Computers, Materials & Continua
DOI:10.32604/cmc.2020.012619
Article

Tech Science Press

Exploiting Structural Similarities to Classify Citations

Muhammad Saboor Ahmed* **and Muhammad Tanvir Afzal**

Capital University of Science and Technology, Islamabad, Pakistan
*Corresponding Author: Muhammad Saboor Ahmed. Email: saboor@cust.edu.pk
Received: 07 July 2020; Accepted: 30 August 2020

Abstract: Citations play an important role in the scientific community by assisting in measuring multifarious policies like the impact of journals, researchers, institutions, and countries. Authors cite papers for different reasons, such as extending previous work, comparing their study with the state-of-the-art, providing background of the field, etc. In recent years, researchers have tried to conceptualize all citations into two broad categories, important and incidental. Such a categorization is very important to enhance scientific output in multiple ways, for instance, (1) Helping a researcher in identifying meaningful citations from a list of 100 to 1000 citations (2) Enhancing the impact factor calculation mechanism by more strongly weighting important citations, and (3) Improving researcher, institutional, and university rankings by only considering important citations. All of these uses depend upon correctly identifying the important citations from the list of all citations in a paper. To date, researchers have utilized many features to classify citations into these broad categories: cue phrases, in-text citation counts, and metadata features, etc. However, contemporary approaches are based on identification of in-text citation counts, mapping sections onto the Introduction, Methods, Results, and Discussion (IMRAD) structure, identifying cue phrases, etc. Identifying such features accurately is a challenging task and is normally conducted manually, with the accuracy of citation classification demonstrated in terms of these manually extracted features. This research proposes to examine the content of the cited and citing pair to identify important citing papers for each cited paper. This content similarity approach was adopted from research paper recommendation approaches. Furthermore, a novel section-based content similarity approach is also proposed. The results show that solely using the abstract of the cited and citing papers can achieve similar accuracy as the state-of-the-art approaches. This makes the proposed approach a viable technique that does not depend on manual identification of complex features.

Keywords: Section-wise similarity; citation classification; content similarity; important citation

1 Introduction

Researchers in all disciplines build upon the foundations laid by former researchers. This notion is succinctly summed up in Ziman's statement that "a scientific paper does not stand alone; it is embedded

in the literature of a subject" [1]. Research in the same field is interlinked, which means that existing research must always be brought in relation to former researches. New findings must be written up in the form of a scientific research paper. This research paper is then shared with other researchers so that the research process can be validated and can be continued. Therefore, while writing research findings, scholars acknowledge the scientific support they have received from former work. These acknowledgements are found in the reference section and termed as 'citations'. Ziman [1] and Narin [2] highlighted the true strength of analyzing citations can aid in producing and authenticating different research studies. They argue that the popularity and significance of a scientific work is expressecd through the frequency with which it is cited. The citations are considered an important tool for assessing the academic and scientific strength of institutions and individuals. They can also be used to investigate authors' or institutions' reputations within the overall scientific community [1,2].

The utility of citation-based measures is multifaceted. They are used to decide award nominees such as the Nobel prize [3] as well as research funding [4]. They can also be used to evaluate peer judgments [5] rank researchers [6–9] and countries [10]. In the late 1960s, Garfield, the founder of Thomson ISI, defined a number of reasons for citations [11,12]. This definition offers numerous opportunities to critically investigate citation behaviour [13,14].

Although citations are included to achieve specific objectives, citation count approaches [15] have never tried to distinguish between these objectives. Consequently, such approaches fail to maintain a balance between the act of citation itself and the purpose for which a citation is made. Instead, they blindly consider all citations equal. This discrepancy has led to achieve research in this area [16,17]. A detailed examination of citation counts was carried out by Benedictus et al. [18]. They concluded that citation count-based measures have inherited problems for example, they shift focus from quality to quantity.

Researchers have developed recommendations for improving the quality and reducing the emphasis on quantity in citation counts [19,20]. Generally, researchers believe that the reasons for citations must be critically considered in order to acknowledge the quality of different scholars' work [21]. Is it possible to differentiate between various reasons for citations? Existing citation annotation approaches proceed manually. The manual approaches rely upon interviewing the citer. Usually, authors are interviewed to share the reason for citing a particular piece of work on two different occasions: after the publication process is over and while writing the article [22,23].

Finney [24] argues that the citation classification process can be automated. Confirming this argument, later researchers took steps to classify citations into various categories [21,25]. However, while this idea made a significant contribution, it also brought a discouraging element to the fore, as citations were classified based on several ambiguous reasons. As a result of this ambiguity, the major limitation of a simple citation count approach was not effectively addressed. Presently, two major types of citations have been identified: important and un-important classes [16,17,26].

What do we mean by important and un-important classes? Generally, during the process of writing a paper, only a few citations in the reference list have a significant impact on the citing study. This impact needs to be precisely described. Zhu et al. [17] has provided a solution by arguing that an influential research study convinces the research community to adopt or extend the presented idea [17]. To establish a clear distinction between important and un-important citations, we need to examine contemporary citation classification mechanisms. Garzone et al. [25] extends the work of Finney [24] by implementing her suggestion of "associating cue words with citation function and using citation location in the classification algorithm" [25]. Both [16] and [17] argue that the citation relations discussed by Finney [24] and Garzone et al. [25] are important. In contrast, Garzone et al. [25] also cite several other studies as background information, such as the citation categories introduced by Garfield [12]. Based on the aforementioned discussion, the studies [16,17] classify citations into two major categories. The first

category of citations aims to provide background knowledge, which forms the foundation of the proposed study. Researchers such as Zhu et al. [17] have termed this category as "non-influential and incidental", whereas Valenzuela et al. [16] have termed it as non-important and incidental. We use the term "non-important" for this category. The second category of citations seeks to extend or apply the cited work. This category is termed as "influential" by Zhu et al. [17] and "important" by Valenzuela et al. [16]. We use the term "important" for this category.

Researchers have recently proposed different features and strategies to identify the important categories. For example, Valenzuela et al. [16] evaluated 12 features and concluded that in-text citation count was the most accurate feature, with a precision of 0.65. However, identifying citation tags from research papers is a challenge [27]. The Valenzuela's approach was further extended recently by Nazir et al. [28] wherein in-text citation counts within different logical sections of the paper (Introduction, Related Work, Methodology, and Results) were examined. This approach has achieved a precision of 0.84. However, there are also two major issues with this approach: (1) Accurately identifiying logical sections and mapping section headings onto the logical sections, and (2) Accurately identifying in-text citations [27]. The best-known approach for mapping section headings onto section categories has an accuracy of 78% [29]. Another binary citation classification approach presented by Qayyum et al. [26] has achieved a precision of 0.72 by examining metadata and cue phrases. However, this approach again involves the construction of cue phrases and identification of in-text citation frequencies. All of these recent approaches have certain limitations resulting from their reliance on the accurate identification of the following parameters: in-text citation counts, an updated dictionary of cue phrases, in-text citation extraction from sentences, and mapping section headings to logical sections. The extraction accuracy of each of the above parameters is around 70% [27,29]. However, the above approaches extract these parameters in a semi-automatic way, which has been demonstrated to be accurate when the parameters are readily and accurately available.

This critical discussion highlights the need for an approach that does not involve such a complex extraction of parameters, which is often inaccurate. Examinination of the relevant related literature shows that a content-based approaches were successfully employed by 55% of around 200 papers applied in the research paper recommendation domain in the last 16 years [30]. This motivated us to evaluate the suitability of the content-based approach for identifying important citations. Moreover, in addition to evaluating the existing content-based approach, this paper further proposes a novel section-wise content-based approach. The results indicate significant precision and recall values without any manual identification of complex parameters. The study's in-depth analysis of papers' complete content and content within different sections suggests that content-related similarities in the abstracts of the cited and citing papers be used to classify the citing paper as an important/non-important citation for the cited paper. Therefore, the proposed approach has a great potential to be applied in citation indexes and open new horizons for future researches in citation classification.

