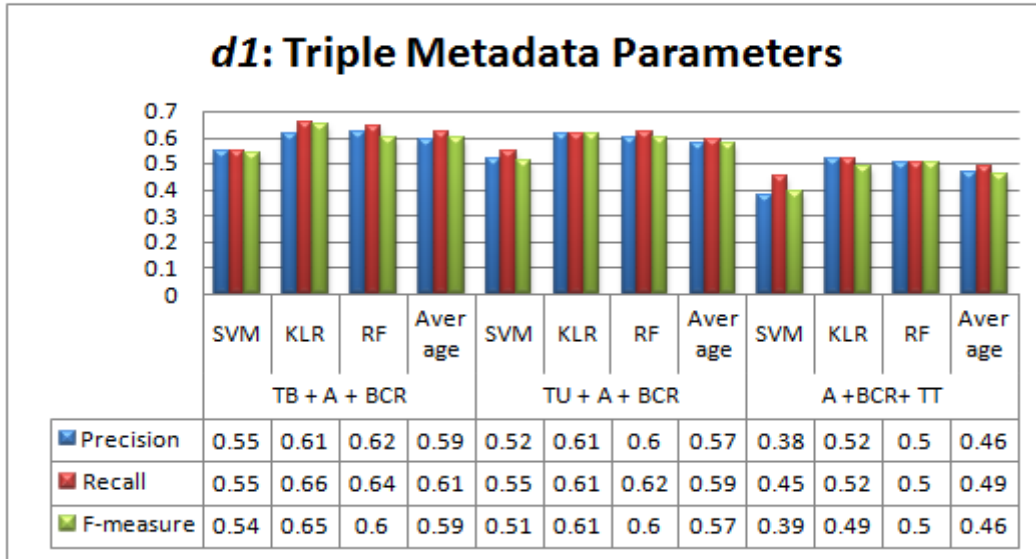


Figure 4-9 PRF Bar Chart for *d1* Triple Parameters

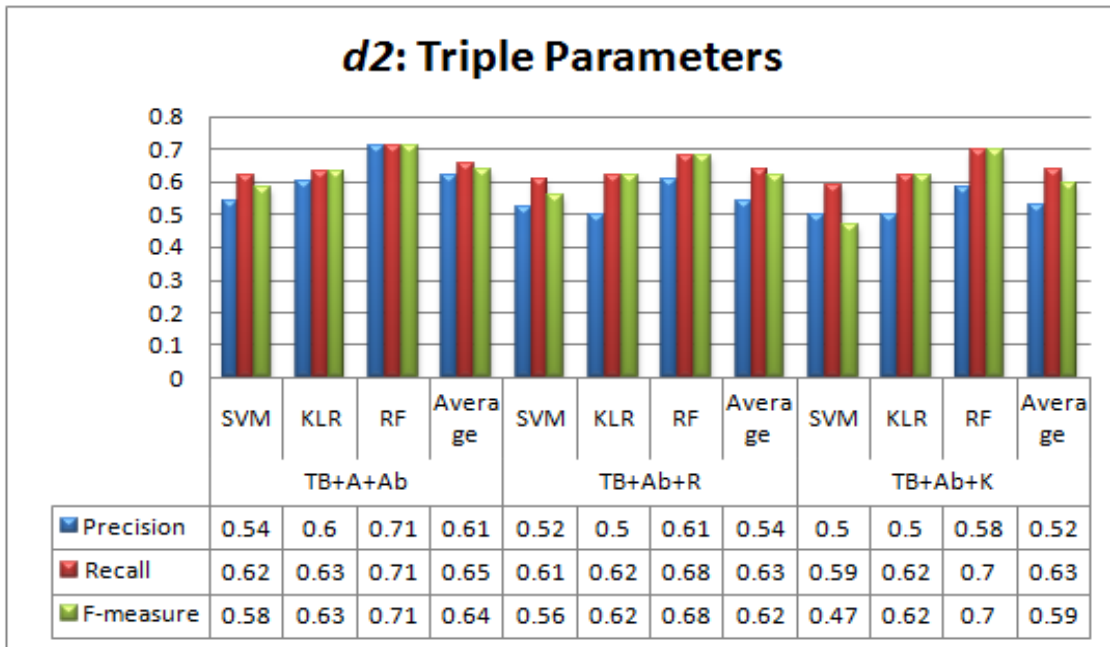


Figure 4-10 PRF Bar Chart for *d2* Triple Parameters

#### 4.8.4 Quadruple Metadata Parameters

In quadruple metadata parameters, every possible combination of four metadata parameters is exploited to obtain Precision, Recall and F-measure scores against three classifiers. For *d1*, The best scored combination is “*Title\_Bigram + Authors + Abstract + Bibliographically Coupled*

