Computers, Materials & Continua
DOI:10.32604/cmc.2021.013809
Article

Tech Science Press

Extension of Direct Citation Model Using In-Text Citations

Abdul Shahid[1,*], Muhammad Tanvir Afzal[2], Muhammad Qaiser Saleem[3], M. S. Elsayed Idrees[3] and Majzoob K. Omer[3]

[1]Institute of Computing, Kohat University of Science and Technology, Kohat, 26000, Pakistan
[2]Department of Computer Science, NAMAL Institute, Mianwali, 42250, Pakistan
[3]College of Computer Science and Information Technology, Albaha University, Al Baha, Saudi Arabia
[*]Corresponding Author: Abdul Shahid. Email: ashahid@kust.edu.pk

Received: 22 August 2020; Accepted: 19 October 2020

Abstract: Citations based relevant research paper recommendations can be generated primarily with the assistance of three citation models: (1) Bibliographic Coupling, (2) Co-Citation, and (3) Direct Citations. Millions of new scholarly articles are published every year. This flux of scientific information has made it a challenging task to devise techniques that could help researchers to find the most relevant research papers for the paper at hand. In this study, we have deployed an in-text citation analysis that extends the Direct Citation Model to discover the nature of the relationship degree-of-relevancy among scientific papers. For this purpose, the relationship between citing and cited articles is categorized into three categories: weak, medium, and strong. As an experiment, around 5,000 research papers were crawled from the CiteSeerX. These research papers were parsed for the identification of in-text citation frequencies. Subsequently, 0.1 million references of those articles were extracted, and their in-text citation frequencies were computed. A comprehensive benchmark dataset was established based on the user study. Afterwards, the results were validated with the help of Least Square Approximation by Quadratic Polynomial method. It was found that degree-of-relevancy between scientific papers is a quadratic increasing/decreasing polynomial with respect to-increase/decrease in the in-text citation frequencies of a cited article. Furthermore, the results of the proposed model were compared with state-of-the-art techniques by utilizing a well-known measure, known as the normalized Discount Cumulative Gain (nDCG). The proposed method received an nDCG score of 0.89, whereas the state-of-the-art models such as the Content, Bibliographic-coupling, and Metadata-based Models were able to acquire the nDCG values of 0.65, 0.54, and 0.51 respectively. These results indicate that the proposed mechanism may be applied in future information retrieval systems for better results.

Keywords: Direct citation model; in-text citations frequencies; normalized discount cumulative gain; least square approximation

1 Introduction

With the advent of the Web, a user can have access to enormous scientific knowledge repositories [1]. There is a rapid increase in research articles, journals, conferences, and open archives resulting in tremendous growth in the quantity and diversity of publications [2]. This scholarly work is built upon other works; hence it creates a citations network. This citation network information plays a crucial role in the scientific community.

The scientific community makes use of citations for various practical applications, ranging from extracting relevant information for devising several quality measures to acquire the effective assistance of citations in different scientific policies such as promotions of faculty, faculty hiring, ranking journals, and ranking of relevant documents. Furthermore, citations are also used to recommend relevant papers for the paper at hand. Apart from citations, current state-of-the-art methods exploit different data sources for the identification of related papers. These sources include the content, metadata of the papers, and user profiles. Such approaches and their critical analysis have been thoroughly covered in [3].

The widely known state-of-the-art citations-based models for relevant document identifications are Bibliographic Coupling [4], Co-Citations [5], and Direct Citations [6,7]. For a long period of time, these citation's models are used by the scientific community for discovering relevant documents at the surface level. In the current state-of-the-art techniques, all citations are treated equally for various tasks, such as finding relevant documents or quality measures *e.g.*, impact factor and h-index. It is a challenging issue to identify in-text citation with enough accuracy and also with a proper reason for citation, and probably this was the reason that it had not been used in such important measures [8]. In literature, it has also been reported that all citations are not of equal importance and thus should be considered differently [9]. The same notion was reinforced by distinctive research studies such as Moravcsik et al. [10], who found that 40% of the citations were perfunctory. Therefore, in recent times, these models have been studied and improved by exploiting citation's textual details [11,12]. Furthermore, The Direct Citation model has been found more accurate in the representation of knowledge taxonomies as compared to Co-citation and Bibliographic Coupling [13].

Therefore, in this study, a detailed study is presented which is based on the Direct Citation Model. The proposed work is based on citations and their in-text citation frequencies to find out the most relevant research articles. Furthermore, citations have been categorized into two major classes, i.e., (a) citations which are methodologically related, and (b) citations which are non-methodologically related. Each category is further explained in the light of the previous research on the reasons for citations [14]. For the evaluation of the proposed model, different experiments were conducted on the CiteSeerX dataset. The CiteSeerX is a well-known automatic citation indexing system that covers all topics of the computer science domain [15]. For experimental purposes, a crawler was developed to crawl and download research articles found on CiteSeerX related a given topic. A total of 5,000 documents were downloaded and, later on, converted into XML, using PDFx [16]. Subsequently, xPath and xQuery based solution was proposed to compute in-text citation frequencies. A total of around 105,000 references were extracted from these research articles. At the end, the sample data was annotated with the help of a user study to prepare the benchmark data. These results were validated with the help of the Least Square Approximation Method for discovering the nearest polynomial function. This method is commonly used for defining a generic continuous function of discrete data.

During the experiments, it was found that higher in-text citation frequency values corresponded to a strong relationship between citing and cited papers. For example, 77% of the time, strong relationship was identified between the cited and citing paper where the in-text citation frequencies of cited papers in the content of the citing paper were greater than or equal to five. Furthermore, 87% of the time, weak relationship was found between citing and cited papers when cited paper was cited less than five time in the citing paper. The results of the proposed model were compared with Content, Bibliographic Coupling, and Metadata-Based Models. The nDCG measure is utilized for the evaluation process. The proposed model produced the nDCG value of 0.89, whereas 0.65, 0.54, and 0.51 nDCG scores were recorded in the case of Content, Bibliographic Coupling, and Metadata-Based Techniques respectively.