2 Literature Review

The citation count is utilized to conduct various types of bibliometric analyses with multidimensional utilities. Such analyses have been used to build indexing systems [31,32] and formulate various academic policies and present awards such as Noble prizes [3]. These analyses have also been used to rank researchers [6–9] and countries [10]. However, researchers believe that all citations cannot be conisedered equal, and each citation should be treated according to its true standing [1,14,16,17,21,33].

A specific study might be cited for myriad reasons. Garfield [12] was the first researcher to analyze citation behaviour [12]. He identified 15 citation reasons by examining various factors such as the citation's location in the paper and scrutinizing the differences and patterns. Some of the reasons include (1) acknowledging the contributions of predecessors, (2) highlighting fundamental contextual details,

(3) extending existing work and targeting expanded objective(s), etc. Later, Liptez [34] identified various classes of citations [34]. However, while both studies appropriately conceptualized the notion of citation reasons, no statistical measures were introduced [14]. Nevertheless, despite this shortcoming, these studies attracted enormous attention from the research community. Consequently, many empirical investigations have been carried out to identify citation reasons. Subsequently, other studies have attempted to capture actual citation behaviour [13,35]. However, common to all of these approaches is the treatment of all citations as being on the same level of importance.

According to Zhu et al. [17], studies involving straightforward quantitative citation analysis can be enhanced by eliminating incidental citations from the citation count. Additionally, maintaining a list of only important citations can be of substantial help for scholars seeking to identify influential studies on a specific topic. Until the mid-1990s, citation reasons were manually identified. For example, a general trend at that time was to interview authors during the process of writing an article or after their proposed article had been formally published, requesting them to describe the specific reasons for citing particular works [22,23]. However, differentiating scholars' citing behaviour using cognitive approaches seemed rather impractical. Therefore, researchers have realized the need for an automated system to identify and classify citation reasons.

Finney [24] demonstrated that the citation classification process can be automated. She created a citation function on an experimental basis. Later, she posited a relationship between cue words and citation location and combined it with the citation function. Though her approach was not fully automatic, however, it underscored the probability of developing a fully automatic citation classification mechanism in the future [25]. However, other researchers were a bit reluctant to acknowledge Finney's contribution because it was a doctoral thesis rather than a formal publication [14].

Drawing inspiration from Finney's approach, Garzone et al. [25] took their place among the trendsetters by creating an automated citation classification system. The authors argued that Finney's approach had several limitations, which they addressed by creating 35 categories to classify citations. They were able to successfully solve the classification task by introducing 14 parsing rules and 195 lexical matching rules. The dataset comprised 9 biochemistry and 11 physics articles. The system found to be stable. It produced average results on unseen articles and appreciative results on previously seen articles. Though the system produced encouraging results, there was also one concern: due to many classes, the system was unable to neatly distinguish between divergent classes. Pham et al. [36] classified citations from 482 citation contexts into four categories. They employed the ripple-down rules (RDR) hierarchy using cue phrases.

Our proposed model is similar to Zhu et al. and Valenzuela et al. [16,17]. Both studies have performed a similar binary citation classification based on the same distinction between important and non-important citations as the present study. The classification task was carried out using (1) in-text count-based features, (2) similarity-based features, (3) context-based features, (4) position-based features, and (5) miscellaneous features. The fundamental idea behind the proposed technique was to recognize the set of references that has an academic influence on the citing paper.

Zhu et al. [17] characterized academic influence as a reference that serves as a source for extracting an idea, problem, method, or experiment. They generated a total of 3143 paper-reference pairs from 100 papers extracted from the Association of Computational and Linguistics (ACL) anthology. These pairs were annotated by the authors of the citing papers. In contrast, Valenzuela et al. [16] introduced a supervised classification approach to identify important and non-important citations. The authors have extracted 465 paper-citation pairs from the ACL anthology. Two domain experts annotated these pairs as either important and non-important citations. Inter-annotator agreement was between two annotators was 93.9%. Twelve features were used to classify the citations into important and non-important classes. These features include the total number of direct citations, the number of direct citations per section, the total

number of indirect citations, the number of indirect citations per section, author overlap, etc. Two different classifiers, random forest and support vector machine (SVM), were used to train these features. Both models attained 0.90 recall and 0.65 precision. Zhu's approach [17] was criticized by Valenzuela et al. [16]. The latter claimed that biased annotation cannot be ruled out if citations are coded by the citing authors.

Another approach was proposed by Qayyum et al. [26]. They utilized metadata and cue terms to discover important citations. However, a limitation of this approach is that cue terms are identified from the papers' content and thus need to be updated for different datasets and domains. There is a need for a domain expert to manually identify cue terms for each domain and keep them updated. Nazir et al. extended Valenzuela's approach by identifying suitable weights for in-text citation frequencies in different sections. This approach has outperformed the previous approaches. However, a critical examination highlights limitations regarding accurately mapping section headings onto logical sections and accurately identifying in-text citations. Although there are many approaches to identifying important citations [16,26,28], in order to practically apply those approaches, there is a need to accurately identify the following information: (1) Accurately identify in-text citations [16,28], (2) Accurately map section headings onto logical sections [28], and (3) Create an updated accurate list of cue terms [26]. Existing approaches have reported precisions up to 0.84. However, these approaches depend on the accurate extraction of the above parameters and either ignore inaccurate results and correct the missing values manually to demonstrate the power of these parameters. The automatic extraction of such parameters is still a challenge [27,29]. This motivates us to fill this research gap by creating a novel apprach that does not require these parameters to be extracted. A critical examination of the related domain of research paper recommendations motivated us to use the content of the cited and citing papers. A survey paper by Beel et al. [30] indicates that more than 55% of the more than 200 articles on research paper recommendation in the last two decades used a content-based filtering approach.

Therefore, this research applies two types of content-based filtering methods. Firstly, the complete content of both the citing and cited paper is used to categorize the citing paper as an important/non-important citation for the cited paper. Furthermore, a novel section-based approach to citation classification is proposed. The results highlight the significance of the proposed approach.

3 Methodology

This research proposes identifying important citing papers for a cited paper by using the content of the pair (cited and citing paper). The content-based approach has been successfully applied in the last two decades for relevant research paper recommendations [30,37]. Drawing inspiration from this research, this study evaluates two types of content-based comparisons between the citing and cited paper pairs. In the literature, the documents' entire content has been used to identify relevant papers. However, in this study, we not only adapt the standard content-based approach for the task of important citation identification, but also propose a novel approach termed as section-wise content-based similarity. Fig. 1 depicts the complete methodology proposed in this study. In the first step of Fig. 1 a benchmark dataset is selected which provides input for both approaches (content-based approach and section-wise content-based approach). The left side of Fig. 1 presents methodological steps of the section-wise content-based approach, which produces a similarity score for each section. The right side of Fig. 1 depicts the content-based approach, which produces an overall content similarity score between the two papers. Both approaches produce top-recommended papers. Thus, Fig. 1 depicts a state-of-the-art evaluation and comparison strategy. The following sections elaborate on each step of this process in detail.