*References*” with the average Precision of 0.67, Recall of 0.73 and F-measure of 0.70, the second top scored combination is “*Title\_Unigram + Authors + Abstract + Bibliographically Coupled References*” and the third one is “*Title\_Trigram + Authors + Abstract + Bibliographically Coupled References*” as shown in figure 4-11. The analysis of all combinations for *d1* has been completed. For *d2*, “*Title Bigram + Authors + Abstract + Keywords*” outperformed other combinations with the average Precision of 0.65, Recall of 0.63 and F-measure of 0.62, the second top scored combination is “*Title\_Unigram + Authors + Abstract + Keywords*” and the third one is “*Authors + Abstract + Keywords + Bibliographically Coupled References*” as shown in figure 4-12. For both *d1* and *d2*, the Random Forest classifier achieved best PRF scores among other classifiers.

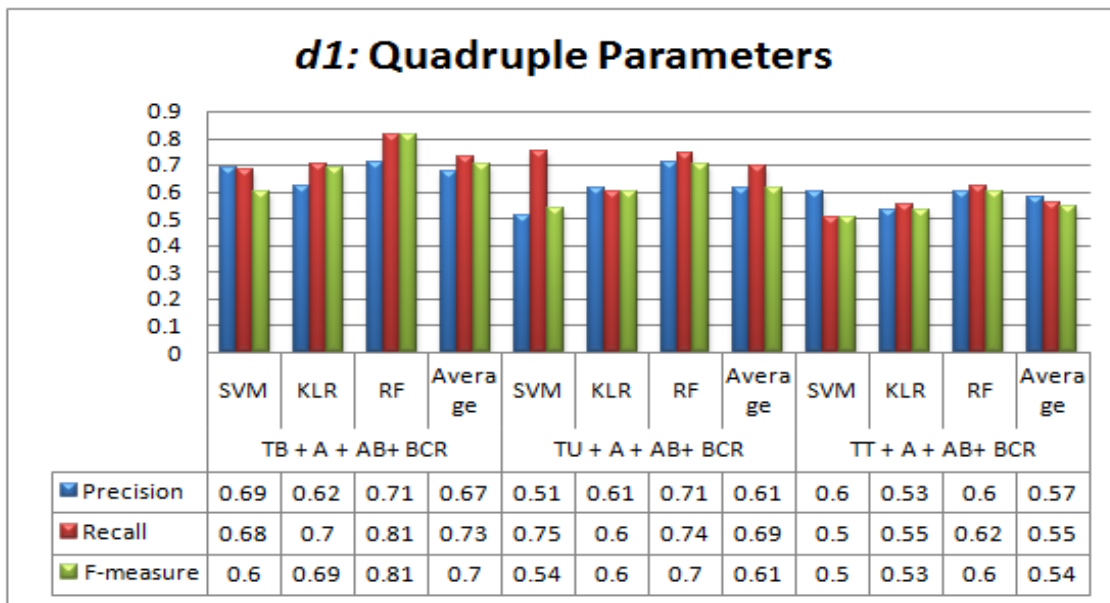
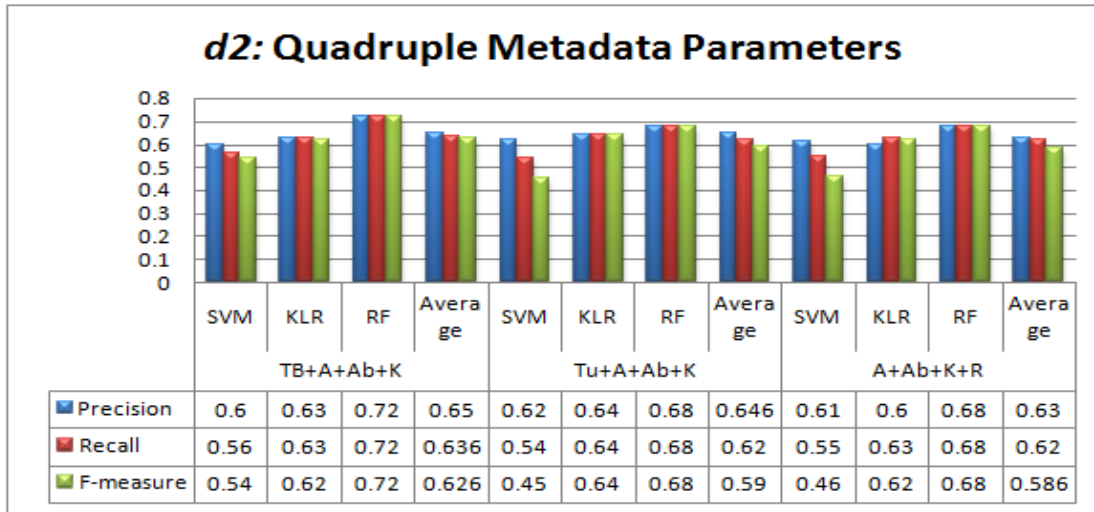