2 Proposed Model

The proposed model is an extended version of the Direct Citation Model. It is based on an essential feature, i.e., in-text citation frequency. Basically, it is a kind of statistical technique that considers in-text citation frequencies of a cited paper in the citing paper. A scenario of citing and cited articles is shown in Fig. 1. In this scenario, in-text citation frequencies refer to the total numbers of occurrences of a cited paper in the citing paper. For example, in Fig. 1, the cited article ("[1]") in the reference list has been referred four times in the body text of the citing article.

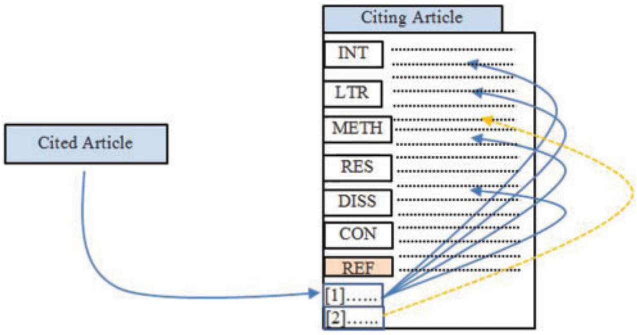


Figure 1: Citing paper, cited papers, and in-text citation frequencies

Similarly, the article ("[2]") has been referred once in the body text of the citing article. In this paper, the proposed model has been presented in terms of determining the degree-of-relevancy between cited and citing papers.

2.1 Mathematical Formulation

In this section, we provide a mathematical description of the hypothesis. This description consists of various concepts such as Citing article as (A), Cited article as (R), and the in-text citation count of R in Body Text (BT) of the A.

Suppose we have an article represented as $\boldsymbol{A}$ Whereas $\boldsymbol{A}$ consists of References ($\boldsymbol{R}$) and Body-Text ($\boldsymbol{BT}$) such that:

$R_{i=1}^{n}$ has a total number of in-text citations (N) >= 1 in *BT* and thus here is the hypothesis that

$$Relevance\,(\mathrm{A}) \propto N\left(R_{i=1}^{n}\right) \tag{1}$$

Eq. (1) states that the relevance of a cited and citing article increases with increase in the in-text citations frequencies in the body text of the citing article. In the result section, this has been verified with the help of Least Square Approximation by Quadratic Polynomial method. To validate the hypothesis, a feasible methodology was adopted as explained below.

2.2 Dataset

To the best of our knowledge, in the field of relevant document identification, no standard dataset is available for the evaluation of the proposed approaches. Therefore, like the previous researchers, we had no choice other than creating our own dataset. For this purpose, we developed a system for acquiring and preparing the dataset using CiteSeerX [15]. The CiteSeerX is an open-access indexing service that has indexed research papers from the computer science domain. Furthermore, it covers all topics of the computer science domain and indexes many journals and conferences. Therefore, it is a suitable resource for preparing the dataset and conducting research. The overall system architecture of data acquisition is shown in Fig. 2. The system consists of two main parts, i.e., the CiteSeer Crawler and XML Modules.

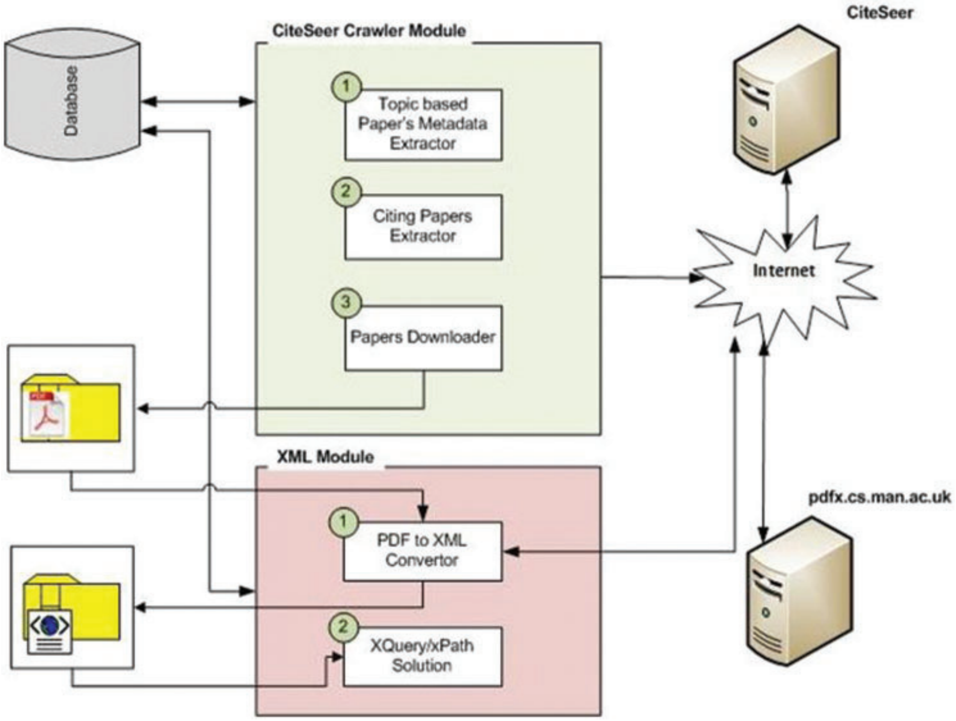


Figure 2: System architecture for data preparation

2.2.1 CiteSeer Crawler

The purpose of this module is to crawl the CiteSeerX for the given terms and then download those research papers for further analysis. This module starts working by extracting topic/searched terms from the database. The topics or searched terms persisted in the database. This module loads a term and then poses a query on the CiteSeerX. The used terms and their respective downloaded papers are shown in Tab. 1. The reason of variance in numbers is due to the CiteSeerX, which provides 500 links to the retrieved results. Furthermore, pagination has been applied to retrieve results, that's why ten records per page are displayed to the end-user.

This module traverses those records page by page, extracts metadata of each paper, and stores them in the MySQL database. CiteSeerX provides access to the top 500 citing papers. Thus, when an article has more than 500 citations, at most 500 are retrievable. Here we can see Papers downloader, which is a separate script that iterates over the metadata of the crawled papers and downloads the research articles.