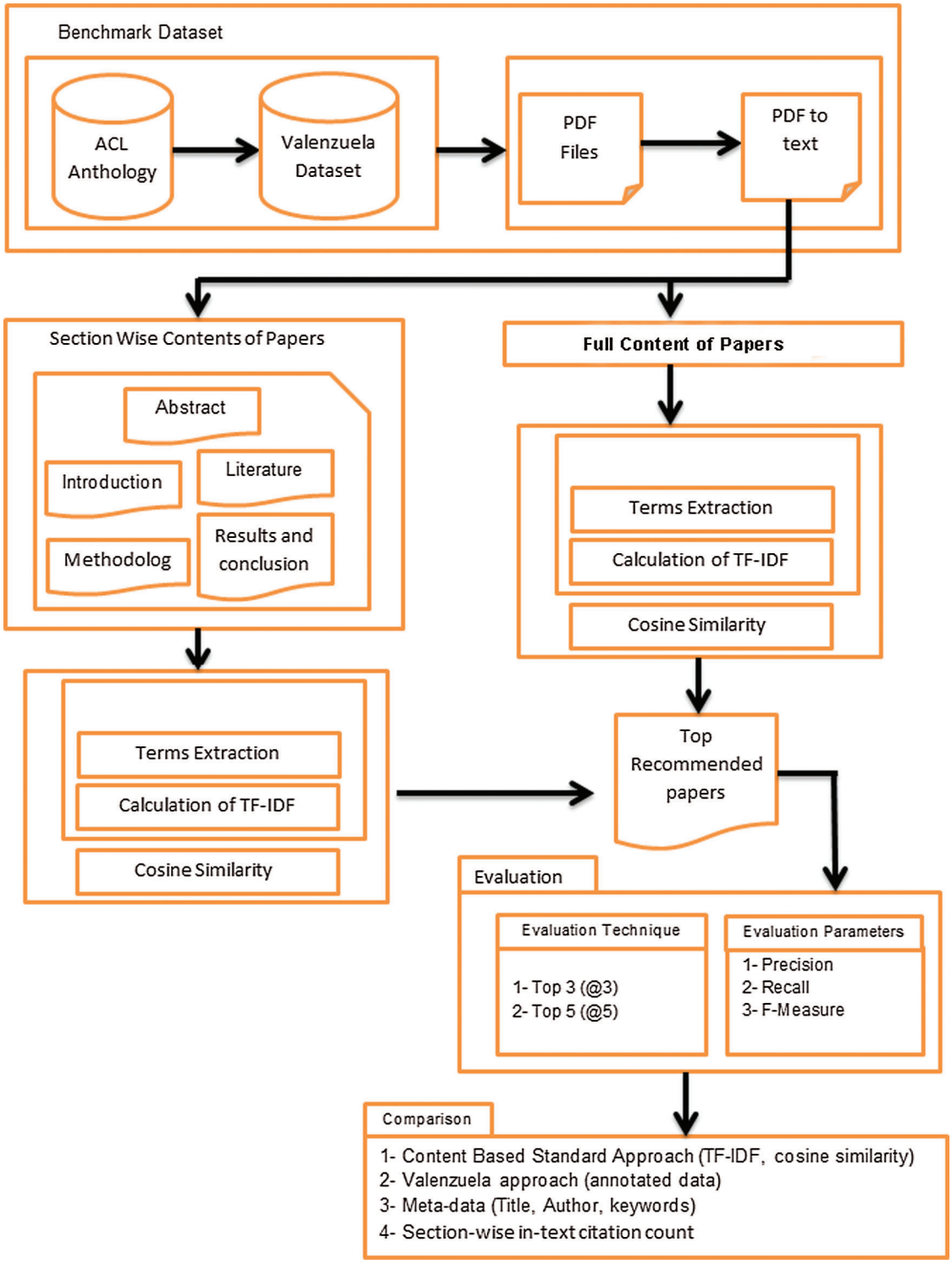


Figure 1: Methodological steps

3.1 Benchmark Datasets

The dataset selected to perform the experiments is the benchmark dataset developed by Valenzuela et al. [16]. This benchmark dataset is freely available online. The dataset stems from the field of information systems and encompasses 465 annotated paper-citation pairs collected from the Association of Computational and Linguistics (ACL) anthology. The ACL anthology is a digital archive of research papers in computational linguistics and a citation network containing only those papers and citations which are published in the ACL anthology itself. Tab. 1 provides a clear description of the dataset. The first column represents the two domain experts who annotated the dataset, denoted "A" and "B" in the Annotator column. The second column contains the source paper ID from the ACL anthology. The third column contains the IDs of the citing papers for the source paper. The fourth column "Follow-up" contains the score assigned by the annotators (i.e., 0 for incidental and 1 for important paper-citation pairs). The dataset also contains Portable Document Format (PDF) files, which were converted into text files to extract the full content and sections of the papers.

Table 1: Benchmark dataset

Annotator	Paper	Cited by	Follow-up
A	A00-1043	C00-2140	0
A	A00-1043	P02-1057	0
A	A97-1011	W09-1118	1
A	A97-1011	A00-2017	1
B	P05-1045	C10-1083	1
B	P05-1045	C10-1087	0
B	P05-1045	C10-1105	1
B	P05-1045	C10-1131	1

3.2 Content-Based Approach

The content-based approach is the most dominant method for the task of relevant paper recommendation [30,37]. In this study, we propose using the content-based approach to find important citing papers for each cited paper. The implementation steps for the content-based similarity approach are shown on the right side of Fig. 1. This study employs Lucene indexing. The Apache Lucene application programming interface (API) is considered the standard software for term indexing. It is widely used by researchers for indexing and finding content similarities [17]. For the extraction of important terms, the papers' full content are provided to the Apache Lucene API. Apache Lucene API indexes all terms within the content. Subsequently, the term frequency–inverse document frequency (TF-IDF) scheme is used to extract important terms from the indexed terms. The term extractor TF-IDF can be mathematically defined as given in Eq. (1). This equation is implemented for all citing and cited papers in the dataset. The basic idea of the TF–IDF technique is elaborated with the following example. For instance, the term T_1 frequently occurs in document D_1, but T_1 is not found frequently in the other documents D_2 to D_n. Thus, the conclusion is reached that term T_1 is the most important term for document D_1. Conversely, if any term T_2 frequently exists in all documents D_1 to D_n, it means that term T_2 is not important at all to distinguish documents.

$$tf - Idf(t,d,D) = tf(t,d) * Idf(t,d) \tag{1}$$

The next step is to measure the similarity of the papers' content. For this purpose, the cosine similarity technique is used. Eq. (2) shows the mathematical model for cosine vector similarity computation. The important terms extracted from document D1 are presented as vector A and the important terms from document D2 as vector B.

$$Content\_Similarity = \frac{A.B}{|A||B|} \tag{2}$$

Cosine similarity was computed for each document compared to all other documents. The generated similarity scores lie between 0 and 1. All text files received a similarity score using the cosine technique. After calculating the cosine similarity scores, the results were sorted in descending order to obtain a ranked list of the top 3 (T@3) and top 5 (T@5) recommended papers.

3.3 Section-wise Content Similarity Approach

The implementation steps for the section-wise content similarity approach are shown on the left side of Fig. 1. The Apache Lucene API was again used to index the terms. The similarities between corresponding sections of the papers were identified. To extract the important terms, the content of corresponding paper sections were provided to the Apache Lucene API. Apache Lucene API indexed all terms in the section. Then, the TF-IDF technique was used to identify the most important indexed terms. The term extractor TF-IDF can be mathematically defined as given in Eq. (3). This equation was implemented for all corresponding sections of the citing and cited papers in the dataset.

$$tf - Idf(t, s, S) = tf(t, s) * Idf(t, s) \tag{3}$$

In Eq. (3), "t" represents the important terms whereas "s" represents the section content of the cited paper and "S" the content of the corresponding section of the citing paper. The section-wise similarity technique is used to measure the similarity between corresponding sections. The mathematical model for section-wise similarity is given in Eq. (4). The vector V1s refers to the extracted important terms for section 's' of cited paper P1, while the vector V2s refers to the important terms for the corresponding section 's' of citing paper P2.

$$Section\_wise_{Similarity} = \frac{V1s.V2s}{|V1s||V2s|} \tag{4}$$

The section-wise similarity was computed for each section and was compared to all the corresponding sections. The generated section-wise similarity scores lies between 0 and 1. All text files received a similarity score using the section-wise similarity technique. After obtaining the similarity scores, the results were sorted in descending order to create a ranked list of the top 3 and top 5 recommended papers.