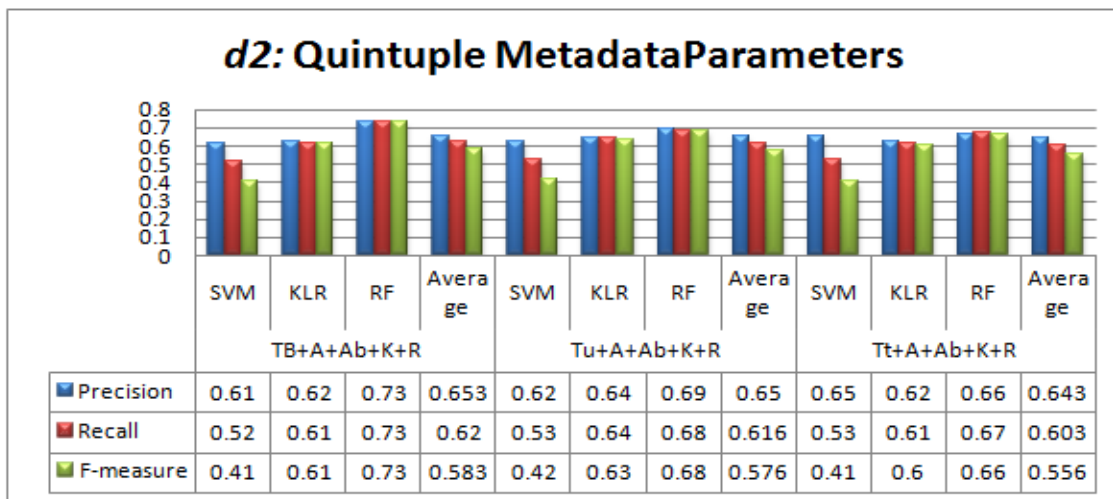
Figure 4-11 PRF Bar Chart for *d1* Quadruple Parameters



**Figure 4-12 PRF Bar Chart for d2 Quadruple Parameters**

#### 4.8.5 Quintuple Metadata Parameters

As *d2* contains “Keywords” parameter, therefore it has more metadata combinations than *d1*. In case of quintuple metadata parameters for *d2*, the combination “*Title\_Bigram + Authors + Abstract + Keywords + References*” outperformed other combination with the average Precision of 0.65, Recall of 0.62 and F-measure of 0.58 as shown in figure 4-13. In all parameters combinations, the Random Forest classifier has contributed more in achieving best average value of Precision, Recall and F-measure.