2.2.2 XML Module

The purpose of this module is to convert research papers into XML formats. Although different tools are available for converting PDF into XML but we do not require simple conversion from PDF to EML as the research paper contains structured/semi-structured information that is needed to be extracted. Therefore, it is necessary to either directly acquire information from PDF documents or convert them into programmable-friendly versions such as XML. Manchester University has developed a tool known as PDFx [16], which takes the research paper as input and converts it into XML using ontologies, e.g., DoCo and DEo. This tool also identifies in-text citations in a research paper. Thus, this module sends a research paper in PDF format to the PDFx tool, using CURL to get it converted into XML format.

Table 1: Term-wise downloaded paper statistics

Terms	Total papers (cited and citing papers)
Ontology engineering	1320
Semantic computing	1224
Digital libraries	1336
Semantic web	1120

After the conversion of papers into XML format, detailed information about in-text citations and their frequencies are required. We developed the xQuery and xPath expression-based solution that extracts all citations of a research article and calculates in-text citation frequencies for each reference. With the help of this module, a total of 5,000 documents are converted into XML format; whereas around 105,000 references were extracted from the translated documents.

For the evaluation of the experiments, it was necessary to have a benchmark dataset. To the best of our knowledge, such a dataset is not available in the field of identification of relevant research articles. Therefore, it was mandatory to have a benchmark dataset, to which the results of our proposed system could be compared.

2.3 Annotation Scheme

The current state-of-the-art systems treat all citations equally [3,4,11]. All the citations are not equally important for the citing paper. For example, sometimes a paper is cited because the citing paper works on the same topic or builds its technique, based on the techniques mentioned in the cited paper and, sometimes, the citing document cites a particular paper to give the background study. Therefore, identification of relevant articles, based on the nature-of-relationship between cited and the citing documents help the scientific community. This issue has been raised over the decades in the literature.

Garfield, the pioneer in the citation analysis, has earlier described 15 different citation reasons to answer this question [14]. However, the identification of such a relationship between cited and citing documents requires an extensive analysis of the content. Such a relationship between the cited and citing documents can be classified into two major categories: (1) methodological relationship and (2) non-methodological relationship. We have explained the methodologically relevant and non-methodologically relevant relevance as below.

2.3.1 Methodologically Relevant

The cited and citing papers are methodologically related when citing papers:

- Have worked on the same problem as the cited work has done (SPRB).
- Have extended/compared their work with cited work (ECW).
- Have used some concepts, definitions of the cited work defined for the same problem (UCD).

2.3.2 Non-Methodologically Relevant

The cited and citing papers are Non-methodologically related when citing articles:

- Have referred to the cited document only to give background study or highlight the importance of the research (UP).
- Have used the cited text partially (cited work is used to complete citing paper's methodology) (UBI).

In the light of the previous research, we have grouped various reasons for citation. This mapping deals with the strength of the relationship between citing and cited papers. We refer to this type of relationship as degree-of-relevancy between citing and cited articles. In this type of mapping (degree-of-relevancy), the citations are further classified into three levels i.e., highly relevant, moderately relevant, and weakly relevant.

2.4 Gold Standard Dataset

As explained earlier in Section 2.2, a total of 5,000 documents were downloaded and their in-text citation frequencies were parsed. The overall contribution of in-text citation frequencies of paper's references i.e., 105,000 in these 5,000 papers is shown in Fig. 3. It was found that significant contributions are of in-text citation frequencies = 1, which is about 60% of extensive data.

The results followed an intuitive pattern that higher values of in-text citation frequencies cover a small number of portions. An exciting part of this result was that in-text citation frequencies = 0 had a significant value. It showed that some of the references given in a paper were not even referred a single time throughout the body text of the citing paper. It validated our old results that some references were given in the article to pay undue credit to some authors [17]. These results indicate that the current state-of-the-art techniques should exercise the role of in-text citation frequencies in their overall calculation.

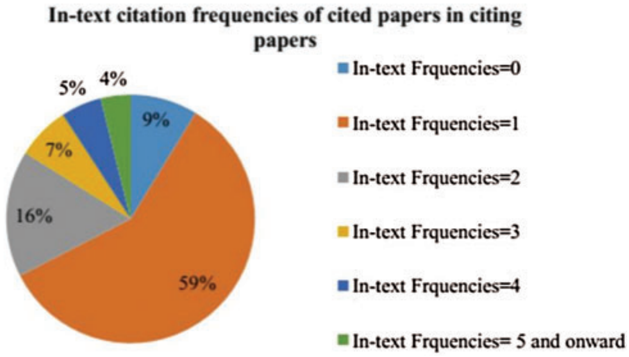


Figure 3: Contribution of citations having various in-text citation frequencies

In our case, annotation of the whole dataset of around 0.1 million citations was not feasible, as it required much time for the specialized resources (researchers). Therefore, as a first step, citations pairs were randomly selected in such a manner that they had coverage from all groups mentioned in Fig. 3. Thus, the total filtered records were 12,000 citations pairs. From these 12,000 citations pairs, 400 sample citations pairs were randomly selected for the user study. The overall distribution of the in-text citation frequencies among these 400 citations pairs is given in Tab. 2.

