Scientometrics (2019) 118:21–43
https://doi.org/10.1007/s11192-018-2961-x

Identification of important citations by exploiting research articles' metadata and cue-terms from content

Faiza Qayyum[1] · **Muhammad Tanvir Afzal**[1]

Received: 23 October 2017 / Published online: 22 November 2018

Abstract
Citations play a pivotal role in indicating various aspects of scientific literature. Quantitative citation analysis approaches have been used over the decades to measure the impact factor of journals, to rank researchers or institutions, to discover evolving research topics etc. Researchers doubted the pure quantitative citation analysis approaches and argued that all citations are not equally important; citation reasons must be considered while counting. In the recent past, researchers have focused on identifying important citation reasons by classifying them into *important* and *non-important* classes rather than individually classifying each reason. Most of contemporary citation classification techniques either rely on full content of articles, or they are dominated by content based features. However, most of the time content is not freely available as various journal publishers do not provide open access to articles. This paper presents a binary citation classification scheme, which is dominated by metadata based parameters. The study demonstrates the significance of metadata and content based parameters in varying scenarios. The experiments are performed on two annotated data sets, which are evaluated by employing SVM, KLR, Random Forest machine learning classifiers. The results are compared with the contemporary study that has performed similar classification employing rich list of content-based features. The results of comparisons revealed that the proposed model has attained improved value of precision (i.e., 0.68) just by relying on freely available metadata. We claim that the proposed approach can serve as the best alternative in the scenarios wherein content in unavailable.

Keywords Citation classification · Metadata · Information retrieval · Support vector machine · Kernel logistic regression · Random forest

✉ Faiza Qayyum
faizaqayyum@cust.edu.pk

Muhammad Tanvir Afzal
mafzal@cust.edu.pk

[1] Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan

Introduction

Researchers always conduct research by relying on the legendary work of eminent predecessors in the field. The statement is justified further by Ziman indicating that *"a scientific paper does not stand alone; it is embedded in the literature of a subject"* (Ziman 1968, p. 58). Whenever researchers discuss someone's work in their article, they always acknowledge that in the references section, this acknowledgement is known as a *citation* (Narin 1976). Ziman (1968) has described 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. Citations are reckoned as a substantial measure to analyze multifarious aspects of individuals or institutions, such as scrutinizing the academic influence of authors or institutions over the scientific community. The citation based measures are being utilized in formulation of different academic policies such as *awards and Nobel Prizes allocation* (Inhaber and Przednowek 1976), *research funds allocation* (Anderson et al. 1978), *peer judgements* (Smith and Eysenck 2002), *ranking of researchers* (Hirsch 2005; Raheel et al. 2018; Ayaz and Afzal 2016), *ranking of countries* (Mazloumian et al. 2013), and so on.

In the late 90s, identification of citation reasons by Garfield (Garfield 1965) opened extensive dimensions of research towards the scrutiny of citation behavior (Spiegel-Rusing 1977; Bornmann and Daniel 2008). Researchers argued that each citation serves different purpose, but all are treated equally in citation count approaches (Moravcsik and Murugesan 1975). Until now, studies regarding usage of pure quantitative citation analysis for different purposes (i.e., author ranking etc.) are deemed as hot topic of interest (Raheel et al. 2018; Ayaz and Afzal 2016). The study Benedictus et al. (2016) has examined the role of citation count and concluded that quantity wins over quality when pure citation count based measures are considered to analyze different factors. Researchers have recommended certain improvements to scale up the quality and weigh down the quantity (Wilsdon et al. 2015; MacRoberts and MacRoberts 2018). Various researchers have argued that the reason of citations must be considered to determine quality of someone's work (Teufel et al. 2006; Valenzuela et al. 2015; Zhu et al. 2015).

Can we automatically differentiate between citation reasons? The old citations annotation approaches work manually by interviewing the citer, sometimes after publication of article, to recall why he cited the work (Brooks 1985); or by interviewing the scholars at the time of writing the article that why they are citing the particular work (Case and Higgins 2000). Finney (1979) suggested an idea in her master's thesis that citation classification process can be automated. Her idea has been adopted by different researchers to classify citations into different categories (Garzone and Mercer 2000; Teufel et al. 2006). However, most of these studies have classified citations into multiple vague reasons that cannot precisely serve the purpose of overcoming the said limitation of mere citation count approaches. Gradually, scientific community paid attention on receding the number of citation classes; currently, the community has concurred on classifying citation into two broad categories *important and non-important* classes (Valenzuela et al. 2015; Zhu et al. 2015). This paper also classifies citations into *important* and *non-important* categories; however, by using different set of features than contemporary approaches.

What do we mean by important and non-important classes? Generally, while writing a research article, an author knows that among list of cited references, only a few of them have influenced more to the citing study than other references. But how to depict this *influence* clearly? According to Zhu et al. (2015) an influential research study is the one that

inspires the scholarly community in terms of adoption or extension of the presented idea (Zhu et al. 2015). To clearly understand the meaning of *important* and *non-important* citations, let us contemplate the window of contemporary state-of-the-art citation classification studies. Consider the study (Garzone and Mercer 2000) wherein authors have extended the study of Finney (1979) by implementing her idea of *"associating cue words with citation function and using citation location in the classification algorithm"* (Garzone and Mercer 2000, p. 339) to automatically classify citations. According to the literature (Valenzuela et al. 2015; Zhu et al. 2015), the citation relation between (Finney 1979) and (Garzone and Mercer 2000) is *important.* On the other hand, Garzone and Mercer (2000) have also cited various other studies, such as the study of (Garfield 1965) to provide the background knowledge of the proposed study, i.e., by explaining the no. of citation categories presented by (Garfield 1965). Here the citation relation between (Garfield 1965) and (Garzone and Mercer 2000) is *non-important* as per the concept of literature (Valenzuela et al. 2015; Zhu et al. 2015). Based on above-stated definitions of important and non-important citations, researchers (Valenzuela et al. 2015; Zhu et al. 2015) have classified citations into two broad categories. (1) The category of the citations which are given just to provide background knowledge of the proposed study. This category is denoted by the terms *non-Influential* and *incidental* in (Zhu et al. 2015) and by the terms *non-important* and *incidental* in the study of (Valenzuela et al. 2015). (2) Another category is of the citations that have been inspired by the cited work in context of using or extending the cited work. This category is denoted by the terms Influential in (Zhu et al. 2015) and important by (Valenzuela et al. 2015). We use the term *important* for this category.

Currently, all citation classification schemes are either fully content dependent or are dominated by the content-based features. However, most of the time content is not freely available. Major journal publishers: IEEE, ACM, Springer, Elsevier etc. do not provide open access to articles. In such scenarios, there should be an alternative way to classify citations. One of the best possible substitutes could be the exploitation of freely available metadata. Different kinds of useful metadata such as *title, authors, keywords and references* are almost freely available and hold the potential to identify meaningful citations.

In this paper, we present a model to classify citations into *important* and *non-important* categories. The primary concern of this study is to analyze the extent to which metadata can behave similar to content based features. Most of the parameters in proposed study involve metadata. From the family of content-based features, we have only opted for few parameters that could identify important citations. The content-based features include abstract and cue-terms from citing sentences. The *abstract* is part of a content but most of the time it is also accessible similar to metadata. Another content-based feature is the cue-terms that reside in the sentence wherein the particular citation has been made in body of the paper (i.e., in-text citation). The sentence is also referred as citing sentence (Jeong et al. 2014). The cue-terms can hint the class of a citation (i.e., important and non-important) in a static nature (i.e., un-matching/matching of cue-terms is not domain-dependent). Therefore, we have decided to incorporate this parameter after critical scrutiny of available content-based parameters.