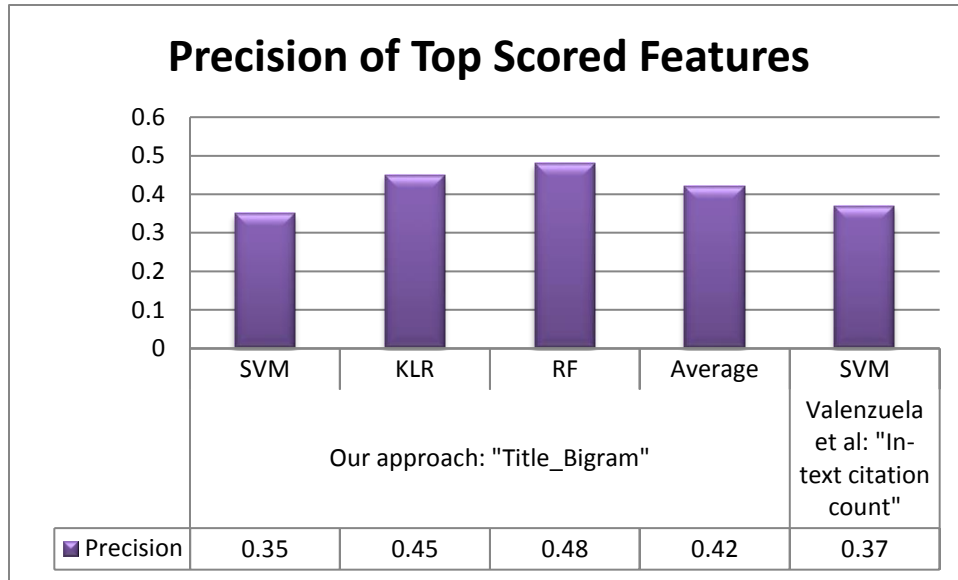


**Figure 4-13 PRF Bar Chart for d2 Quintuple Parameters**



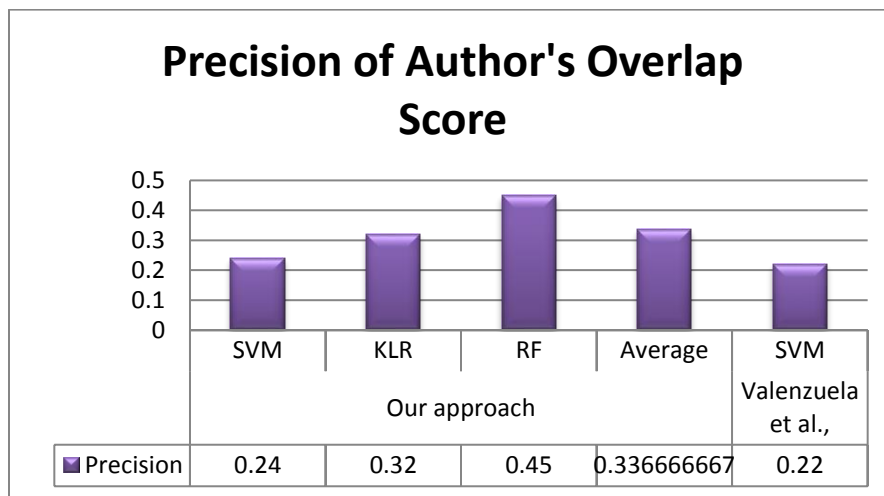
## 4.9 COMPARISONS

In citation classification community, the valenzuela's (Valenzuela, Ha, & Etzioni, 2015) have proposed first approach to tackle the problem of *important* citations identification. They proposed twelve different features from which most of them are relying on the contents of the articles. There can be scenarios when content is not available for example, major journal publishers like IEEE, ACM, Springer Elsevier etc hold financial and legal barriers to provide access to the content of articles. In such cases, there should be an alternative way to solve the problem of important citations identification. This can be done with the help of freely available information about the articles. Such information is available in the form of metadata. Different kinds of useful metadata such as titles, authors, keywords, categories and references are almost freely available and different ways of their exploitation can actually help to narrate useful results. Our approach relies on the metadata of same articles as of Valenzuela's approach. Since their proposed features are different than ours, but we have used their dataset for experimentation so it would be justified to make possible comparisons with their approach. In this chapter, we have considered all the possibilities to compare our results with their results. The similarity ratio of all metadata parameters has been calculated with the help of different proposed formulas and the obtained values are assigned as features for supervised machine learning. Three classifiers Support Vector Machine, Kernel Logistic Regression and Random Forest have been used to classify citations by using 10-fold cross validation. We have evaluated every possibility of metadata parameters combinations by defining the hierarchy of single parameters combinations up to quintuple metadata parameters to discover useful combinations. The average Precision, Recall and F-measure score is calculated by taking the average of PRF values achieved by three classifiers. Their top scored single feature *In-text citation count* obtained Precision of 0.37 and our top scored single parameter *Title\_Bigram* obtained Precision of 0.42. Considering the fact that this Precision is obtained just by exploiting freely available metadata, so the Precision of 0.42 is quite high as compared to Precision of 0.37. The difference can be seen in figure 4-14 more clearly.