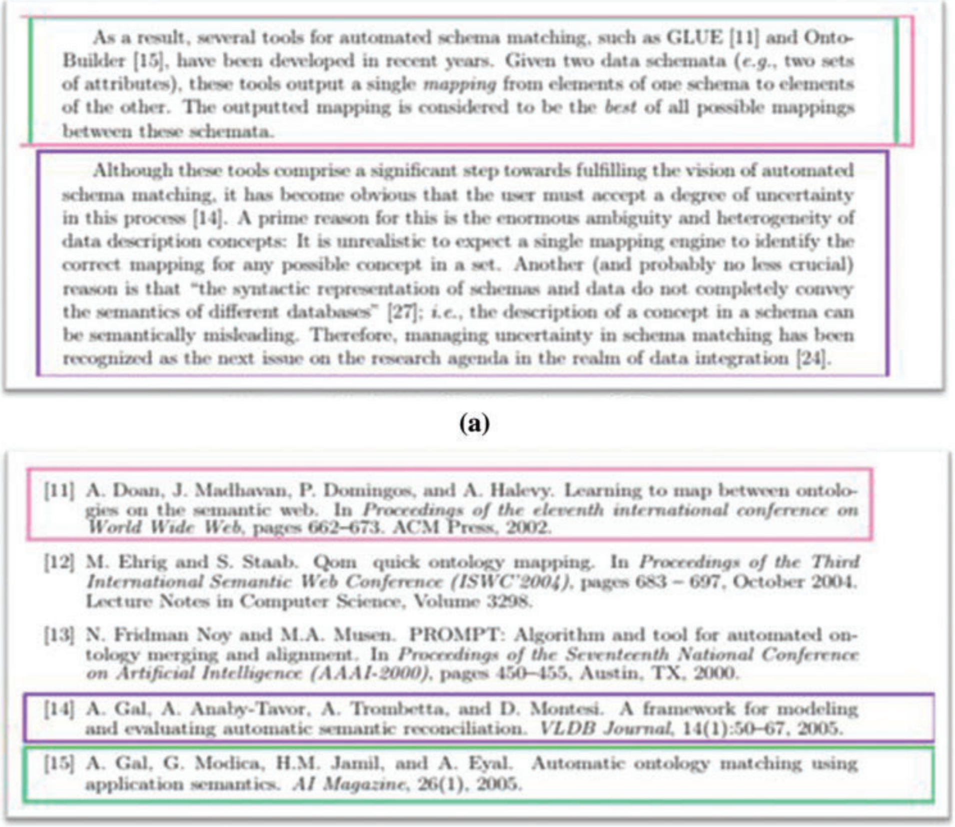
Table 2: Total citations context in citing papers for the selected dataset

In-text frequencies	Total instances	Citation context
1	94	94
2	64	124
3	74	229
4	56	224
>=5	112	560
Total citation contexts		1231

A citation pair means citing and cited paper, for example, CitationPair (p100, p34) means that the first entry represents the citing paper, and the second entry represents the cited paper. It means that paper number "p34" is cited by paper number "p100." Therefore, a user study was conducted on 400 citation pairs to prepare a benchmark dataset. The targeted users in this study were Ph.D. and MS students who were actively involved in their research activities. More than 120 students were approached; out of them, only those students were selected who had enough knowledge of conducting their research. Thus, only 66% i.e., 80 users were selected to participate on their will. The selected papers and their selected citations were given to the users for annotation purposes. The selection of the citing papers was made on its relevancy to the researcher's profile so that they did not feel any difficulty in understanding that paper. Authors cite the papers of other researchers for some reason. Therefore, the author's sentiments for the cited article can always be found around the text of their in-text citations in the paper, which is called the citation context. Different researchers also exploited the citation context for discovering the sentiments of the authors for the cited paper [18].

The backgrounds of the selected 400 citation-pairs were marked for an easy and quick decision about the relationship between citing and cited papers. In Fig. 4a, it is shown how the citation contexts were marked for the users. In Fig. 4a, in-text citations marked in a paper titled "Managing Uncertainty in Schema Matching with Top-K Schema Mappings" are shown, while in Fig. 4b, their references are shown. Similarly, wherever in-text citations were found, they were marked for a more in-depth analysis of the users.

Thus, more than 1,230 citation contexts were exploited in our user study. Along with citation context, the full contents of the papers were also available for users to further explore the context of the cited papers. The overall details of citations and their contexts are shown in Tab. 2. For annotation purposes, three things were provided to the users: first is the selected paper where references were marked along with their in-text citations in the citing paper (as shown in Fig. 4a); second is the "Citation reasons form" as shown in Tab. 3; third is the list of selected citations list from the paper.



As a result, several tools for automated schema matching, such as GLUE [11] and OntoBuilder [15], have been developed in recent years. Given two data schemata (*e.g.*, two sets of attributes), these tools output a single *mapping* from elements of one schema to elements of the other. The outputted mapping is considered to be the *best* of all possible mappings between these schemata.

Although these tools comprise a significant step towards fulfilling the vision of automated schema matching, it has become obvious that the user must accept a degree of uncertainty in this process [14]. A prime reason for this is the enormous ambiguity and heterogeneity of data description concepts: It is unrealistic to expect a single mapping engine to identify the correct mapping for any possible concept in a set. Another (and probably no less crucial) reason is that "the syntactic representation of schemas and data do not completely convey the semantics of different databases" [27]; *i.e.*, the description of a concept in a schema can be semantically misleading. Therefore, managing uncertainty in schema matching has been recognized as the next issue on the research agenda in the realm of data integration [24].

(a)

[11] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to map between ontologies on the semantic web. In *Proceedings of the eleventh international conference on World Wide Web*, pages 662–673. ACM Press, 2002.

[12] M. Ehrig and S. Staab. Qom quick ontology mapping. In *Proceedings of the Third International Semantic Web Conference (ISWC'2004)*, pages 683 – 697, October 2004. Lecture Notes in Computer Science, Volume 3298.

[13] N. Fridman Noy and M.A. Musen. PROMPT: Algorithm and tool for automated ontology merging and alignment. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, pages 450–455, Austin, TX, 2000.

[14] A. Gal, A. Anaby-Tavor, A. Trombetta, and D. Montesi. A framework for modeling and evaluating automatic semantic reconciliation. *VLDB Journal*, 14(1):50–67, 2005.

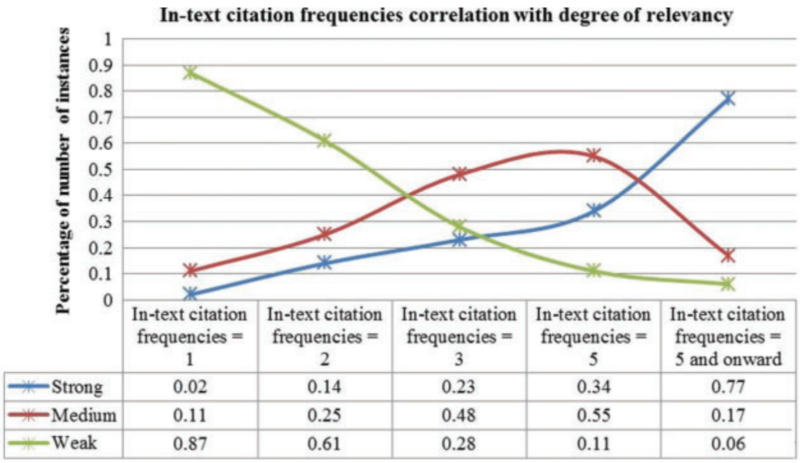
[15] A. Gal, G. Modica, H.M. Jamil, and A. Eyal. Automatic ontology matching using application semantics. *AI Magazine*, 26(1), 2005.