Diversified information seeking behaviors have been presented in the extant literature (Krikelas 1983; Ellis 1993; Mai 2016). The study of Ellis (1993) has analyzed different studies that employ quantitative and qualitative measures for information seeking. He narrated six main features of information seeking pattern such as *starting, chaining, browsing, differentiating, monitoring, and extracting*. Chaining is a process of seeking information from bibliographies of a material. Our proposed study can assist the scholarly community in terms of seeking information through *chaining*. For instance, consider a scenario

wherein a researcher has some research papers (i.e., a source paper) and he/she intend to seek the research articles of the same nature to conduct a literature survey. In this regard, one of the best possible sources would be to follow the chaining process, i.e., exploiting the citations/bibliographies of source paper. Now the exploitation of bibliography requires cognitive effort to discern only important citations *(definition of important citations has been explained this section).* An efficient method to track the important citations can make the literature seeking process much efficient. In this research, we intend to scrutinize the behavior of our hybrid metadata parameters to identify the important citations.

The pioneer approach towards important citations identification was proposed by (Valenzuela et al. 2015) which is the combination of content and metadata based features. Our work is near to (Valenzuela et al. 2015) approach (i.e., binary citation classification) due to following common factors: (1) *we have picked metadata and abstract information of the same articles from which they have extracted content and metadata based features,* and (2) *we have employed their annotated data set to evaluate and compare the proposed model.* The results of comparison demonstrate that our system has attained improved value of precision when all features have been combined (0.68 vs. 0.65). However, the value of recall is lower than their recall value (0.90 vs. 0.70), but still a significant one as it has been obtained relying on freely available information (i.e., metadata). The cue-terms parameter alone has outperformed all other parameters when all have been evaluated individually. Beside this, comparison results of same metadata parameters between both approaches also signify the potential of the proposed scheme. We claim that the proposed scheme can identify important citations dominantly by the metadata parameters.

Literature review

Manual citation classification

Dating back from late 90s to date, plethora of studies have been conducted that employ pure count of citations to conduct different sort of bibliometrics analysis such as citation indexing systems (Giles et al. 1998; Lawrence et al. 1999), formulation of different academic policies such as awards and Nobel prizes allocation (Inhaber and Przednowek 1976), ranking researchers (Hirsch 2005; Raheel et al. 2018; Ayaz and Afzal 2016) ranking countries (Mazloumian et al. 2013) etc. There could be multiple reasons behind citing a particular study. Garfield (1965) was the pioneer who analyzed the citation behavior and listed 15 citation reasons by analyzing the location of text in the paper, scrutinizing the differences and patterns. The reasons include (1) paying homage to pioneers (2) providing background knowledge (3) extending the work etc. Afterwards, (Liptez 1965) presented a similar study mentioning different classes of citations. However, both of the studies have not presented any statistical measure, rather they have just narrated the concept theoretically (Bornmann and Daniel 2008). Nonetheless, these studies have served as a foundation for scholarly community to perform empirical investigation for citation reasons identification. Thereafter, various other studies have also focused on capturing the citation behavior (Oppenheim and Renn 1978; Spiegel-Rosing 1977). Almost all the citation analysis based studies provide equal importance to all the citations regardless of their diverging behavior. From past several decades, there has been an extensive debate regarding usage of pure citation count; researchers have argued that all citations are not equal and therefore should be treated as

per their importance (Bonzi 1982; Ziman 1986; Bornmann and Daniel 2008; Teufel et al. 2006; Zhu et al. 2015; Valenzuela et al. 2015). According to (Zhu et al. 2015) if incidental citations are filtered out from citation count, then it could positively contribute towards enhancing the scope of mere quantitative citation analysis based studies. Moreover, having list of only important citations *(the meaning of important and non-important has already been explained in Sect. 1)* can also help scholars to find influential studies pertaining to the particular topic for the sake of reviewing the contemporary state-of-the-art. Till late mid-90′s, the process of citation reasons identification was restricted to manual investigation. For instance, the scholars were interviewed at the time of writing an article or after its publication to describe the purpose of citing the particular work (Brooks 1985; Case and Higgins 2000). Though it was not practical to discern the citing behavior through cognitive approaches. Therefore, scientific community pondered about automating the process to tackle the citation reasons. Let us shed a light on some of the prominent automatic citation classification schemes.

Automated citation classification

Finney (1979) coined an idea that the process of citation classification can be automated. She associated the cue words and citation location with the citation function. Her approach was not fully automated, but is pertinent to mention that this study has served as an inspiration for first fully automatic citation classification approach (Garzone and Mercer 2000). However, researchers have hardly acknowledged Finney's approach as the pioneer approach of automatic citation classification, due to argument that it was a Master's thesis, not a published study (Bornaman and Daniel 2016). Garzone and Mercer (2000) claimed that they are among the pioneers of initiating fully automated citation classification scheme. The main theme of the study was inspired by Finney's approach. Authors argued that Finney's idea does not cover all aspects of being cited. They incorporated those aspects and classified citations into 35 categories. The classification was done by forming 195 lexical matching rules and 14 parsing rules, which were based on cue words and section location of the citation. The data set contained 11 physics and 9 biochemistry articles. The system attained good results on seen articles and average results on unseen articles. However, the number of classes are so large in number that most of them are conflicting. Pham and Hoffman (2003) considered 482 citation contexts and classified citations into four categories using cue-phrases based Ripple Down Rules (the "RDR") hierarchy. They employed 150 citation contexts for classification and claimed 95.2% accuracy. Similarly, Teufel et al. (2006) proposed a supervised learning approach to classify citations into four categories. Their study is an inspiration of (Spiegel-Rusing 1977) scheme. The citations were differentiated on the basis of cue-phrases based linguistic rules formed using 26 articles and their 548 citations. The studies of (Abu-Jbara and Radev 2011; Dong and Sch¨afer 2011; Jochim and Schütze 2012; Li et al. 2013) have determined the sentiments of citing behavior by analyzing the polarity (i.e., positive and negative) of citations employing content based features including, citation count, cue-phrases etc.

Critical analysis of contemporary approaches

However, all of these studies solely rely on linguistic features of content following the identical norms of cue-phrases exploitation. On contrary, our study introduces exclusively distinct

set of features (i.e., potential metadata based features). Furthermore, it is noteworthy that with the passage of time, the number of citation categories has receded. Authors have focused more on the richness of features rather than classifying citations into vague or conflicting categories. CiTO (Peroni and Shotton 2012; Shotton 2010) identified 90 semantic relations between papers and citations. According to Zhu et al. (2015), the best categorization of these relations would be their division into two broad classes, (1) important and (2) non-important.

Our work is closest to Zhu et al. (2015) and Valenzuela et al. (2015). These studies have performed similar binary citation classification following the same meaning of important and non-important citations as the proposed study. Zhu et al. (2015) classified citations into two categories (1) Influential and (2) Non-influential. The classification was done using (1) In-text count based (2) Similarity based (3) Context based (4) Position-based and (5) Miscellaneous features. The main idea behind this technique was to identify those references that have an academic influence to the citing paper. The authors defined the academic influence as a reference from which the idea, problem, method, or experiment has been adopted. Total 3143 paper-reference pairs were generated from 100 papers, extracted from ACL anthology. They annotated these pairs from the authors of citing papers. Valenzuela et al. (2015) presented a supervised classification approach to identify important and non-important citations. They extracted 465 paper-citation pairs from ACL anthology. These pairs were annotated as important and non-important by two domain experts having 93.9% inter-annotator agreement. The classification was done using 12 features including, total number of direct citations, number of direct citations per section, total number of indirect citations, number of indirect citations per section, author overlap and so on. These features were trained on SVM and Random Forest classifier and the model attained 0.65 precision and 0.90 recall.