**Figure 4-14 Comparisons between Top Scored Features**

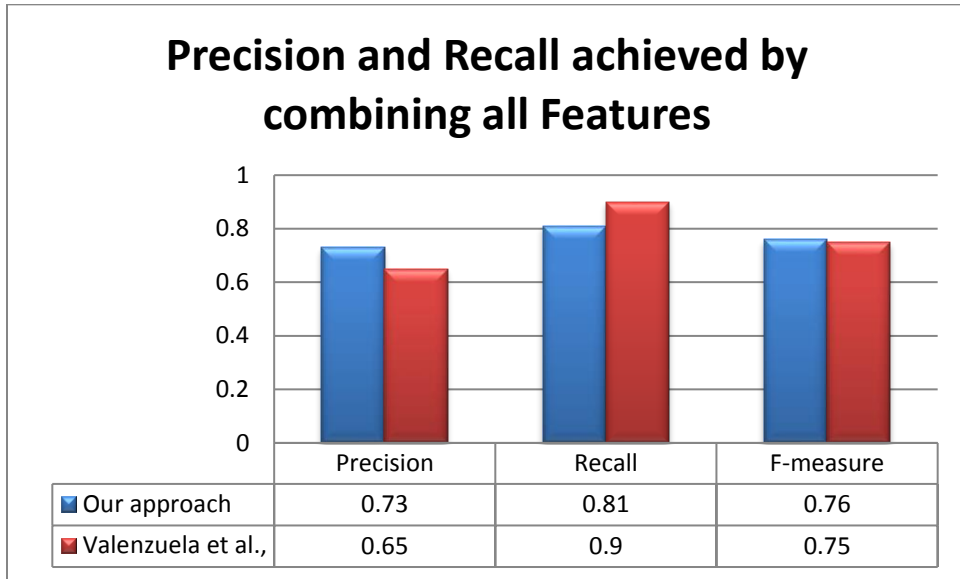
In the case of author's overlap score, their author's overlap score is Boolean; 0 for having no common authors between pairs and 1 for having one or more common authors. Since their scheme assign same weight to more than one common author between pairs. However, our proposed formula does not treat more than one common author equally; it calculates the ratio of all common authors and uncommon authors. Their authors overlap feature obtained Precision of 0.22 and our authors overlap feature obtained Precision of 0.34 which demonstrates that our way of calculating authors overlap score is more significant than theirs. The difference can be seen in figure 4-15 more clearly.



**Figure 4-15 Comparisons between Author's Overlap Score**

;

Their system achieved average Precision of 0.65 by combining all features. We have presented the Precision of each classifier individually and at the end by taking average of them. Our system achieved Precision of 0.68 against SVM, Precision of 0.69 against KLR, Precision of 0.82 against Random Forest and average Precision of all classifiers is 0.73. Hence, this result is *Important* as our system achieved improved Precision just by relying on freely available metadata. The results are envisioned in figure 4-16.



**Figure 4-16 Comparisons between Results**

## Chapter 5

### CONCLUSION AND FUTURE WORK

#### 5.1 CONCLUSION

The citation analysis is used to formulate different scientific policies such as to measure academic influence of a journal or a researcher and identification of evolving research topics etc. In literature, researchers have questioned the reliability of results achieved by using citation count approach. They critically reviewed citations of different articles and argued that all citations are not equal, some can be important and some can be non-important. The importance of citations can be measured by considering the reasons of citations discovered by numerous researchers. Those reasons are providing background knowledge, using or extending existing work etc. In citation classification community, the most recent approaches focus on merging different reasons into two categories (1) Important and (2) Non-important. This binary citation classification (i.e., important and non-important) can ensure the reliability of citation count approach via counting only those citations which are important. Moreover, the identification of important citations can help to have an idea about the emerging research trends and to get the articles which are closely related to the state-of-the art approaches for research work.