3.4 Evaluation Parameters

The proposed approach was evaluated using standard evaluation parameters used by state-of-the-art approaches, namely by (1) Valenzuela et al. [16], (2) Qayyum et al. [26], and Nazir et al. [28]. The evaluation parameters used by state-of-the-art approaches are precision, recall, and F-measure. The definition of each parameter is given below:

The formula to calculate the precision is shown in Eq. (5). The dataset contains citations classified as important and non-important. Precision in identifying important citation using the proposed technique is defined as the ratio of citations correctly classified as "important citations" to the total number of citations classified by the technique as "important citations".

$$Precision = \frac{relevant\ retrieved}{Total\ retrieved} \tag{5}$$

The formula to calculate the recall is depicted in Eq. (6). Recall in identifying important citations using the proposed technique is defined as the ratio of citations correctly classified as "important citations" to the total number of citations which are in actual fact "important citations".

$$Recall = \frac{relevant\ retrieved}{Total\ relevant} \tag{6}$$

The F-measure is the harmonic mean of precision and recall and is calculated as shown in Eq. (7).

$$F - measure = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{7}$$

The precision, recall, and F-measure for each cited paper were calculated against all of its citing papers by considering the classification presented in Valenzuela et al. [16]. Subsequently, the average precision, recall, and F-measure were calculated for the full dataset. Then, the precision, recall, and F-measures for the state-of-the-art approaches were taken from the original published papers [16,26,28], all of which worked on the same dataset.

4 Results and Comparisons

This section presents the analysis of results and comparisons for both the proposed approaches. The objectives of this section are twofold. First, we sought to identify the applicability of using content to identify important citations. This refers to both approaches, i.e., using the full content and evaluating individual sections. Secondly, we sought to compare the results with existing state-of-the-art approaches proposed by Qayyum et al. [26], Valenzuela et al. [16], and Nazir et al. [28]. Section 4.1 presents the results and evaluation of the two proposed approaches, whereas Section 4.2 compares the best results from the proposed approach with existing state-of-the-art approaches.

4.1 Results and Evaluation of the Proposed Approaches

Our first approach was to adapt a content-based filtering technique for citation classification. The second approach was to apply the same content-based technique to individual paper sections, namely the abstract, introduction, literature review, methodology, and results sections. Subsequently, we discuss which section plays the best role in classifying citations into two classes, important and non-important. Specifically, the following six similarity-based rankings were computed.

1. Full content similarity-based ranking
2. Abstract-section similarity-based ranking
3. Introduction-section similarity-based ranking
4. Literature section similarity-based ranking
5. Methodology-section similarity-based ranking
6. Results-section similarity-based ranking

In this section, the section-based similarity rankings (No. 2 to No. 6 above) will be compared to the full-content-based ranking (No. 1 above).

4.1.1 Evaluation of Proposed Six Rankings

This section presents the results for all six similarity rankings proposed in this research and listed above. The first ranking is a full-content similarity-based ranking. In this ranking, the full content of the cited document is taken and compared to the full content of the citing documents in the list. Similarity scores are calculated for comparison purposes. Afterwards, the similarity scores are sorted in descending order to rank the top 3 as well as top 5 citing documents for each cited paper. Then precision, recall, and F-measure scores are calculated. In the end, a cumulative F-measure for the top 3 and top 5 documents is calculated and compared to the F-measure for the top 3 and top 5 documents identified using the other similarity rankings listed above. The cumulative F-measure for the full-content similarity-based ranking was 0.63 for the top 3 documents and 0.65 for the top 5 documents, respectively. This is a significant result obtained by solely examining the content of research papers. It will be compared with existing state-of-the-art approaches in the next section.

The second ranking was produced by computing the similarity between the abstracts of the cited and citing documents. The cumulative F-measures for the abstract-section similarity-based ranking were 0.70 for the top 3 documents and 0.69 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. Comparisons between the abstract section similarity and full-content similarity are shown in Figs. 2a and 2b for the top 3 and top 5 ranked documents, respectively. The cited paper number is listed on the x-axis and the F-measure score on the y-axis. The red line shows the F-measure for each cited paper using the full-content approach, whereas the blue line shows the abstract section-based F-measure. For most of the cited papers, the red line and blue line follow the same path, meaning that both approaches produced the same results in these cases. It is also clear from Figs. 2a and 2b that when the results of abstract-based similarity and content-based similarity differ, abstract-based similarity produces more accurate results.

This result clearly shows that abstract-section similarity-based ranking outperforms full-content similarity-based ranking. This is because the abstract is a concisely written paper section of just a few hundred words in which the author has to explain the whole idea of the research paper, including motivation, research gap, state-of-the-art, research question, methodology, results, and comparisons.

Thus, the abstract has more descriptive power regarding the context and contribution of a research paper. In contrast, the full content of a paper encompasses many different sections, including the introduction, literature review, etc., which might not be feasible to compare and might not deliver such strong results. Research papers' abstracts are normally available for free for both citing papers as well as cited papers.

The third proposed ranking involves the introduction sections of the cited and citing papers. The cumulative F-measures for the introduction-section similarity-based ranking were 0.57 for the top 3 documents and 0.59 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. This result clearly shows that full-content similarity-based ranking outperforms introduction-section similarity-based ranking. The results are shown in Figs. 3a and 3b. An in-depth analysis of the terms identified as important for the introduction sections of some of the randomly selected papers illustrates the reason for such results. The introduction section is usually an extended version of the abstract. The general flow of the introduction section is as follows: (1) background of the problem, (2) existing state-of-the-art approaches, (3) research gap, (4) methodology, and (5) results and comparisons. Accordingly, most of the content in papers' introduction sections tends to be very similar, leading towards the citing paper to be considered an important citation for the cited paper. Moreover, this section is typically not very long.

The fourth ranking was produced by examining the content of the literature review sections of the cited and citing paper pairs. The cumulative F-measures for the literature section similarity-based ranking were 0.59 for the top 3 documents and 0.62 for the top 5 documents, respectively. Recall that the cumulative

F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. The results are shown in Figs. 4a and 4b. This result clearly shows that full-content similarity-based ranking outperforms literature-section similarity-based ranking. This is because the literature review section contains very generic terms to explain others' work. Every author has a unique way of writing the literature review section by explaining existing approaches in the respective research areas and critical analyzing the literature. Therefore, the literature review section is not significant for identifying contextual similarities between cited and citing pairs.