Valenzuela et al. (2015), criticizes Zhu's approach and argues that labeling the citations form actual citing authors could cause the biased annotation. To validate this argument, our study employs two data sets. One data set (D1), is taken from Valenzuela et al. (2015) which is annotated by two domain experts. And another data set (D2), which we collected and get it labelled form the actual citing authors following the assumption of Zhu et al. (2015) that actual authors of the papers are in the best position to label their citing work. The evaluation results obtained against both data set can better provide an idea regarding best annotation method.

By a large, number of features in aforementioned studies exploit the content of research articles. Most of them solely rely on the linguistic features from content, following the typical method of cue-phrases exploitation. We argue that scope of these studies gets limited in the scenarios wherein content in unavailable. Major journal publishers such as IEEE, Springer, Elsevier etc. do not provide open access to articles. In this regard, our study is dominated by metadata-based features. The metadata of research papers like title, authors, keywords etc. are almost freely available. Though we agree with the fact that metadata might not be as strong candidate as content, but we assume that it could serve as the best alternative in the scenarios wherein content is unavailable. We further compare our results with the study of Valenzuela et al. (2015) which employs most of the content based features. The results demonstrate that metadata holds potential of obtaining accuracy closer to the content based measures.

The Table 1 concretely recapitulates the existing state-of-the-art approaches.

Table 1 List of contemporary citation classification techniques

Sr #	Authors name	Classes	Accuracy	Content-based features	Metadata-based features
1	Finney (1979)	(1) Background knowledge (2) Tentative references (3) Methodological references (4) Conformational references (5) Negational references (6) Interpretational references (7) Future research references	–	✓	✗
2	Garzone and Mercer (2000)	(1) Negational (2) Affirmational (3) Assumptive (4) Tentative (5) Methodological (6) Interpretational (7) Developmental (8) Future research (9) Use of conceptual, (10) Contrastive, and 25 more	Good results for seen articles and average results for unseen articles	✓	✗
3	Pham and Hoffmann (2003)	(1) Basic (2) Support (3) Limitation (4) Comparison	–	✓	✗
4	Teufel et al. (2006)	(1) Neutral (2) Weakness (3) Comparisons (2) Compatibility	F-measure = 0.71	✓	✗
6	(Dong and Sch¨afer 2011)	(1) Negative (2) Positive (3) Neutral	F-measure = 0.66	✓	✗
7	Jochim and Schütze (2012)	(1) Negative (2) Positive	F-measure = 0.68	✓	✗
8	(Abu-Jbara 2013)	(1) Negative (2) Positive	F-measure = 0.68	✓	✗
9	Meyers (2013)	(1) Corroborate an (2) Contrast	67% Recall for contrast category and 83% for corroborate category.	✓	✗
10	(Li et al. 2013)	(4) Negative (5) Positive (6) Neutral	F-measure = 0.67	✓	✗
11	Zhu et al. (2015)	(1) Influential and (2) Incidental	Precision = 0.35 for countsInPaper_whole	✓	✓
12	Valenzuela et al. (2015)	(1) Important and (2) Incidental	Precision = 0.65 Recall = 0.90	✓	✓

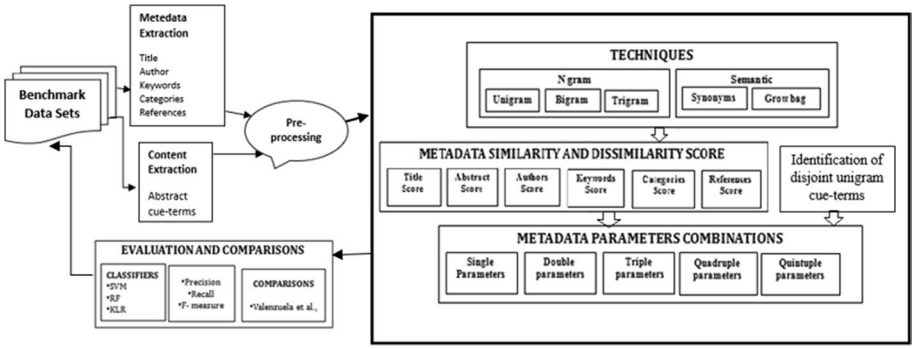


Fig. 1 Context diagram of proposed system

Approach

To the best of our knowledge, no citation classification technique exists that predominantly relies on metadata parameters. The existing schemes (Valenzuela et al. 2015; Zhu et al. 2015) are hybrid approaches that employ most of the parameters from content of research articles. On the contrary, our hybrid scheme contains five metadata based parameters and only two content based parameters.

The proposed study focuses on addressing the following questions:

a. *To what extent the similarities and dissimilarities between metadata parameters can serve as useful indicators for important citation tracking?*
b. *Which metadata parameters or their combinations are helpful in achieving good results?*
c. *Whether our approach can behave closer to the content based approach?*

Figure 1 presents the overall flow of proposed system.

Data set

To classify citations into *important* and *non-important* categories, two data sets *(D*1 and *D*2*)* are employed by considering different factors, which are delineated below.

D1 (Data set1) The classification should be performed on some standard data set of the required nature. By the term 'standard' we mean the data set that has been employed previously in research study that has focused on binary citation classification and is published in some well reputed journal or a conference (so that the reliability of data can be ensured). In this regard, we have employed the data set collected by Valenzuela et al. which uses corpus of 20,527 papers along with their citation graph from the ACL anthology. There are 106,509 citations between these papers. From this collection, they have labeled 465 paper-citation (i.e., cited paper- citing paper) pairs as *important* and *incidental* from two domain experts. To the best of our knowledge, this is the only annotated data of similar nature (i.e.,

automatic binary citation classification) which is publicly available. This data set contains only 14.6% *important* paper-citation pairs.

*D*2 *(Data set2)* Since *D*1 contains only 465 paper-citation pairs that might not be sufficient to analyze the outcomes. The overall conclusion can be drawn more accurately by analyzing the performance behavior between different data sets. Therefore, we have built another data set (referred as *D*2) by considering the best possible source and annotators (i.e., citing authors). *D*2 comprises of 488 paper-citation pairs, which are formed by considering the research papers of Computer Science faculty members from *Capital University of Science and Technology (CUST), Islamabad.* These faculty members are associated with different disciplines of Computer Science such as, Networks, Database, Semantic Web, Information Retrieval, and Software Testing. We have picked two of their research papers and paired them with all of their references that formed total 488 paper-citation pairs. We have explained the same definition of *important* and *non-important* citations to them, that has already been defined by the scientific community (Valenzuela et al. 2015; Zhu et al. 2015). To ensure the accurate annotation of pairs, we have provided them abstract and keywords so that they can accurately recall their citing work. We asked them to label the pair by providing the score of 1 for the references they think *important/influential* and score of 0 for *non-important/incidental* citations. This annotation has formed only 18.4% *important* paper-citation pairs.