We have comprehensively studied more than 40 research articles to critically review state-of-the-art approaches. In citation classification community, researches have proposed different techniques relying on the content of articles. They proposed different features such as in-text citation count, cue words/phrases, authors overlap etc. These approaches have their own limitations as in case of cue words, the list of cue-phrases or cue-words needed to be updated manually for every dataset and high in-text citation count of cited paper in citing paper does not guarantee that the particular cited paper is an important citation etc as discussed with the help of case studies in chapter 2.

The limitations discussed above are the limitations of content based approaches. As per our knowledge, all the existing schemes are content dependent. There should be an alternative approach to classify citations when the content is not available. Sometimes we do not have an open access to get full articles as there are technical, financial and legal barriers. On the other hand, various kinds of useful metadata are almost freely available such as title, authors, keywords categories and references etc. The similarity between metadata of citing and cited

paper can help us to capture the type of relation (i.e., Important or Non-important) between citing and cited paper. To address the issue of important citations identification we have proposed a comprehensive methodology to analyze whether exploitation of metadata can help us to achieve the result closer or better than content based approach.

We have proposed different formulas to obtain ratio of similarity between metadata paper-citation pairs. Two benchmark datasets have been used for experiments. One is taken by the recently published paper in 2015 by Valenzuela et al, published in AAAI workshop on Scholarly Big Data (referred as *d1* in this thesis). The second one is personally collected and annotated by the actual authors of the citing and cited articles from CUST Computer Science faculty members (referred as *d2* in this thesis). Further pre-processing is done on datasets where we removed stop words from titles, stemming of titles, keywords, synonyms and Grow bag datasets and extraction of titles from references.

The score against each formula is calculated and assigned as a feature for supervised machine learning for a binary classification. The classification is performed by using state-of-the-art classifiers which are being used in such research works like: SVM, KLR and Random Forest classifier. We have combined metadata parameters into 5 levels to identify which combinations help to achieve high Precision. We have evaluated the results achieved against each classifier and at the end by taking average of them as we want to analyze the every possibility of achieving the best results.

First level is of single metadata parameters combinations in which the performance of each metadata parameter is analyzed. In both datasets the top scored metadata parameters is *Title\_Bigram* which achieved Precision of 0.42 and 0.54 for *d1* and *d2* respectively. The similar behavior of bigram in both datasets show that Bigram terms similarity between titles hold a potential to identify important citations. In the second level, two metadata parameters are combined, since the *Title\_bigram* has achieved highest value of Precision individually, therefore, it is almost found the top scored combination in every level. The combination of *Title\_Bigram* and *Bibliographically Coupled References* outperformed for *d2* with Precision 0.52 , and *Title\_Bigram* and *Authors* with Precision 0.61 for *d2*. In the third level the *Title\_Bigram* + *Authors* + *Bibliographically Coupled References* with average Precision of 0.59 for *d1* and “*Title\_Bigram* + *Authors* + *Abstract*” with Precision 0.61 for *d2*. In fourth level, *Title\_Bigram* + *Authors* + *Abstract* + *Bibliographically Coupled References*” for *d1* with average Precision of

0.68 and “Title Bigram + Authors + Abstract + Keywords for  $d2$  with average Precision of 0.61. In the fifth level the *Title\_Bigram + Authors + Abstract + Keywords + References* for  $d2$  with average Precision 0.67. The important finding of our results is the identification of important parameter bigram terms similarity, and keywords similarity achieved average Precision of 0.47, this is the important results because this Precision is achieved just by having 53% keywords in datasets.