(b)

Figure 4: Contribution of citations having various in-text citation frequencies. (a) In-text citations marking (citations context) of selected references in a paper. (b) List of selected references in a paper for a user study

The users were asked to fill in the citation reason code after the analysis of citation reasons for a citation. To obtain multiple judgments on the same citations pair, we assigned the same citation pair to two different users.

Based on the criteria mentioned above, 400 citations were classified into three categories i.e., “Strong,” “Medium,” and “Weak” relationship with the citing paper in a degree-of-relevancy based classification. It means that every annotated citation has a specific group i.e., either methodologically relevant or non-methodologically relevant. Among the total of 400 citation pairs, there was a difference of opinion among annotators on 82 citation pairs in the degree-of-relevancy grouping. These disputed citation pairs were not considered in the results of the final experiments.



	In-text citation frequencies = 1	In-text citation frequencies = 2	In-text citation frequencies = 3	In-text citation frequencies = 5	In-text citation frequencies = 5 and onward
Strong	0.02	0.14	0.23	0.34	0.77
Medium	0.11	0.25	0.48	0.55	0.17
Weak	0.87	0.61	0.28	0.11	0.06

Figure 5: In-text citation frequencies mapping over the degree of relevancy

Table 3: Citation reasons form

Citation reasons	Code
Paper has worked on the same problem as the cited work has done.	SPRB
Citing paper has extended/compared their work with cited work.	ECW
Citing paper has used some concepts, definitions of the cited work.	UCD
Citing paper has used the cited document partially (cited work is used to complete citing paper methodology).	UP
Citing paper has referred cited document only as a background study or highlighting the importance of the research.	UBI

3 Results

To validate the hypothesis, several results were obtained which show the impact of in-text citation frequencies and their contribution towards the degree-of-relevancy.

3.1 In-Text Citation Frequencies' Correlation with Degree-of-Relevancy

In this experiment, the degree-of-relevancy correlation with in-text citation frequencies was explored. In-text citation frequencies were mapped over the degree-of-relevancy between citing and cited papers. Results are shown in Tab. 4. The first column represents the in-text citation frequencies. The next column with labels "Strong," "Medium," and "Weak" represents the number of agreed-upon instances that are classified in respective classes. The last column presents the number of cases where the inter-annotator agreement is not identical. We did not consider the disputed cases for further processing.

These results indicated that lower in-text citation frequencies typically represent a weaker relationship between citing and cited paper. For example, total of 85 citation instances having in-text citation frequencies = 1, a fragile relationship was found for 74 cases. Similarly, these patterns hold for in-text citation frequencies = 2. Most of the time, a moderate correlation between citing and cited paper was reported for in-text citation frequencies = 3 and 4. However, a good number of strong relationships were also recorded for in-text citation frequencies = 3 and 4. Lastly, the strong relationships were found for in-text citation frequencies = 5 and more significant. However, in that case, some medium relationship was also recorded, that is 17%. In this experiment, the degree-of-relevancy correlation with in-text citation frequencies was explored. The in-text citation frequencies were mapped over the degree-of-relevancy between citing and cited papers. The results are shown in Tab. 4. The In-text citation frequencies' correlation with degree-of-relevancy is shown in Fig. 5 with the help of a line graph.

Table 4: Mapping of in-text citation frequencies on the degree of relevancy

Frequencies	Strong	Medium	Weak	Un-decided
In-text citation frequencies = 1	2	9	74	9
In-text citation frequencies = 2	7	13	31	13
In-text citation frequencies = 3	14	29	17	14
In-text citation frequencies = 4	13	21	4	18
In-text citation frequencies = 5 and onward	65	14	5	28

The results can be summarized that the lower in-text citation frequencies represent a weak relationship between citing and cited documents. The text citation frequencies of values 3 and 4 represent a medium relationship between citing and cited documents. Finally, for in-text frequencies = 5 and onward, there exists a strong correlation between cited and citing papers. In Fig. 5, the percentage of the total number of instances is shown on Y-axis, whereas X-axis represents the in-text citation frequencies ranges.

3.2 Mathematical Validation of Hypothesis

The results shown in Fig. 5 indicate that the in-text citation frequencies contribute towards the degree-of-relevancy between citing and cited articles are polynomial in nature. Therefore, it can be described as below in Eq. (2).

$$f(x) = ax^2 + bx + c \tag{2}$$

As per the procedure of the Least Square Approximation Quadratic Polynomial, the objective is to define an equation such that the polynomial is shown in Fig. 5 (for Strong, Medium, and Weak relations) can be correctly mapped through it. In other words, the error should be minimized while defining a function for assigning the values of in-text citation frequencies over Strong, Medium, and Weak relationships. Thus, the error E is shown in Eq. (3).

$$\mathrm{E} = \sum \left(ax_i^2 + bx_i + c - f_i(x)\right)^2 \tag{3}$$

The error function depends on "a", "b" and "c", and hence its variation with respect to these parameters should be equal to zero. The objective is to minimize the function defined in 3. Now, the error with respect to the constant C, can be determined as shown in Eqs. (4) and (5).

$$\frac{\partial E}{\partial c} = 0 => a\sum x_i^2 + b\sum x_i + \sum c_i = \sum f_i \tag{4}$$

$$=> a\sum x_i^2 + b\sum x_i + cn = \sum f_i \tag{5}$$

As we have considered only "n" points, the rate of change in error can be computed with respect "b" and "a", and the final equations are shown in Eqs. (6) and (7).

$$a\sum x_i^3 + b\sum x_i^2 + c\sum x_i = \sum (f_i)(x_i) \tag{6}$$

$$a\sum x_i^4 + b\sum x_i^3 + c\sum x_i^2 = \sum (f_i)(x_i^2) \tag{7}$$

Now equations are ready, it's the time to compute the values for the required components in each equation. These components are:

$x_i, f_i, x_i^2, x_i^3, x_i^4, f_i x_i$ *and* $f_i x_i^2$

The list of values of each component for strong relationship is shown in Tab. 5.