Synonyms and growbag

Usually, two persons use two different words for a same thing, which could be synonyms of each other (e.g. distributed and dispersed etc.). On the other hand, some words are strongly related to each other but they are not synonyms of each other (e.g. semantic web and RDF). In this study, Synonyms and Growbag technique is utilized for maximum matching of titles and keywords to examine which pairs hold more similarity. The synonyms of titles and keywords terms are extracted using WordNet[1] library. To match the strongly related terms, the collection of 0.5 million first ordered co-occurrence DBLP Growbag keywords/terms produced by (Diederich and Balke 2007) is utilized.

Parameters

Two types of parameters are employed in this study: (1) Metadata based parameters and (2) Content based parameters. The metadata parameters include, *Titles, Authors name, Keywords, Categories*, and *References* and content based parameters include *(1) Abstract and (2) Cue-phrases.* We have picked these parameters due to following assumptions:

(1) More chance of similarity between titles of *important* pairs.
(2) Common authors between pair increase the chance of being an important citation of the cited paper.
(3) More similar keywords are more likely to be present in *important* pairs than in *incidental* ones.
(4) Different publishers that publish articles related to Computer Science domain uses the ACM categorization system to assign suited category to the article. Similar to keywords, we believe more similarity exists between categories of *important* pairs than in *incidental* ones.

[1] https://wordnet.princeton.edu/.

(5) Most of the time, most relevant papers cite same work in their references section. So frequently the references of citing and cited papers are matched, the chance of being its *important* paper-citation pair increases.
(6) Similar to other freely available metadata parameters, the *abstract* of most of the research papers is freely available and similarity between abstract of two papers could increase the chance of citations being important.
(7) Although the primary objective of this study is to identify the potential of metadata based parameters in the scenarios wherein content is unavailable. However, if one intends to identify important citations from interdisciplinary domain then there could be less similarity between metadata parameters. Contemplating this issue, we have incorporated unigram cue-phrases from citing sentences of pairs. Further detail regarding cue-phrases is presented in the following section.

Since *D*1 does not contain keywords and categories, therefore, its parameter list includes *title, authors, abstract and references.*

Data availability

For *D*1, 432 pairs out of 465 are found from ACL anthology from which 13.2% are *important.* While building the pairs in *D*2, we have considered only those citations which are research articles. All the citations other than research articles like URL, tool information etc. have been skipped and experiments are performed on remaining 324 pairs from which 14% of the pairs are *important*. All the metadata for *D*2 are extracted successfully, however, the abstract of the paper having ID: *P98-1106* was not found because this article does not contain the abstract. For *D*2, 98.7% of the abstracts, 58.3% of the Keywords, 4.3% of the Categories, 93.2% of the References and 100% of the Titles and 100% of the Authors are extracted successfully. The category parameter has been skipped from experiments due to its less availability (i.e., 4.3%). 3% out of 4.3% categories were matched between *important* pairs but still we cannot draw a generic conclusion based on such small amount.

The following necessary pre-processing steps are performed to calculate the parameters scores for supervised machine learning.

1. The stop words are removed from titles by using *Onix Text Retrieval Toolkit.*[2]
2. The terms of titles, keywords and Growbag data set are converted into their root terms by using porter stemmer algorithm (Porter 1980), in R by importing snowball library. This step is necessary in order to have better precision.
3. The terms of titles are split into unigram, bigram and trigram by using NLP library in R.
4. The unmatched n-gram terms of titles and terms of keywords are matched with their synonym terms and then with the Growbag terms. Although, only 3 synonym terms of *D*1 titles are acquired from WordNet library, therefore, the synonyms matching scheme has been skipped from our experiments. This could be due to the fact that WordNet library contains synonyms of routinely used English language words, not specific words that belong to the domain of Computer Science.
5. The variation was found in order of author's first and last name in reference of same articles. Therefore, we have preferred to match only titles of all references because

[2] http://www.lextek.com/onix/.

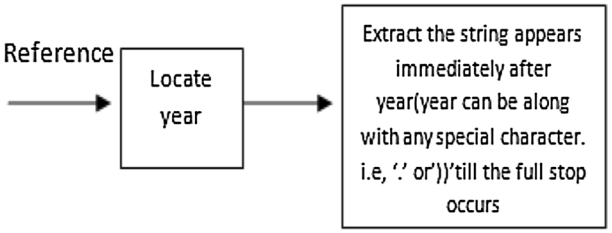


Fig. 2 References' title extraction heuristic

Root_Paper_ID	Cited-by_Paper_ID	Citing_Sentence	Follow-up
A00-1043	C00-2140	Since we only use shallow methods for textual analysis that do not generate a dependency structure, we cannot use complex methods for text reduction as described, e.g., in	0
A00-1043	P02-1057	Sentence simplification systems are capable of compressing long sentences by deleting unimportant words and phrases	0
A97-1011	W06-0202	Three dependency parsers were used for these experiments: MINIPAR the Machinese Syntax parser from Connexor Oy and the Stanford parser	1
A97-1011	P01-1006	The current version of the evaluation workbench employs one of the high performance "super-taggers" for English Conexor's FDG Parser	1

Fig. 3 Extracted citing sentences

we believe every article holds a unique title. The titles of references are extracted by applying heuristic approach described in Fig. 2. To ensure the correction, the extracted titles are manually cross-checked with the titles of references. This heuristic helped us to extract 89% of the titles. The remaining 11% of the titles are extracted manually.

6. Extraction of Cue-phrases: Cue phrases are extracted from the citing sentences. Citing sentences are those sentences in which in-text citation appears in body of the paper (Jeong et al. 2014). The terms utilized in these sentences hold the strong potential to hint about the reasons of citation (Zhu et al. 2015). Since the automatic citation extraction algorithms do not provide adequate accuracy (Shahid et al. 2011). Therefore, we have preferred to extract citing sentences manually. Total of 693 citing sentences are extracted from the pairs in $D1$. The authors' names have been removed from the citing sentences to reduce the noise. An overview of extracted citing sentences is presented in the Fig. 3.

To extract the cue-phrases from citing sentences, following pre-processing steps are performed.

a. *Stop Words Removal* In Natural Language Processing ("NLP"), the frequently appeared words like 'is', 'which', 'and', 'it', 'at' etc. are considered as stop words. In various NLP problems, the existence of stop words causes extra noise which may adversely affect the accuracy of results. In the first pre-processing step, the stop words are removed from citing sentences.
b. *Stemming* In this step, all the terms in citing sentences are converted in their root terms. For instance, the terms cooking, cooked, and cookery get stemmed into a term cook. For this, we have employed PorterStemmer algorithm.

c. *Unigram terms extraction* All the unigram terms from citing sentences are extracted using NLTK library in R.

Similarity and dissimilarity score calculation

Metadata based features

In information retrieval systems, usually researchers only focus on computation of similarity and do not consider dissimilarity which can actually be helpful to achieve improved results, as also explained by Mehmood et al. (2014) wherein they obtained better results than similarity based approaches to find relatedness between research papers. The similarity and dissimilarity scores between these metadata parameters are calculated using the formulas described below:

Let $<S_i, C_{ij}>$ be a paper-citation pair, where S_i is the ith source (cited) paper and C_{ij} is the jth citing paper of S_i.