We have compared our results with the Valenzuela et al., (Valenzuela, Ha, & Etzioni, 2015) results which are content based approach. Their top scored feature in-text citation count achieved Precision of 0.37 and our top scored feature Title\_Bigram achieved Precision of 0.42. Considering the fact that this best feature is from content of the article and our best feature is based on metadata of the article, so the Precision of 0.42 is quite higher than 0.37. Our one feature is author overlap, since their author’s overlap score is Boolean, 0 for having no common authors between pairs and 1 for having at least one common author. Since their scheme assign same weight to more than one common authors between pairs, While our proposed formula do not treat more than one common authors equally, it calculates the ratio of all common authors and un common authors. Their authors overlap feature obtained precision of 0.22 and our authors overlap feature obtained average precision of 0.34 which show that our way of calculating authors overlap score is more significant than theirs. Their precision achieved by combining the entire features is 0.65 and our average precision is 0.73. The best performed classifier is Random Forest for both  $d1$  and  $d2$ , we have used default configurations setting in WEKA in which 10 trees are build having maximum depth of the trees 0. While the time taken to build the model for  $d1$  is 0.37 seconds and 0.39 seconds for  $d2$ . The overall finding of this thesis is that in the scenarios when content is not available; the similar behavior between content and metadata based approach and in some cases better than content based approaches can be seen.

## **5.2 FUTURE WORK**

In this thesis, we have used the dataset of Valenzuela’s approach. Currently, this is the standard dataset from state-of-the-art approaches which is freely available. This dataset contains only 465 annotated paper-citation pairs. Since this is the small amount of data to draw a generic conclusion. In this domain, there is lack of availability of large annotated dataset. So the one future direction could be the generation of large gold standard dataset covering different

domains, having the different geographical locations of authors and covering maximum amount of metadata, only then we can analyze the behavior of metadata in citation classification. The second future direction could be the combination of top scored features of both metadata and content based approaches.

## BIBLIOGRAPHY

Abu-Jbara, A., & Radev, D. (2011). Coherent citation-based summarization of scientific papers. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 1*, pp. 500-509. Stroudsburg, PA, USA: Association for Computational Linguistics.

Agarwal, S., Choubey, L., & Yu, H. (2010). Automatically classifying the role of citations in biomedical articles. *In Proceedings of American Medical Informatics Association Fall Symposium (AMIA). 2010*, pp. 11-15. Washington, DC : American Medical Informatics Association.

Anderson, R., Narin, F., & McAllister, P. (1978). Publication ratings versus peer ratings of universities. *Journal of the American Society for Information Science* , 29 (2), 91-10.

Balaban, A. T. (2012). Positive and negative aspects of citation indices and journal impact factors. *Scientometrics* , 92 (2), 241-247.

Benedictus, R., Miedema, F., & Ferguson, M. (2016). Fewer numbers,better science. *Nature* , 538 (7626), 453-455.

Bonzi. (1982). Characteristics of a literature as predictors of relatedness between cited and citing works. *Journal of the American Society for Information Science* , 33 (4), 208–216.

Bonzi, S., & Snyder, H. (1991). A comparison of self citation and citation to others. *Scientometrics* , 21 (2), 245-254.

Brooks, T. (1985). Private acts and public objects: An investigation of citer motivations. *Journal of the American Society for Information Science* , 6 (4), 223-229.

Case, D. O., & Higgins, G. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science* , 51 (7), 635-645.

Chubin, D. E., & Moitra., S. D. (1975). Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science* , 5 (4), 423-441.

Ciancarini, P., Iorio, A. D., Nuzzolese, A. G., Peroni, S., & Vitali, F. (2013). Semantic annotation of scholarly documents and citations. *In Congress of the Italian Association for Artificial Intelligence* (pp. 336-347). Springer International Publishing.

Diederich, J., & Balke, W. T. (2007). The semantic growbag algorithm: Automatically deriving categorization systems. . *In International Conference on Theory and Practice of Digital Libraries* (pp. 1-13). Berlin: Springer Berlin Heidelberg.

Ding, Y. (2011). Applying weighted pagerank to author citation networks. *Journal of the American Society for Information Science and Technology* , 62 (2), 236-245.