Table 5: The computed values for strong relationship in Fig. 7

x_i	f_i	x_i^2	x_i^3	x_i^4	$f_i x_i$	$f_i x_i^2$
1	0.02	1	1	1	0.02	0.02
2	0.14	4	8	16	0.28	0.56
3	0.23	9	27	81	0.69	2.07
4	0.34	16	64	256	1.36	5.44
5	0.77	25	125	625	3.85	19.25
15	1.5	55	255	979	6.2	27.34

Using the data of Tab. 5, the system of equations from Eqs. (5), (6), and (7) would become as below:

$$55a + 15b + 5c = 1.5$$

$$225a + 55b + 15c = 6.2$$

$$979a + 225b + 55c = 27.34$$

These equations can be solved in various methods to compute the values for a, b, and c.

$$f(x) = 0.04571428542x^2 - 0.1042857124x + 0.10999999995 \tag{8}$$

Thus, the final equation for the strong relationship between citing and cited article becomes as shown in Eq. (9).

$$f(x) = -0.08571428571428571x^2 + 0.5562857142857143x - 0.414 \tag{9}$$

This is a quadratic increasing function for any value of x. However, the increasing effect will eventually stop as in-text citation frequencies cannot be infinite. For Moderate relationship, we have the following system of equations:

$$55a + 15b + 5c = 1.56$$

$$225a + 55b + 15c = 5.1$$

$$979a + 225b + 55c = 18.48$$

In the similar fashion, the formulation for Weak relevance is computed and it is shown in Eq. (10).

$$55a + 15b + 5c = 1.93$$

$$225a + 55b + 15c = 3.65$$

$$979a + 225b + 55c = 9.09$$

$$f(x) = 0.05x^2 - 0.514x + 1.378 \tag{10}$$

This equation is quadratic decreasing function. In theory, it will work for an infinite number of values of in-text citation frequencies. However, as already discussed, these values (i.e., in-text citations of a cited article) cannot be infinite. Thus, the decreasing effect will stop when in-text citation frequency reaches the lower value of a cited article. These derived equations are in support of the formulated hypothesis, i.e., the higher in-text citation frequencies of a paper correspond to a strong relationship, whereas, lower in-text citation frequencies depict a weak correlation between citing and cited articles.

In the following section, a comparison of the proposed approach with state-of-the-art techniques has been presented.

3.3 Comparative Analysis of the Proposed Approach

This section presents a comparative analysis of the proposed approach. The comparison is performed based on the nDCG with state-of-the-art approaches such as Content, Bibliographic, and Metadata-Based Models.

The nDCG value was computed for each query document using Eq. (11) and then the results were normalized with the help of Ideal Discount Cumulative Gain (IDCG) as shown in Eq. (12). In the first equation, i.e., Eq. (11), the ranking value for citing and cited papers was computed whereas, in the second equation (i.e., Eq. (12)), the ideal ranking value was calculated for citing and cited papers' pairs. Finally, all the results were averaged to get a single value.

$$DCG_p = \sum_{i=1}^{p} \frac{2^{\mathrm{Relevance_i}} - 1}{\log_2(i+1)} \tag{11}$$

$$\mathrm{nDCG_p} = \frac{\mathrm{DCG_p}}{\mathrm{IDCG_p}} \tag{12}$$

In following sections, detail results are discussed for each Models.

3.3.1 In-Text Citation Frequencies' Based Results

For this experiment, selected references were ranked based on in-text citation frequencies. Thus, citations having higher in-text citation frequencies were ranked at the top, whereas the references with lower in-text citation frequencies were placed at the bottom of the list. Various disjoint sets of queries such as nDCG@5, nDCG@10, nDCG@15, nDCG@20, and nDCG@25 were prepared and executed. The nDCG values for a different sets of query documents are shown in Fig. 6. The overall results for different set of queries are stable and there is no considerable variations.

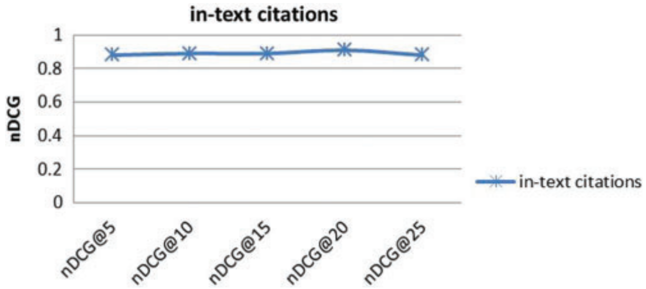


Figure 6: nDCG values of the proposed approach

One noteworthy point of the results is that a situation of a tie may occur while mapping two citation pairs of the same in-text citation frequency over different classes i.e., strong, medium, and weak citations. In the total 400 citation pairs, such situations were carefully observed and a total of 44 instances were found where in-text citation frequencies between cited and citing papers were same. Furthermore, out of these 44 citation pairs, 35 cases were classified by the users in the same category, whereas disagreement was recorded on 7instances between the users. Therefore, only 2 cases were found in which they were classified into different categories; therefore we chose to rank the lowest group on the top, to avoid the biases towards the proposed technique.

3.3.2 Content-Based Results

The "Content" based recommendations are sometimes referred to as word-level similarity in literature. The word-level likeness is used by more than 53% of the researchers who worked in the area of research paper recommendations. For computing word-level similarity, the abstracts of cited and citing papers were retrieved. The abstracts of the selected dataset were then indexed using the apache Lucene platform.

The reason for selecting the Lucene is that it provides a proven, robust, and scalable indexing and retrieval functionality and has been used by many other techniques in this domain. The Lucene accepts documents like a basic unit of information that is used for indexing, storage, and retrieval. The TF-IDF term vectors were acquired for all the papers in the selected dataset i.e., 400 annotated pairs of citation. Finally, cosine similarity was applied to compute document similarity. The Lucene provides support for extracting terms from the indexed documents. By default, Lucene excludes stop words such as "the," "is," and "and," etc. while retrieving terms. The "Content" based Model produced many recommendations for each of the source paper (higher recall). It is considered the strength of "content" based systems that they require only two documents to compute relevancy between them. Once these values were calculated to produce the ranked list. This rank list was normalized with the help of gold standard ranking. The nDCG values for different sets of queries are shown in Fig. 7. The average nDCG value of 0.65 was recorded for this technique for a different set of queries.