Let p_n be the nth parameter is our parameters set and let $v(S_i, C_{ij}, p_n)$ be the value of parameter p_n in paper-citation/reference pair $<S_i, C_{ij}>$

Suppose that S_i contains q different citations/references $(S_i, C_{i1}),\ldots,(S_i, C_{iq})$, resulting in m values for p_n, $v(S_i, C_{i1}, p_n),\ldots, v(S_i, C_{in}, p_n)$. Let P be a set of parameters. $P=\{p_u, p_b, p_t, p_a, p_{ab}, p_k, p_c, p_r\}$ where

p_u={List of unigram terms present in titles of S_i, and C_{ij}}
p_b={List of bigram terms present in titles of S_i and C_{ij}}
p_t={List of trigram terms present in titles of S_i and C_{ij}}
p_a={List of authors present in *Si* and C_{ij}}
p_k={List of keywords present in S_i *and* C_{ij}}
p_c={List of categories present in S_i *and* C_{ij}}
p_r={List of titles of references present in S_i *and* C_{ij}}
m=Total no. of citation papers

Title similarity and dissimilarity

The similarity and dissimilarity score between titles of pairs are calculated separately for each n-gram: *unigram similarity and dissimilarity score between titles, bigram similarity and dissimilarity score between titles and trigram similarity and dissimilarity score between titles* are calculated by using formulas in Eqs. 1, 2 and 3 respectively.

$$P1_{ij}(\text{title}_{\text{unigram}}) = \frac{|(S_i(p_u)) \cap \sum_{j=1}^{m} (C_{ij}(p_u))|}{|(S_i(p_u)) \cup \sum_{j=1}^{m} (C_{ij}(p_u))|} \tag{1}$$

$$P2_{ij}(\text{title}_{\text{bigram}}) = \frac{|(S_i(p_b)) \cap \sum_{j=1}^{m} (C_{ij}(p_b))|}{|(S_i(p_b)) \cup \sum_{j=1}^{m} (C_{ij}(p_b))|} \tag{2}$$

$$P3_{ij}(\text{title}_{\text{trigram}}) = \frac{|(S_i(p_t)) \cap \sum_{j=1}^{m} (C_{ij}(p_t))|}{|(S_i(p_t)) \cup \sum_{j=1}^{m} (C_{ij}(p_t))|} \tag{3}$$

Authors overlap

The approach of (Valenzuela et al. 2015), treats more than one common author equally by assigning value of 1 to one or more common authors and value of 0 for no common authors. While our scheme does not treat more than one common author equally, it measures the ratio of all common and uncommon authors by using formula in Eq. 4.

$$P4_{ij}(\text{Authors}) = \frac{|(S_i(p_a)) \cap \sum_{j=1}^{m}(C_{ij}(p_a))|}{|(S_i(p_a)) \cup \sum_{j=1}^{m}(C_{ij}(p_a))|} \tag{4}$$

Keywords

Similar to keywords, this formula calculates the ratio of similarity and dissimilarity between fully matched terms of keywords between pair, which is calculated using formula in Eq. 5.

$$P5_{ij}(\text{Keywords}) = \frac{|(S_i(p_k)) \cap \sum_{j=1}^{m}(C_{ij}(p_k))|}{|(S_i(p_k)) \cup \sum_{j=1}^{m}(C_{ij}(p_k))|} \tag{5}$$

Bibliographically coupled references

We are interested to examine the existence of same articles in the bibliography of *important* pairs. We assume that *important* pairs are more likely to cite same work than *non-important* ones. In case of paper-citation pairs, the references are matched after removing the citation of *cited paper* in *cited by* paper. The score is calculated using formula in Eq. 6.

$$P6_{ij}(\text{References}) = \frac{|(S_i(p_r)) \cap \sum_{j=1}^{m}(C_{ij}(p_r))|}{|(S_i(p_r)) \cup \sum_{j=1}^{m}(C_{ij}(p_r))|} \tag{6}$$

Content based parameters

Abstract

The abstract similarity is calculated by measuring the cosine of *tf-idf* scores. The cosine similarity is computed using Apache *Lucene indexing*[3] using formula in Eq. 7.

$$P7\,(\text{Abstract}) = \cos(x, y) = \frac{x \cdot y}{\|x\|\|y\|} \tag{7}$$

where x and y are *tf-idf* vectors of documents whose similarity has to be calculated.

[3] https://lucene.apache.org/core/.

Cue terms

The disjoint unigram cue terms are extracted from citing sentences of 60% of the paper-citation pairs from $D1$. These extracted terms are then identified from remaining 40% of the pairs. Disjoint words/phrases are those words/phrases which appear in the important citations and do not appear in the non-important citations and vice versa. Following steps are performed for disjoint cue words extraction.

- First of all, two separate lists of phrases for important and non-important citations are generated.
- The duplication is removed from both the lists.
- The common words in the both lists are removed.

Now both the lists contain only disjoint phrases. These two lists are parsed for classification, details about which are delineated in section IV. The example of disjoint cue-terms from important and non-important citation is shown in the Table 2.

Table 2 Disjoint cue-terms

Disjoint words for important class	Disjoint words for non-important class
Supper	Stem
Reason	Appear
Attract	Length
Deficit	Fetch
Slow	Deep

Results

The calculated scores are assigned for supervised machine learning by using WEKA tool. We have evaluated combinations of every parameter where n is the no. of parameters and r is the size of combination (i.e., combination of 2 up to 5). While building the combinations, we further split the title parameter into three features: *title unigram*, *title bigram* and *title trigram* and treated them as an independent feature. We have not harnessed them collectively in order to discover the potential of each n-gram individually. The Support Vector Machine with RBF kernel and degree 2, Random Forest with total no. of trees 10 and maximum depth 0 and Kernel Logistic Regression with degree 2, are utilized. Since we have class imbalanced problem (i.e., no. of *incidental* class is more than *important* class), which can lead to biasness of results by always guessing the *incidental* class accurately. Therefore, we have equalized the number of both classes by applying SMOTE filter[4] in WEKA. It generates artificial instances based on the original instances. We have performed 10-fold cross validation for all combinations by using machine learning tool *WEKA*. We have macro averaged the results of precision, recall and F-measure for all combinations and presented top scored combinations. Since, we have macro averaged the precision, recall and F-measure score of both classes. Therefore, it is not mandatory that value of F-measure

[4] http://weka.sourceforge.net/doc.packages/SMOTE/weka/filters/supervised/instance/SMOTE.html.

lies in the middle of precision and recall. The following sections address the first two questions (see "Approach" section).

Single metadata parameters

In single metadata parameters, it is analyzed that out of title unigram, title bigram and title trigram, authors, abstract, keywords, and references, which metadata parameter has produced the best result. There are total 6 and 7 single combinations in case of $D1$ and $D2$ respectively. From results, it can be seen that $P2$ (Title Bigrams) has individually performed significantly better than other features for both the data sets by achieving precision of 0.35 for $D1$ and 0.38 for $D2$ (see Figs. 3 and 4). Similar behavior of bigram in both data sets makes it a strong parameter for *important* citations identification. Usually, the maximum match occurs between unigram but bigram terms are more meaningful than unigrams. Similarly, trigram terms are more meaningful than bigram terms, but trigram combinations are rarely found in both data sets. Although, the variations have been seen in case of other parameters like bibliographically coupled references placed on number 3 in case of $D1$ and on last position for $D2$.

Double metadata parameters

In double metadata parameters, every possible combination of two metadata parameters is examined. There are total 12 and 18 double combinations for $D1$ and $D2$ respectively. In case of $D1$, the combination $P2$ and $P6$ has achieved the highest score by obtaining precision of 0.38 and for $D2$ the $P2$ and $P6$ have performed well by obtaining precision of 0.41 (see Figs. 4 and 5).