- Dong, C., & Sch afer, U. (2011). Ensemble-style self-training on citation classification. *In Proceedings of 5th International Joint Conference on Natural Language Processing*, (pp. 623–631). Chiang Mai, Thailand: Asian Federation of Natural Language Processing.
- Finney, B. (1979). The reference characteristics of scientific texts. Master's thesis. London: The City University of London.
- Frost, C. (1979). The use of citations in literary research: a preliminary classification of citation functions. *Library Quarterly*, 49 (4), 399-414.
- Garfield, E. (1965). Can citation indexing be automated. *In Statistical association methods for mechanized documentation, symposium proceedings*. 269, pp. 189-192. Washington, DC: National Bureau of Standards, Miscellaneous Publication 269.
- Garfield, E., & Merton, R. K. (1979). *Citation indexing: Its theory and application in science, technology, and humanities* (Vol. 8). New York: Wiley.
- Garfield, E., Sher, I. H., & Torpie, R. J. (1964). The use of citation data in writing the history of science. *INSTITUTE FOR SCIENTIFIC INFORMATION INC PHILADELPHIA PA.*
- Garzone, M., & Mercer, R. (2000). Towards an automated citation classifier. *In Conference of the Canadian Society for Computational Studies of Intelligence* (pp. 346-337). Springer Berlin Heidelberg.
- Giles, L. C., Bollacker, K., & Lawrence, S. (1998). CiteSeer: An automatic citation indexing system. *In Proceedings of the third ACM conference on Digital libraries* (pp. 88-98). ACM.
- Hou, W. R., Li, M., & Niu, D. K. (2011). Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution. *BioEssays*, 33 (10), 724-727.
- Inhaber, H., & Przednowek, K. (1976). Quality of research and the Nobel prizes. *Social Studies of Science*, 6 (1), 33-50.
- Jochim, C., & Sch tze, H. (2012). Towards a generic and flexible citation classifier based on a faceted classification scheme. *In Proceedings of COLING'12* (pp. 1343–1358). Mumbai, India: COLING'12.
- Li, X., He, Y., Meyers, A., & Grishman, R. (2013). Towards fine-grained citation function classification. *In Proceedings of Recent Advances in Natural Language Processing*, (pp. 402–407). Hissar, Bulgaria.
- Liakata, M., Saha, S., Dobnik, S., Batchelor, C., & Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28 (7), 991-1000.
- Locke, W. N. (1956). Machine translation of languages. *American Documentation*, 7 (2), 135.
- Meyers, A. (2013). Contrasting and corroborating citations in journal articles. *In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP* (pp. 460-466). Hissar, Bulgaria: RANLP.

- Moravcsik, J. M., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science* , 5 (1), 88–91.
- Myers, C. R. (1970). Journal citations and scientific eminence in contemporary psychology. *American Psychologist* , 25 (11), 1041.
- Nanba, H., & Okumura, M. (1999). Towards multi-paper summarization using reference information. *In IJCI*, 99, pp. 926-931.
- Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Washington, D. C: Computer Horizons.
- Oppenheim, C., & Renn., S. P. (1978). Cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information* , 29 (5), 227-231.
- Paice, C. D. (1981). The automatic generation of literary abstracts: an approach based on the identification of self indicating phrases. In R. Oddy, C. S. Robertson, & v. Rijsbergen (Ed.), *In Proceedings of the 3rd annual ACM conference on Research and development in information retrieval* (pp. 172-191). Butterworth & Co.
- Pham, S., & Hoffmann, A. (2003). A new approach for scientific citation classification using cue phrases. *In Australasian Joint Conference on Artificial Intelligence* (pp. 759-771). Berlin: Springer.
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program* , 14 (3), 130-137.
- Radoulov, R. (2008). Exploring automatic citation classification. Waterloo: University of Waterloo
- Shahid, A., Afzal, M. T., & Qadir, M. A. (2011). Discovering Semantic Relatedness between Scientific Articles through Citation. *Australian Journal of Basic and Applied Sciences* , 5 (6), 1599-1604.
- Smith, A. T., & Eysenck, M. (2002). *The correlation between RAE ratings and citation counts in psychology*. University of London, Royal Holloway.
- Smith, L. C. (1981). Citation Analysis. *Library trends* , 30 (1), 83-106.
- Spiegel-Rusing, I. (1977). Science studies: Bibliometric and content analysis. *Social Studies of Science* , 7 (1), 97-113.
- Swales, J. (1990). *Genre Analysis: English in academic and research settings*. Cambridge University Press.
- Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. *In Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 103-110). Association for Computational Linguistics.
- Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying Meaningful Citations. *Workshops at the*

*Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI.*

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S. H., Jones, R., et al. (2015). *The metric tide: report of the independent review of the role of metrics in research assessment and management*. Publisher Full Text.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66 (2), 408-427.

Ziman, J. M. (1968). *Public knowledge: An essay concerning the social dimension of science* (Vol. 519). CUP Archive.