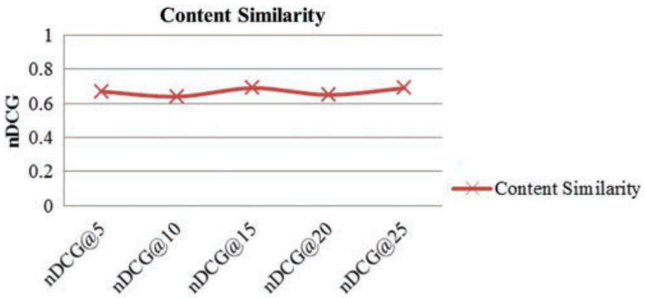


Figure 7: nDCG values of content similarity approach

3.3.3 Bibliographic Analysis Based Results

Bibliographic Coupling and Co-citation Models are widely known citation techniques for the identification of relevant documents. The Co-citation-based recommendations can vary over time, as in the future other documents can co-cite those papers. Therefore, in our current study, the possible choice was Bibliographic Coupling. Thus, standard references between citing and cited papers were automatically computed using the edits distance algorithm. Furthermore, they were cross verified manually.

Afterward, the relevant documents were ranked based on several frequent references between citing and cited papers. The relatively low nDCG value was recorded for this technique, which was 0.54. Furthermore, the nDCG values for a different set of queries are shown in Fig. 8. Papers may not have common references, and thus recommendations based on bibliographic coupling may also fail to provide any suggestions, and such limitations will affect the overall recall of the system. For example, in our case, 35% of the time, we have not found any bibliographically coupled paper.

3.3.4 Metadata Based Results

In this experiment, the most relevant papers were identified with the help of different metadata, such as papers' title terms matching, keywords matching, and papers' authors matching.

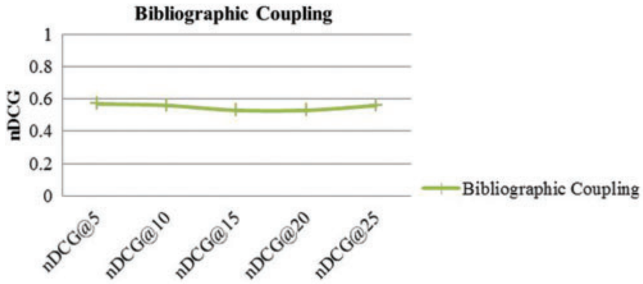


Figure 8: nDCG values of bibliographic coupling approach

The titles of the citing and cited papers were extracted and tokenized based on white spaces. Afterward, stop words (i.e., "for", "a", "an", etc.) were removed. Furthermore, the filtered terms were stemmed using porter stemming algorithm [19]. Along with titles of the paper, some other metadata parameters were also extracted, such as authors of the article and paper's keywords. The purpose of using multiple type of metadata was to increase the total number of recommendations produced by this approach. Once the metadata of the papers was ready for experimentation, the results were produced. So, based on a successful comparison of any of the metadata e.g., paper's title, author, or keywords, would result in a recommendation. Furthermore, articles were ranked in descending order. Finally, the nDCG's values were averaged to compare it with the rest of the techniques. It was found that the gain of title+author+keywords based recommendation was around 0.51. The overall nDCG values for different sets of queries are shown in Fig. 9. All the research papers have titles; thus, title-based recommendations can provide recommendations for all the cases. However, when terms are not matched, the metadata-based technique does not give any guidance. It reflects the overall recommendations made by a different method, as shown in Fig. 10. A total of 60% recommendation were made using a title matching technique.

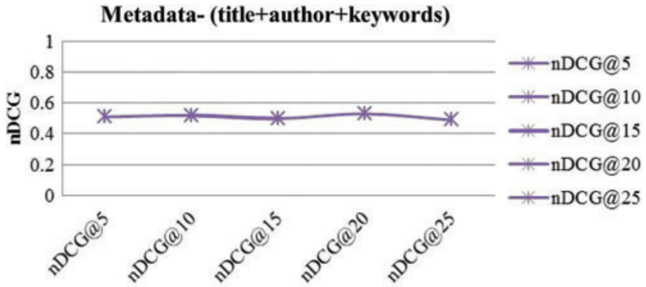


Figure 9: nDCG values of metadata approach

From these results, we may conclude that the proposed approach has higher gain as compared to the rest of the methods. The state-of-the-art techniques were tested against different sets of disjoint queries, and in the end, the results of these sets were averaged. There was no significant change reported in overall results across these sets of questions. Apart from the nDCG values, the total recommendations by different approaches in this experiment are also shown in Fig. 10. The in-text citation frequencies and "content" based techniques have a higher recall by providing recommendations for all possible instances. On the contrary, other methods such as Bibliographic

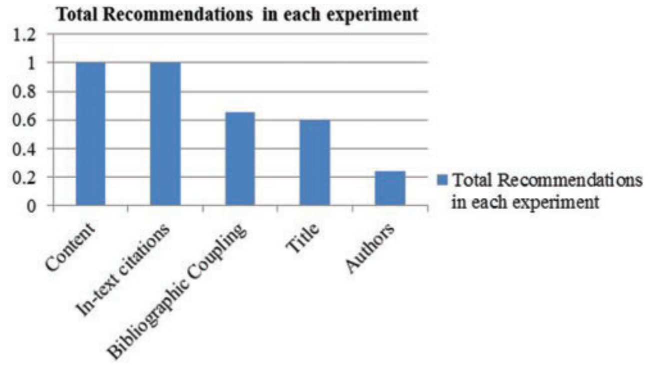


Figure 10: Total recommendations by each technique

Coupling, title terms matching, authors matching provided a fewer number of recommendations i.e., 65%, 60%, and 24%, respectively.

4 Discussion

The overall results indicated that in-text citation frequencies play a vital role in discovering degree-of-relevancy between the citing and cited papers. The overall relationship of in-text citation frequencies with degree-of-relevancy has shown in Fig. 6. The large numbers of weak connections were discovered for in-text citation frequencies = 1, which are 87%. On the other hand, many strong relationships for higher in-text citation frequencies >= 5 were reported, which are 77%. These results indicated that higher in-text citation frequencies corresponded to a more reliable connection between citing and cited papers. However, there were certain cases where in-text citation frequencies cannot classify citations independently. Therefore, there is a need to find some other useful features to enhance the overall results. One such functionality could be the in-text citation frequency distribution in different logical sections of a paper. Apart from this, below are certain other limitations in our work that should be resolved.