Triple metadata parameters

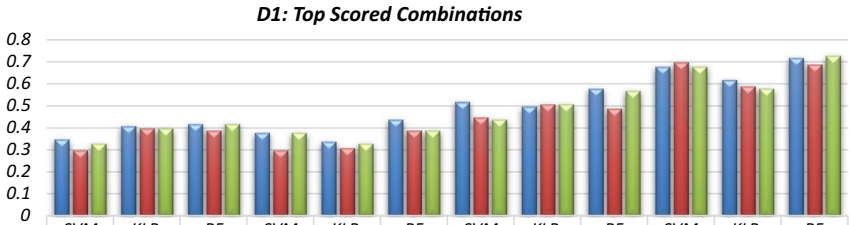
In triple metadata parameters, every possible combination of three metadata parameters is analyzed. There are total 10 and 31 single combinations in case of $D1$ and $D2$ respectively. $P2$, $P4$ and $P6$ have collectively performed better than other triple combinations by obtaining precision of 0.52 for $D1$. For $D2$, the parameters $P2$, $P4$ and $P7$ have performed well by obtaining precision of 0.41 (see Figs. 4 and 5).

Quadruple metadata parameters

In quadruple metadata parameters, every possible combination of four metadata parameters is analyzed. There are total 3 and 16 single combinations in case of $D1$ and $D2$ respectively. Form 3 combinations in $D2$, the parameters $P2$, $P4$, $P6$ and $P7$ have outperformed other parameters with precision of 0.68 (see Fig. 3 and 4). For $D2$, the parameters $P2$, $P4$, $P6$ and $P7$ have collectively outperformed by attaining the precision of 0.50 (see Fig. 4 and 5).

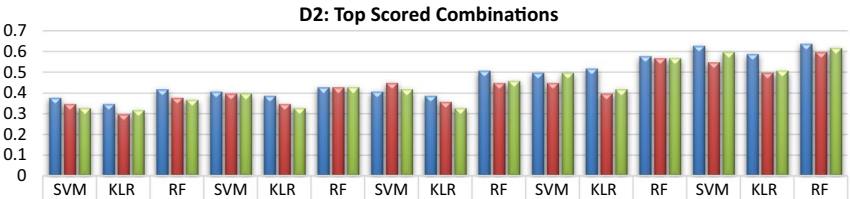
Quintuple metadata parameters

In quintuple parameters, every possible combination of five metadata parameters is formed to analyze which combination produces best results. All the combinations from $D2$ have



	SVM	KLR	RF	SVM	KLR	RF	SVM	KLR	RF	SVM	KLR	RF
Precision	0.35	0.41	0.42	0.38	0.34	0.44	0.52	0.5	0.58	0.68	0.62	0.72
Recall	0.3	0.4	0.39	0.3	0.31	0.39	0.45	0.51	0.49	0.7	0.59	0.69
F-measure	0.33	0.4	0.42	0.38	0.33	0.39	0.44	0.51	0.57	0.68	0.58	0.73

Fig. 4 Evaluation of top scored combinations from *D*1



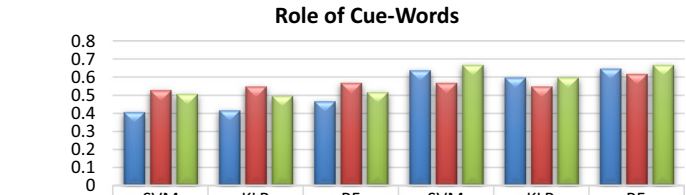
Precision	0.38	0.35	0.42	0.41	0.39	0.43	0.41	0.39	0.51	0.5	0.52	0.58	0.63	0.59	0.64
Recall	0.35	0.3	0.38	0.4	0.35	0.43	0.45	0.36	0.45	0.45	0.4	0.57	0.55	0.5	0.6
F-measure	0.33	0.32	0.37	0.4	0.33	0.43	0.42	0.33	0.46	0.5	0.42	0.57	0.6	0.51	0.62

Fig. 5 Evaluation of top scored combinations from *D*2

been evaluated. For *D*2, there are total 3 combinations from which *P*2, *P*4, *P*5, *P*6 and *P*7 have collectively attained the precision of 0.63 (see Figs. 4 and 5).

So far, the presented results are produced by harnessing metadata based feature and freely available content based parameter (i.e., Abstract). The results have shown that metadata can be deemed as the best alternate of content in the scenarios wherein content in unavailable. However, the pairs in D1 belongs to similar domain so there is a high chance that their metadata would be similar. To ensure the accuracy of our model, we have incorporated cue-words based parameter to ensure its validity in the scenarios where one intends to identify important citations from interdisciplinary domain. For this, we have passed the cue-words parameter to WEKA in two forms: (1) individually and (2) combining it with all other metadata parameters. The obtained results are shown in Fig. 6 below.

The above figure shows that individually cue-terms have performed well than single top scored metadata parameter (i.e., Bigram) by attaining F-measure of 0.52. We have



	SVM	KLR	RF	SVM	KLR	RF
	Cue Terms			P2+P4+P5+P6+P7+Cue-Terms		
Precision	0.41	0.42	0.47	0.64	0.6	0.65
Recall	0.53	0.55	0.57	0.57	0.55	0.62
F-measure	0.51	0.5	0.52	0.67	0.6	0.67

Fig. 6 Role of cue-words

manually investigated the pairs and analyzed that most of the pairs in important citations having distinct terms in titles are identified accurately as important citations due to appearance of specific cue-terms. We have carried out this experiment just to ensure the validity of proposed scheme for the scenarios wherein one intends to tackle the important citations by adopting the proposed model for interdisciplinary domain.

Comparisons

Whether our approach can behave closer to the content based approach?

In citation classification community, Valenzuela et al. (2015) have proposed first approach to tackle the problem of *important* citations identification. They have formed twelve different features from which most of the features are based on the content of articles. The significant justification of the proposed scheme can further be ensured by comparing the outcomes with Valenzuela's approach. *Why to compare the results with this scheme?* We compared the proposed scheme with Valenzuela's approach due to following reasons:

- We have comprehensively scrutinized the contemporary studies in order to find some standard data sets that have been employed in citation classification based studies. To the best of our knowledge, this is the only annotated data set of the required nature, which is publicly available.[5] Therefore, we have utilized this data set by following their predefined (Valenzuela et al. 2015) definition of *important* and *non-important* citations.
- Their scheme also performs similar binary citation classification by using both content and metadata based parameters wherein most of the features are based on content. We claim that similar or closer results can be obtained using most of the features from freely available metadata to perform same sort of binary citation classification.

Top scored features

The top scored single feature of Valenzuela's approach is *direct citations per section* (content based feature) that has obtained precision of 0.37 and our top scored single parameter from metadata parameters is Title_Bigram (P2) that has obtained precision of 0.35. Since there is a minor difference between both values of precision (i.e., 0.35 vs. 0.37) but

[5] http://allenai.org/data.htm.