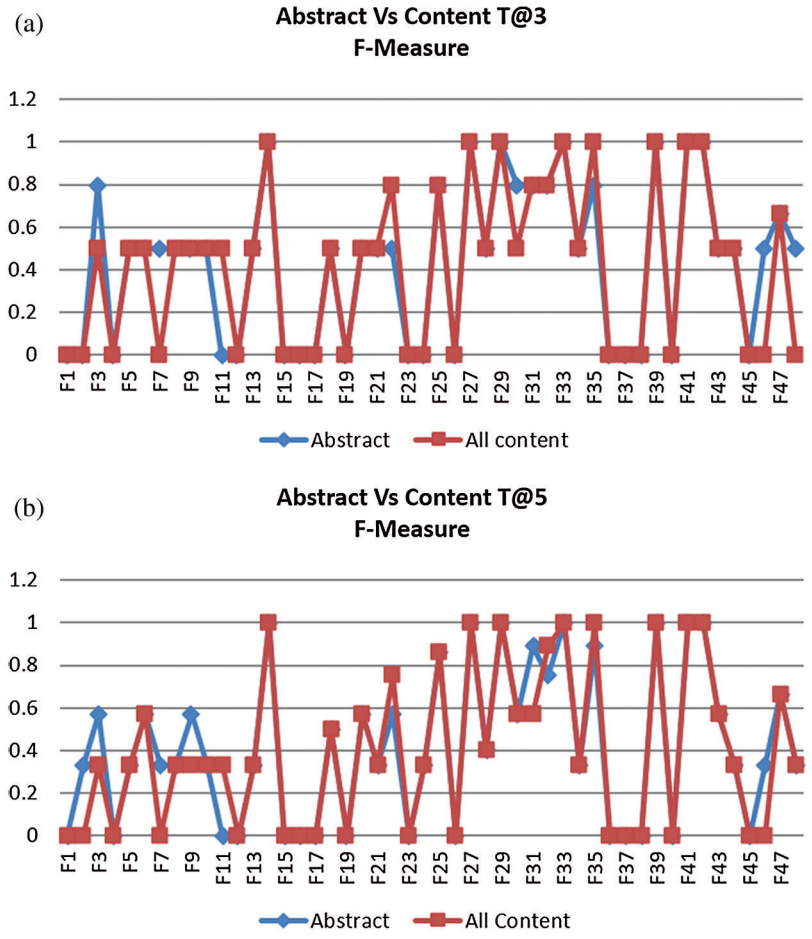


Figure 2: (a) Abstract *vs*. full content top 3 (b) Abstract *vs*. full content top 5

The fifth ranking was achieved by examining the content of the methodology sections of both cited and citing papers. The cumulative F-measures for the methodology-section similarity-based ranking were 0.72 for the top 3 and 0.66 for the top 5 documents, respectively. Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. The results are represented in Figs. 5a and 5b. This results clearly show that methodology-section similarity-based ranking outperforms full-content similarity-based ranking. The results demonstrate the expressive power of the cited paper's methodology section to identify important citations from the list of citing papers. The methodology section presents the study conceptualization of both papers, which involves the use of similar domain-related terms. Nevertheless, while the results for the methodology section were good, the results for the abstract were even better.

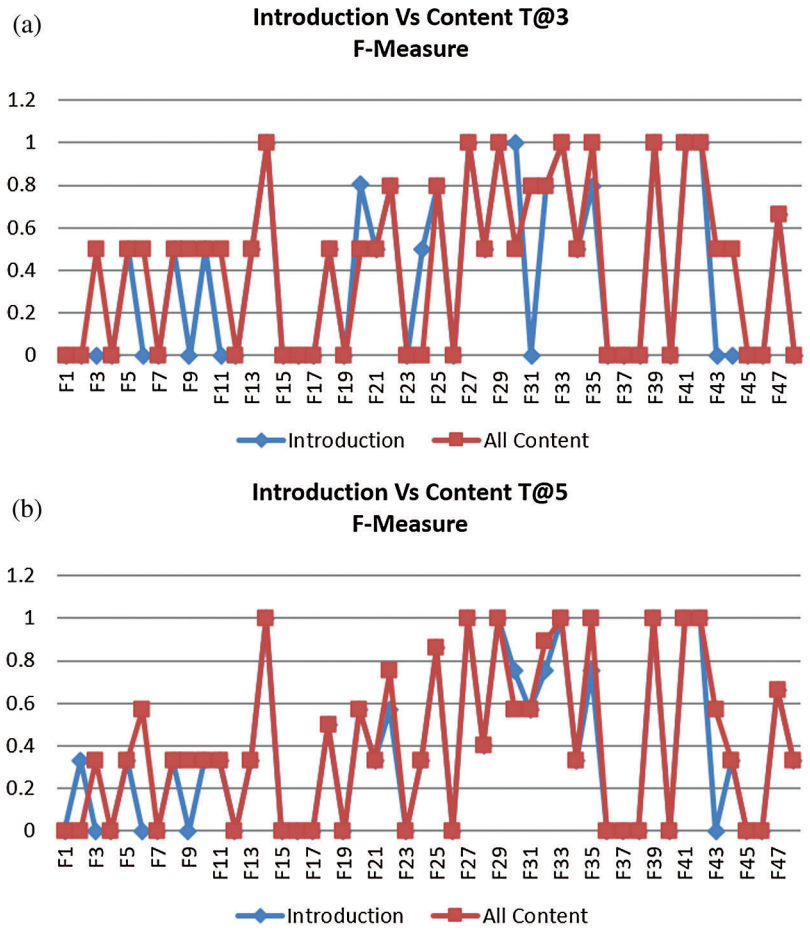


Figure 3: (a) Introduction *vs*. full content top 3 (b) Introduction *vs*. full content top 5

The sixth and final ranking proposed in this research is depicted in Figs. 6a and 6b. This ranking was achieved by comparing the content of the results sections of both papers. The cumulative F-measures for the results-section similarity-based ranking were 0.64 for the top 3 and 0.63 for the top 5 documents, respectively.

Recall that the cumulative F-measures for the full-content similarity-based ranking were 0.63 and 0.65, respectively. This result clearly shows that results-section similarity-based ranking and full-content similarity-based ranking are approximately equal. The results section is also an important section of a research paper, as it reports the important findings of both cited and citing papers. The results section similarity was found to be significant when: (1) both papers address the same topic and use a common vocabulary of terms for the specific domain, (2) both papers use the same dataset, (3) both papers apply similar evaluation metrics, and (4) both papers compare their results with the same/similar research papers.

This section has reported the results of the six proposed rankings. The full-content-based ranking was adapted from the domain of identifying relevant research papers [30,37]. Furthermore, this research proposed the novel approach of section-based similarity. Five further rankings could be calculated when applying this proposed approach. The content of the five major sections of research papers, namely abstract, introduction, literature review, methodology, and results, were systematically compared. The results indicate that examining content alone makes it possible to identify important citations for the cited paper from the list

of citing papers. Furthermore, the abstract, methodology, and results sections achieved similar or better results compared to the full content of research papers. Specifically, the abstract section outperformed the full content of research papers. Therefore, we conclude that papers' abstracts should be used to compute content-based similarity for two reasons: (1) better performance compared to the full-content approach, and (2) abstracts are normally freely available.

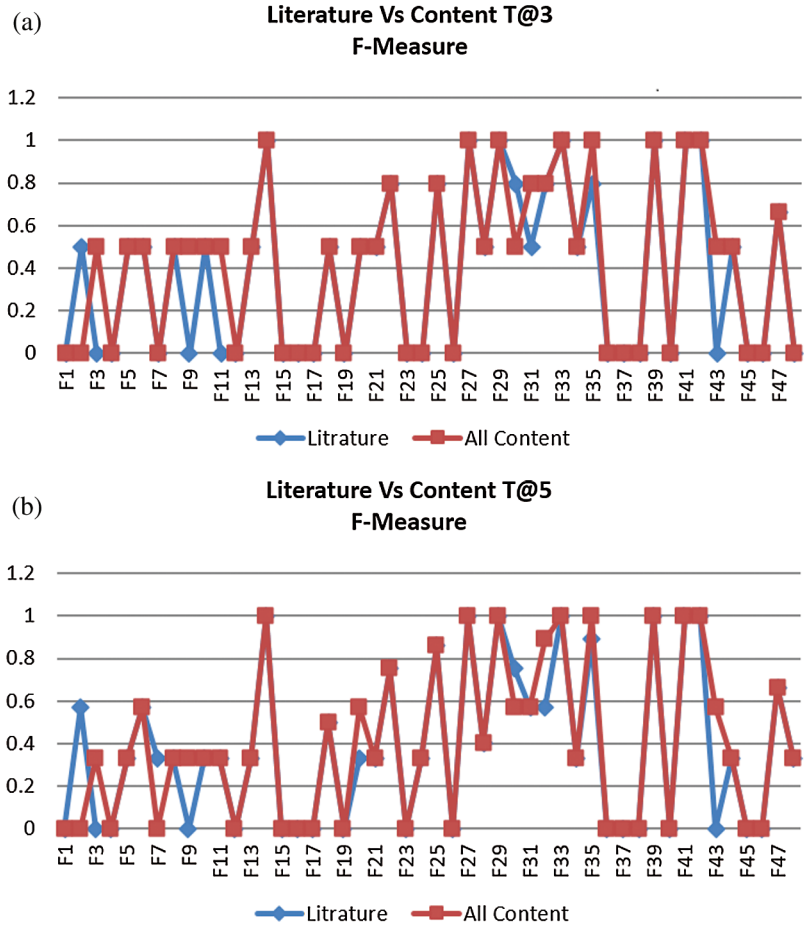


Figure 4: (a) Literature *vs*. full content top 3 (b) Literature *vs*. full content top 5