The TF-IDF based scheme was used for term's extraction; other techniques need to be tested in the future, for example, Yahoo's term extractor and KEA [20]. Thus, Metadata Based Model may be improved by integrating some other Models. For example, keywords can be extracted using some methods when they are not explicitly mentioned. Despite the benefits of in-text citation frequencies-based recommendations, it may also add overhead for computation of accurate identification of in-text citation in body text of the paper. Another limitation of the study is the size and diversity of the dataset. First, it should be large enough, and it would be encouraging to extend this study based on the various disciplines because the citation's behavior may differ among the various disciplines.

In a nutshell, despite these considerations, the in-text citation frequencies, Content-based, Metadata, and Bibliography-based Models are complementary. The strength of the Content-based retrieval Model is that it can retrieve documents that are not even linked-up with each other using, for example, citation-link, whereas the in-text citation may help as reflected in the results for better recommendations.

5 Conclusion

In the literature, different citation-based techniques have been extended with the help of textual analysis to find out the relevant research papers. In-text citation analysis is a dominant approach to find related research papers. The in-text citation analysis provides more insights as compared to the surface level citation analysis techniques. In this study, we deployed an in-text citation analysis technique, which extends the scope of the Direct Citation Model to discover the nature of the relationship between scientific papers. The proposed work was aimed to find the degree-of-relevancy between citing and cited articles by categorizing the kind of relationship between scientific papers into three categories, i.e., weak, medium and strong. The results revealed that in-text citation frequencies play a pivotal role in the identification of degree-of-relevancy between research papers. This identification follows a quantitative pattern of in-text citations i.e., in-text citation frequencies. The pattern produced that a less quantity value reports to a weak relationship between citing and cited paper; whereas, a considerable quantity value corresponds to the strong relationship between citing and cited article for higher in-text citation frequencies. The outcomes of the proposed technique were compared with different state-of-the-art techniques. The comparisons revealed that the proposed extended Direct Citation Model provides better results than other contemporary models.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] K. D. Bollacker, S. Lawrence and C. L. Giles, "Discovering relevant scientific literature on the web," *IEEE Intelligent Systems and their Applications*, vol. 15, no. 2, pp. 42–47, 2000.

[2] G. Zhaoquan, C. Yinyin, W. Sheng, L. Mohan, Q. Jing *et al.*, "Adversarial attacks on content-based filtering journal recommender systems," *Computers Materials & Continua*, vol. 64, no. 3, pp. 1755–1770, 2020.

[3] A. Shahid, M. T. Afzal, M. Abdar, M. E. Basiri, X. Zhou *et al.*, "Insights into relevant knowledge extraction techniques: A comprehensive review," *Journal of Supercomputing*, vol. 76, no. 1, pp. 1–39, 2020.

[4] M. M. Kessler, "Bibliographic coupling between scientific papers," *American Documentation*, vol. 14, no. 1, pp. 10–25, 1963.

[5] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the American Society for Information Science*, vol. 24, no. 4, pp. 265–269, 1973.

[6] W. R. Hou, M. Li and D. K. Niu, "Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution," *BioEssays*, vol. 33, no. 10, pp. 724–727, 2011.

[7] A. Shahid, M. T. Afzal and M. A. Qadir, "Discovering semantic relatedness between scientific articles through citation frequency," *Australian Journal of Basic and Applied Sciences*, vol. 5, no. 6, pp. 1599–1604, 2011.

[8] A. Riaz and M. T. Afzal, "CAD: An algorithm for citation-anchors detection in research papers," *Scientometrics*, vol. 117, no. 3, pp. 1405–1423, 2018.

[9] J. Cole and S. Cole, "Measuring the quality of sociological research. problems in the use of the science citation index," *American Sociologist*, vol. 6, no. 1, pp. 23–29, 1971.

[10] M. J. Moravcsik and P. Murugesan, "Some results on the function and quality of citations," *Social Studies of Science*, vol. 5, no. 1, pp. 86–92, 1975.

[11] K. W. Boyack, H. Small and R. Klavans, “Improving the accuracy of co-citation clustering using full text,” *Journal of the Association for Information Science and Technology*, vol. 64, no. 9, pp. 1759–1767, 2013.

[12] B. Gipp and J. Beel, “Citation proximity analysis (cpa): A new approach for identifying related work based on co-citation analysis,” in *Proc. ISSI*, Rio de Janeiro, Brazil, pp. 571–575, 2009.

[13] R. Klavans and W. B. Kevin, “Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge,” *Journal of the Association for Information Science and Technology*, vol. 68, no. 4, pp. 984–998, 2017.

[14] E. Garfield, “Can citation indexing be automated,” in *Proc. SAMMD*, Washington, DC, USA, pp. 189–192, 1965.

[15] C. L. Giles, K. Bollacker and S. Lawrence, “Citeseer: An automatic citation indexing system,” in *Proc. DL*, Pittsburgh Pennsylvania, USA, pp. 89–98, 1998.

[16] A. Constantin, S. Pettifer and A. Voronkov, “Pdfx. fully automated pdf-to-xml conversion of scientific literature,” in *Proc. DocEng*, Florence, Italy, pp. 177–180, 2013.

[17] A. Shahid, M. T. Afzal and M. A. Qadir, “Lessons learned: The complexity of accurate identification of in-text citations,” *International Arab Journal of Information Technology*, vol. 12, no. 5, pp. 481–488, 2015.

[18] S. Teufel, A. Siddharthan and D. Tidhar, “Automatic classification of citation function,” in *Proc. EMNLP*, Sydney, Australia, pp. 103–110, 2006.

[19] M. F. Porter, “An algorithm for suffix stripping,” *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[20] S. Jones and G. W. Paynter, “Automatic extraction of document keyphrases for use in digital libraries evaluation and applications,” *Journal of America Society of Information Science and Technology*, vol. 53, no. 8, pp. 653–677, 2002.