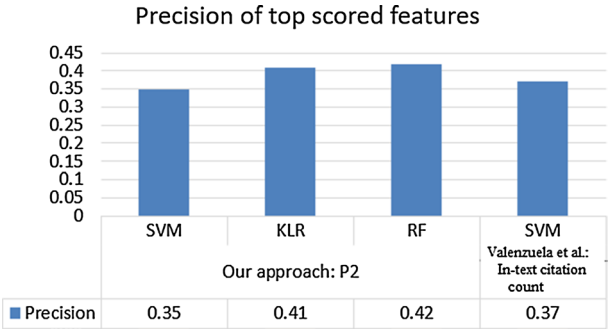


Fig. 7 Top scored combinations comparisons

considering the fact that this precision is obtained just by exploiting freely available metadata, precision of 0.35 seems quite higher (see Fig. 7).

Valenzuela et al. (2015) have also utilized two metadata based features:

1. Author overlap
2. References count

Both of these features are also part of our proposed scheme. However, the method of harnessing these metadata parameters distinguishes both schemes. Let's have a detailed look in the following sections.

Authors overlap

Similar to (Valenzuela et al. 2015) we have also calculated authors overlap score. However, there is a difference between method of this score calculation. For instance, the approach of (Valenzuela et al. 2015) treats more than one common author equally by assigning value of 1 to one or more common authors and value of 0 for no common authors. While our scheme does not treat more than one common author equally, it measures the ratio of all common and uncommon authors. We believe ratio obtained by this method can help in identifying important citations more accurately because more common authors between pairs increase the chances of being *important* citation of the cited paper. Their author overlap score has attained precision of 0.22 and our scheme has attained precision of 0.24 for same classifier SVM, and value of precision improved when we harnessed other classifiers, i.e., 0.34 for KLR and 0.45 for Random Forest (see Fig. 8).

The Valenzuela's scheme has also provided equal weightage to the idea of author matching based on the assumption that if citing and cited papers have same author(s) then there is a fair chance that the citing work has extended or used the cited work. We do not provide importance to the *author* solely based on this assumption. Rather, we have analyzed this factor in both the data sets ($D1$ and $D2$). Most of the time, same authors are found in *important* pairs than in *incidental* ones. From $D1$, 42% of the common authors are found in *important* pairs and 24.5% common authors are found in incidental pairs. From $D2$, 58.4% of the common authors are found in *important* pairs and 33% of the common authors are found in incidental pairs. Based on these facts, we cannot disregard the presence of same author(s) in *important* citations. Therefore, it is a quintessential feature of the proposed scheme. However, if one intends to adopt the proposed scheme specifically for quantitative

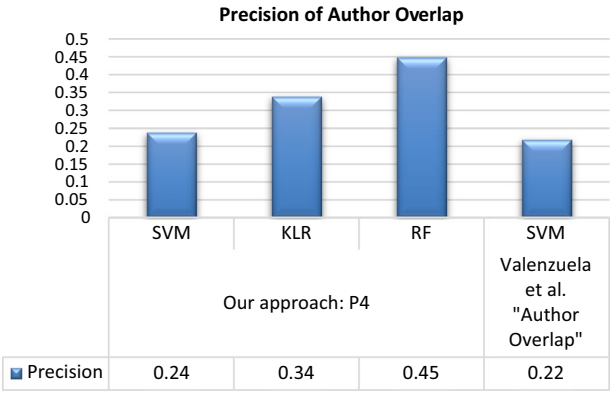


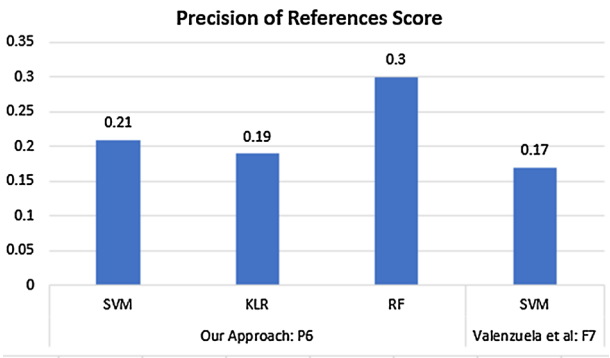
Fig. 8 Author's overlap comparison

citation based approaches (i.e., citation count) then the pairs containing same authors should be excluded to disregard the self-citations.

References count

This reference feature of Valenzuela's scheme computes the inverse of the length of the citing paper's reference list (1/number of references), which hints to the value of receiving one citation, e.g., if it is one citation from a total of two references, this citation is clearly *important*. We have also taken references parameter into account but with different method as of Valenzuela's. We think value of ratio between common and uncommon references can hint the importance of a citation. In our approach, first of all, heuristic approach is applied to extract the titles of papers form reference list of both citing and cited paper. This approach is built by considering the reference syntax of the research articles published in ACL anthology (see Fig. 2). Afterwards, we have calculated the ratio between common and uncommon titles of references between pairs. The score is calculated using formula in Eq. 6. The obtained results are presented in Fig. 6. Their reference score (*f*7) has attained precision of 0.17 and our scheme has attained precision of 0.21 using same classifier (i.e., SVM), 0.19 for KLR and 0.30 for Random Forest classifier. The results are shown in Fig. 9.

Fig. 9 References score comparison



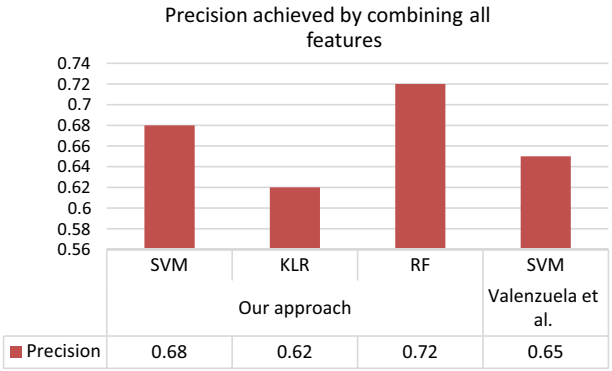


Fig. 10 Comparisons of overall results

Overall results

Valenzuela's approach has achieved average precision of 0.68 by combining all the twelve features. We have presented the precision of each classifier individually. Our system achieved precision of 0.68 against SVM, precision of 0.62 against KLR, and precision of 0.72 against Random Forest. It is pertinent to mention here that our system has achieved improved precision just by relying on freely available metadata. The results are envisioned in Fig. 10.

Another important finding of this study is the potential of bigram terms similarity between *important* paper citation pairs. P2 (similarity and dissimilarity between titles bigram) has outperformed individually with precision of 0.35 and exists in every top scored combination of $D1$ and $D2$. In case of $D2$, the metadata parameter "keywords" is found in top scored combinations of triple, quadruple and quintuple metadata parameters. Considering this result, if we had 100% availability of Keywords then it could be one of the top scored features for D2. However, availability of keywords is 58.3%, therefore we cannot form a generic conclusion. Another important result to be considered is that the random forest classifier has successively attained good values of evaluation measures for all the combinations in $D1$ and $D2$.

Let us recapitulate the obtained results further, continuing the stream of automatic citation classification, we have proposed a model for binary citation classification by using only freely available information of citations (i.e., metadata and abstract).

Usually, citation based approaches have advantage that most of the time, citing author cites those papers that are topically relevant. This minimizes the effort to find relevant papers. It can be assumed that the pair (citing paper-cited paper) are topically relevant. Now, *how to figure that extracted citations are not only relevant but also important citation (following the definition of important citations) of the citing paper?* In the proposed scheme, the degree of relevancy between metadata parameters plays a major role and distinguishes our scheme with simple relevance finding techniques. As in obtained results for both data sets, the unigram terms have extracted both the relevant and *important* citations. However, when we switched to one order higher n-gram (i.e., bigram) then it has specifically tackled more *important* citations than topically relevant citations. To verify further, we have analyzed the behavior of bigram in $D2$ and discovered that bigram behaved in a similar pattern for $D2$ as well.