Figs. 2–6 present the results of the five proposed rankings based on comparing individual paper sections. Fig. 7 compares the results of six proposed ranking approaches. The complete content achieved the highest precision of 0.68; however, the abstract alone achieved a very close precision score of 0.66. Furthermore, the abstract alone was able to achieve a very high recall value of 0.94, second only to the methodology section. Due to this high recall, the abstract outperformed all other approaches in terms of F-measures. Therefore, out of the five proposed section-wise rankings, the abstract section was selected to be compared to existing state-of-the-art approaches in the following sections.

4.2 Comparison to State-of-the-Art Approaches

The previous section proposed six new rankings for identifying important citations for cited papers from the list of citing papers. Comparing abstracts was identified as the best ranking technique based on a critical analysis of the results for all six rankings. This section, in turn, compares the results of the best proposed

ranking to the best parameter rankings achieved by current state-of-the-art approaches. Comparisons of precision and recall are depicted in Figs. 8 and 9, respectively.

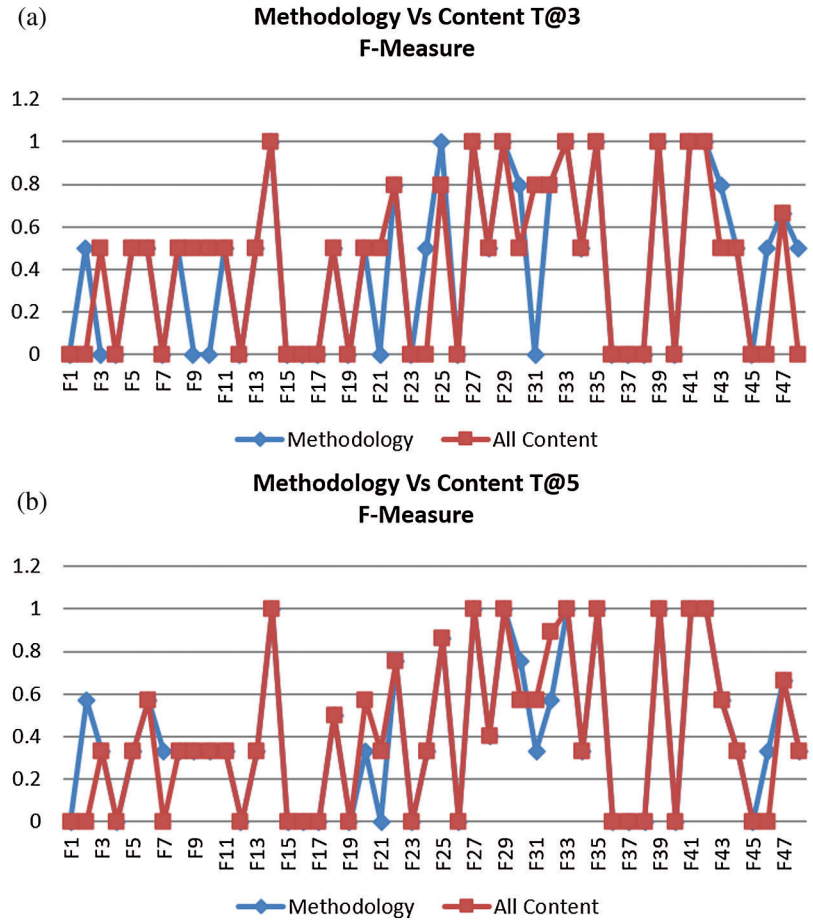


Figure 5: (a) Methodology *vs*. full content top 3 (b) Methodology *vs*. full content top 5

The proposed approach was compared to the following state-of-the-art approaches. The first approach was presented by Valenzuela et al. [16], who conducted the pioneering work in this area and have made their dataset freely available online. This is the same dataset used by the proposed approach and the other approaches represented in Figs. 7 and 8. Valenzuela et al. [16] tested 12 features as identifiers of important citations for the cited papers, with the in-text citation-based feature producing the best results.

The second state-of-the-art approach was proposed by Qayyum et al. [26]. They presented a hybrid approach the uses metadata and content-based features to identify important citations. The third state-of-the-art approach is the technique was proposed by Nazir et al. [28], who extended the approach by Valenzuela et al. [16]. They assigned weights to different sections of the paper to better capture the significance of in-text citation counts.

Fig. 8 compares the precision results for the newly proposed approaches and three existing state-of-the-art approaches. The x-axis lists the approaches' names and the y-axis the precision score. The results for proposed approach #1 (utilizing the full content of the cited-citing pair) is 0.68, and the result of proposed approach #2 (examining content similarity in the abstract sections of the cited-citing pair) is

0.66. Nazir et al. [28] achieved the maximum precision of 0.84, followed by Qayyum and Afzal with a precision score of 0.72. These are the best results from a variety of feature evaluations conducted within each study. The results seem to indicate that the proposed approach outperformed Valenzuela et al.'s [16] approach but was inferior to other state-of-the-art approaches. However, this is not in fact the case. To illustrate why, let us discuss the results of each approach one-by-one.

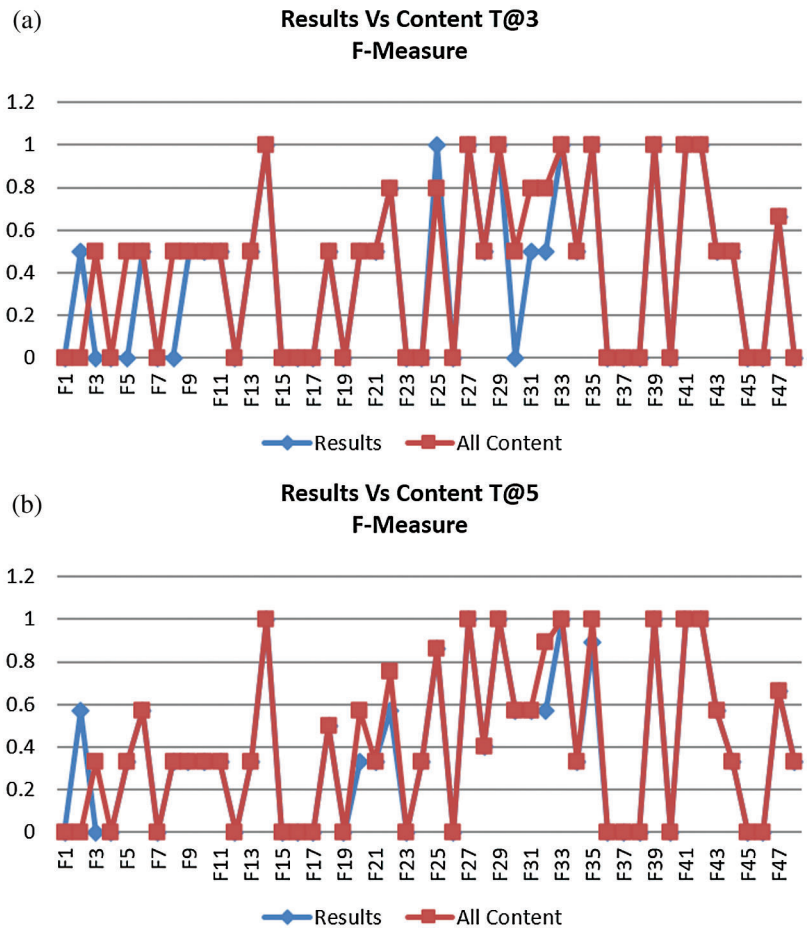
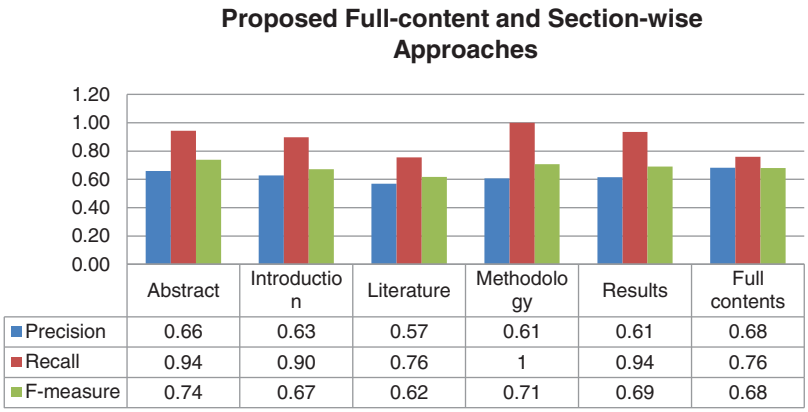


Figure 6: (a) Results *vs*. full content top 3 (b) Results *vs*. full content top 5



	Abstract	Introduction	Literature	Methodology	Results	Full contents
Precision	0.66	0.63	0.57	0.61	0.61	0.68
Recall	0.94	0.90	0.76	1	0.94	0.76
F-measure	0.74	0.67	0.62	0.71	0.69	0.68

Figure 7: Comparison of the six proposed rankings

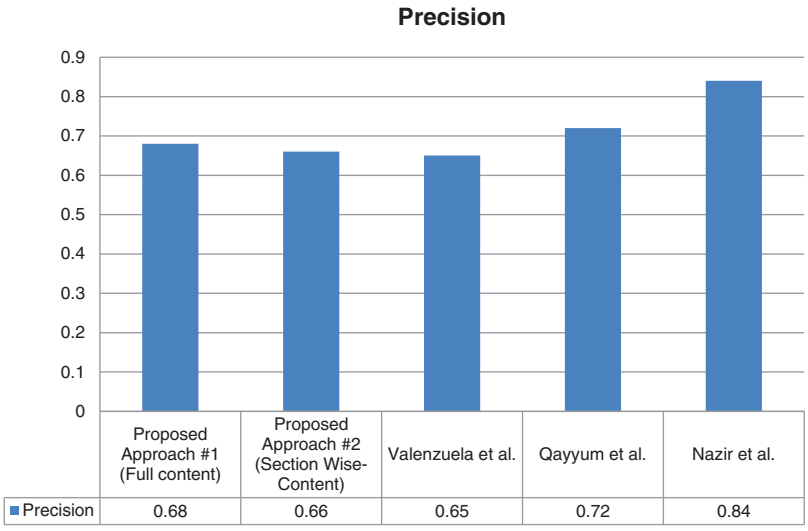


Figure 8: Comparing the precision of the proposed approaches with state-of-the-art rankings

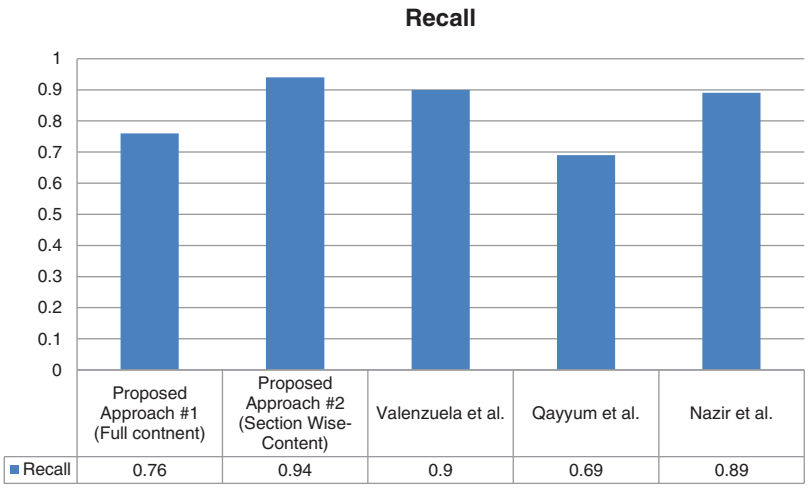


Figure 9: Comparing the recall of the proposed approaches with state-of-the-art rankings

Valenzuela et al. [16] achieved a maximum precision of 0.37 when emplyoing only a single parameter, namely "direct citations per section". When examining only a single parameter, the newly proposed approach focusing solely on the abstract achieved a precision score of 0.66, thus outperforming Valenzuela et al. In comparison, Valenzuela et al. achieved a precision score of 0.65 when aggregating all 12 parameters, still slightly lower than the precision score of 0.66 obtained by Proposed Approach #2. Furthermore, to compare this value, one needs to consider the following facts. Valenzuela et al. have not discussed how accurately they extracted the 12 features. For example, metadata features like keywords are only available around 50% of the time [26]. Furthermore, the accurate extraction of in-text citation counts is not a trivial task and requires very sophisticated algorithms. This has been pointed out by Shahid et al. [11], who achieved 58% accuracy in extracting in-text citations. Although an approach recently proposed by Ahmad et al. [38] raises this accuracy, it still needs to be verified on journals from diverse fields and different publishers' styles. Therefore, the precision score of 0.65 achieved by Valenzuela et al. is dependent on the accurate identification of in-text citation counts. If the approach by Valenzuela et al. [16] were to extract

in-text citation counts automatically using the procedures presented by Shahid et al. [11], the precision score might remain in the range of around 0.3. Standard tools such as Content ExtRactor and MINEr (CERMINE) [39] and GeneRation Of BIbliographic Data (GROBID) [40] could only achieve precision, recall, and F-measure scores in the range of 0.8 to 0.9 when evaluated by Ahmad et al. [38]. Thus, if Valenzuela et al. [16] were to apply the best automated approach to detecting in-text citations, the precision score for finding important citations would drop from 0.65 to less than 0.5. In contrast, the proposed approach does not require any such complex parameter computations; it is based solely on the content of the abstract, which is freely available. Therefore, in terms of real applications, the proposed approach outperforms Valenzuela et al.'s approach in terms of precision score and thus can be considered a viable solution for citation indexes and digital libraries.

The second state-of-the-art approach was presented by Qayyum et al. [26]. They classified citing papers as important/non-important citations for the cited paper using metadata and the papers' content. The best individual feature they examined achieved a precision score of 0.35. Thus, with respect to single features, the proposed approach utilizing the abstract alone outperforms Qayyum et al. [26]. However, when Qayyum et al. [26] aggregated four metadata elements, the precision score using the random forest classifier reached to 0.72. Important to consider here is that this score can only be obtained when all four metadata elements are available. For example, only 58.3% of Qayyum et al. [26] dataset included keywords. This approach is not applicable in the scenarios wherein metadata is not present in equal ratio. Furthermore, cue phrases need to be identified for each individual dataset. This makes the method impractical to use in real systems. In contrast, the newly proposed approach does not rely upon defining a cue phrase dictionary or the availability of keywords.