Valenzuela et al. have also employed few similar metadata parameters (i.e., authors and references) to identify important citations. We have removed the existing deficiencies in their method of using these parameters; our proposed method has shown a significant improvement

in the results. On the basis of overall analysis, we can say that the proposed scheme is coherent, contains feasible methodology, which ensue it as a significant contribution in citation classification community.

Conclusion

From early 90s to date, citations are reckoned as requisite measure to assess different purposes, such as author ranking, formulation of academic policies, deciding reviewer for journal or conference etc. The identification of 15 different citing behaviors by Garfield opened the extensive dimensions of criticism for pure quantitative citation based measures. Researchers argued that citations of perfunctory nature, or given just to provide background knowledge should be filtered out from the citation count to transform it into a reliable measure. In this regard, numerous citation classification studies have been presented to tackle the citation reasons. The trend of citation classification started from their classification into 35 categories (first approach of automatic citation classification) to now classifying citation into only two categories (*important and non-important*). Gradually, the number of these classes receded and researchers payed more attention to form a set of comprehensive features to classify citations exclusively into meaningful classes rather than into vague conflicting classes. However, the plethora of these schemes is dominated with content based features following the typical pattern of content exploitation (i.e., extraction of cue-phrases from citation context). The known phenomenon-free availability of research articles, limits the scope of these studies. Major journal publishers such as IEEE, Elsevier, Springer etc. do not provide open access to their articles. There are financial, legal or technical barriers to acquire the content of research articles. On the other hand, mostly the metadata of research papers is freely available. We have presented a supervised learning binary citation classification technique that is dominated by set of distinct metadata features such as *titles, authors, keywords and references* etc. The main objective of this study is to analyze the extent to which metadata can behave similar to the content. The study has employed two benchmark data sets $D1$, containing 465 paper-citation pairs, and D2 having 488 paper-citation pairs. SVM, KLR and RF, have been employed for classification. The citations are classified using tenfold cross validation in WEKA. We have also resolved the class imbalance problem from both data sets using SMOTE. We have evaluated each possible combination of employed parameters to identify the best performing combination. Individually, among the metadata parameters, the bigram of titles matching strategy between paper-citation pairs has attained F-measure of 0.42 for $D1$ and 0.37 for $D2$. By combining all the features, our system yielded F-measure of 0.73 for $D1$ and 0.62 for $D2$. We have compared our results with the contemporary content based state-of-the-art approach to tackle the similar and diverging behavior between content and metadata. Our top scored metadata feature *(title_bigram)* outperformed their top scored content feature *(in-text citation count)* by attaining the precisions of 0.42 (i.e., 0.42 vs. 0.37). The results of the study signify the potential of metadata parameters. We claim that proposed approach is adequately suitable to be utilized as an alternative of content to classify citations. This study would immensely serve the scholars via providing them list of important papers form best possible accessible source (i.e., citations of the source paper). Moreover, the study can also be employed in the stream of studies focusing on filtering the incidental citations from mere citation count approaches. However, the study has certain limitations discussed in the following section.

Limitations and future work

There could be multiple other reasons of citations that are important, such as, authors have refuted the citing work etc. For this, annotation of data should be reformed according to updated definition that considers each possible *important* and *non-important* reason in both classes. In this scheme, we have employed two annotated data sets as best of our effort. Currently, there is a lack of availability of such sort of large annotated data set. We think that significance of metadata in tracking important citations could further be ensured by forming an extensive annotated data set by covering different domains, maximum amount of metadata and authors from different geographical locations; and more specifically, updating the important and non-important classes by incorporating other suitable reason in both classes.

References

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* (Vol. 1, pp. 500–509). Stroudsburg, PA: Association for Computational Linguistics.

Anderson, R., Narin, F., & McAllister, P. (1978). Publication ratings versus peer ratings of universities. *Journal of the American Society for Information Science, 29*(2), 91–103.

Ayaz, S., & Afzal, M. T. (2016). Identification of conversion factor for completing-h index for the field of mathematics. *Scientometrics, 109*(3), 1511–1524.

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

Bonzi, S. (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.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80.

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.

Diederich, J., & Balke, W. T. (2007). The semantic growbag algorithm: Automatically deriving categorization systems. In *International conference on theory and practice of digital lib*raries (pp. 1–13). Berlin: Springer.

Ellis, D. (1993). Modeling the information-seeking patterns of academic researchers: A grounded theory approach. *The Library Quarterly, 63*(4), 469–486.

Finney, B. (1979). The reference characteristics of scientific texts. Master's thesis. London: The City University of London.

Garfield, E. (1965). Can citation indexing be automated. In *Statistical association methods for mechanized documentation, symposium proceedings* (Vol. 269, pp. 189–192). Washington, DC: National Bureau of Standards, Miscellaneous Publication 269.

Garzone, M., & Mercer, R. (2000).Towards an automated citation classifier. In *Conference of the canadian society for computational studies of intelligence* (pp. 346–337). Berlin: Springer.

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.

Hirsch, Jorge E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569–16572.

Inhaber, H., & Przednowek, K. (1976). Quality of research and the Nobel prizes. *Social Studies of Science, 6*(1), 33–50.

Jeong, Y., Song, M., & Ding, Y. (2014). Content-based Author co-citation analysis. *Journal of Informetrics, 8*(1), 197–211.

Jochim, C., & Schütze, 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.

Krikelas, J. (1983). Information-seeking behavior: Patterns and concepts. *Drexel Library Quarterly, 19*(2), 5–20.

Lawrence, S., Giles, C. L., & Bollacker, K. D. (1999). Digital libraries and autonomous citation indexing. *Computer, 32*(6), 67–71.

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.

MacRoberts, M. H., & MacRoberts, B. R. (2018). The mismeasure of science: Citation analysis. *Journal of the Association for Information Science and Technology, 69*(3), 474–482.

Mai, J. E. (2016). *Looking for information: A survey of research on information seeking, needs, and behavior*. Bingley: Emerald Group Publishing.

Mazloumian, A., Helbing, D., Lozano, S., Light, R. P., & Börner, K. (2013). Global multi-level analysis of the 'Scientific Food Web'. *Scientific, reports,* 3.

Mehmood, Q., Qadir, M., & Afzal, M. (2014). Finding relatedness between research papers using similarity and dissimilarity scores. In 15th international conference Web-Age information Management (pp. 707–710). Macau, China.

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.

Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Washington, DC: 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.

Peroni, S., & Shotton, D. (2012). FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web, 17,* 33–43.

Pham, S., & Hoffmann, A. (2003). A new approach for scientific citation classification using cue phrases. In L. C. C. F. Tam´as Domonkos Gedeon (Ed.), *AI 2003: Advances in artificial intelligence* (Vol. 2903, pp. 759–771)., Lecture notes in computer science Berlin: Springer.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Raheel, M., Ayaz, S., & Afzal, M. T. (2018). Evaluation of h-index, its variants and extensions based on publication age & citation intensity in civil engineering. *Scientometrics, 114*(3), 1107–1127.

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*. London: University of Royal Holloway.

Spiegel-Rusing, I. (1977). Science studies: Bibliometric and content analysis. *Social Studies of Science, 7*(1), 97–113.

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). Cambridge: CUP Archive.