The third approach selected for comparison is the technique proposed by Nazir et al. [28]. They used section-based in-text citation frequencies to classify citations as important or non-important. A further novel element of this approach is their identification of suitable section weights using linear regression. The approach achieved a precision score of 0.84. However, the present comparison demonstrates the pitfalls of this state-of-the-art approach. Specifically, it is necessary to calculate in-text citation frequencies, which is quite challenging to perform automatically, as noted above. Another challenge concerns mapping section headings onto logical sections (such as Introduction, Literature, Methodology, Results and Discussion). Shahid et al. [29] achieved the highest accuracy for this task which is 78%. Considering all these factors, the proposed approach is comparable to the best-known existing approach as it does not require any complex calculations to be performed unlike other state-of-the-art approches.

The recall of both proposed approaches and existing state-of-the-art approaches is compared in Fig. 8. Proposed Approach #2 (section-wise similarity between abstracts) achieved the best recall of 0.94, higher than existing state-of-the-art approaches. Content-based approaches such as those used by search engines and citation indexes are considered the best approaches to obtain maximum recall. This means that 90% of the time, important citations are identified as such by the proposed approach, with some noise. The proposed approach not only achieves better recall, its implementation is also more viable for the following two reasons: (1) it does not require complex calculations of in-text citation frequencies, mapping section headings to logical sections, the availability of all metadata fields, or identifying cue phrases for each dataset, and (2) abstracts generally available for free online.

5 Conclusions

Identifying the set of important citations for the cited paper from the list of citing papers is a challenge that has led the scientific community to propose a wide range of techniques. This research has critically evaluated the literature and identified three state-of-the-art approaches to classifying citations into two classes, namely important and non-important. These existing approaches have utilized a different set of

features than the classification method proposed in this study. The precision of these state-of-the-art approaches range from 0.72 to 0.84. However, they are dependent on the accurate identification of some complex features, such as in-text citation frequencies, mapping section headings onto section labels, availability of metadata elements, and constructing dataset-dependent dictionaries of cue phrases. The values for the state-of-the-art approaches cited above are achieved only when all of these parameters are extracted accurately.

However, a critical analysis shows that the accuracy of identifying in-text citations varies from 58% to 90%, as highlighted by different research and state-of-the-art tools. The accuracy of mapping section headings onto logical sections is just 78%. Keyword metadata is available only 53% of the time. Cue phrases built for one dataset need to be developed anew for another dataset. Currently, state-of-the-art approaches extract such features in a semi-automatic way, and incorrect values are corrected manually. However, when all of these features are extracted fully automatically, the precision score drops to one-third of the reported values.

This paper presents a method that does not require the computation of such complex features. In the similar domain of identifying relevant research papers, papers' content has been successfully used for nearly two decades to identify relevant papers. Based on these findings, this paper adopted the content-based similarity approach to identify the similarity between pairs of cited and citing papers. Furthermore, a novel approach involving section-based similarity was proposed, implemented, and evaluated. An in-depth analysis of both proposed approaches indicated that the abstract alone is sufficient to decide whether the citing paper is important or non-important for the cited paper. The proposed approach achieved precision scores of 0.68 for full content and 0.66 for the abstract section, respectively, outperforming existing state-of-the-art approaches when considering the facts presented above. Furthermore, the recall of existing state-of-the-art approaches range from 0.7 to 0.9, while the proposed approach has achieved a recall score of 0.94. Thus, the proposed approach significantly outperformed existing approaches in terms of recall, particularly when considering that inaccurate calculations of in-text citations, section mapping, metadata availability, and cue phrase construction will significantly reduce recall scores for the state-of-the-art approaches when conducted automatically. In contrast, there is no need for such complex calculations in the proposed approach.

Acknowledgement: We acknowledge the support of Capital University of Science and Technology, Islamabad, Pakistan for providing us with the working environment to complete this research.

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] J. M. Ziman, *Public Knowledge: An Essay Concerning the Social Dimension of Science*. 1st ed., vol. 519. London: Cambridge University Press, 1968.

[2] F. Narin, *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity.* Cherry Hill, NJ: Computer Horizons, 206–219, 1976.

[3] H. Inhaber and K. Przednowek, "Quality of research and the Nobel prizes," *Social Studies of Science*, vol. 6, no. 1, pp. 33–50, 1976.

[4] R. C. Anderson, F. Narin and P. McAllister, "Publication ratings versus peer ratings of universities," *Journal of the American Society for Information Science*, vol. 29, no. 2, pp. 91–103, 1978.

[5] A. T. Smith and M. Eysenck, *The Correlation Between RAE Ratings and Citation Counts in Psychology.* [Departmental Technical Report] (Unpublished).

[6] S. Maqsood, M. A. Islam, M. T. Afzal and N. Masood, "A comprehensive author ranking evaluation of network and bibliographic indices," *Malaysian Journal of Library & Information Science*, vol. 25, no. 1, pp. 31–45, 2020.

[7] M. Raheel, S. Ayaz and M. T. Afzal, "Evaluation of h-index, its variants and extensions based on publication age & citation intensity in civil engineering," *Scientometrics*, vol. 114, no. 3, pp. 1107–1127, 2018.

[8] S. Ayaz and M. T. Afzal, "Identification of conversion factor for completing-h index for the field of mathematics," *Scientometrics*, vol. 109, no. 3, pp. 1511–1524, 2016.

[9] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proc. of the National Academy of Sciences*, vol. 102, no. 46, pp. 16569–16572, 2005.

[10] A. Mazloumian, D. Helbing, S. Lozano, R. P. Light and K. Börner, "Global multi-level analysis of the scientific food web," *Scientific Reports*, vol. 3, no. 1, pp. 1–5, 2013.

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

[12] E. Garfield, "Can citation indexing be automated," in *Statistical Association Methods for Mechanized Documentation, Sym. Proc.*, Washington, vol. 269, pp. 189–192, 1965.

[13] I. Spiegel-Rosing, "Science studies: Bibliometric and content analysis," *Social Studies of Science*, vol. 7, no. 1, pp. 97–113, 1977.

[14] L. Bornmann and H. D. Daniel, "What do citation counts measure? A review of studies on citing behavior," *Journal of Documentation*, vol. 64, no. 1, pp. 45–80, 2008.

[15] 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.

[16] M. Valenzuela, V. Ha and O. Etzioni, "Identifying meaningful citations," in *Workshops at the Twenty-ninth AAAI Conf. on Artificial Intelligence*, Palo Alto, California, 2015.

[17] X. Zhu, P. Turney, D. Lemire and A. Vellino, "Measuring academic influence: Not all citations are equal," *Journal of the Association for Information Science and Technology*, vol. 66, no. 2, pp. 408–427, 2015.

[18] R. Benedictus, F. Miedema and M. W. Ferguson, "Fewer numbers, better science," *Nature*, vol. 538, no. 7626, pp. 453–455, 2016.

[19] M. H. MacRoberts and B. R. MacRoberts, "The mismeasure of science: Citation analysis," *Journal of the Association for Information Science and Technology*, vol. 69, no. 3, pp. 474–482, 2018.

[20] J. Wilsdon, L. Allen, E. Belfiore, P. Campbell, S. Curry *et al., The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. Springer Nature, London, United Kingdom: Publisher Full Text, 2015.

[21] S. Teufel, A. Siddharthan and D. Tidhar, "Automatic classification of citation function," in *Proc. of the 2006 Conf. on Empirical Methods in Natural Language Processing*, Sydney, Australia, pp. 103–110, 2006.

[22] D. O. Case and G. M. Higgins, "How can we investigate citation behavior? A study of reasons for citing literature in communication," *Journal of the American Society for Information Science*, vol. 51, no. 7, pp. 635–645, 2000.

[23] T. A. Brooks, "Private acts and public objects: An investigation of citer motivations," *Journal of the American Society for Information Science*, vol. 36, no. 4, pp. 223–229, 1985.

[24] B. Finney, "The reference characteristics of scientific texts," Ph.D. dissertation. City University, London, England, 1979.

[25] M. Garzone and R. E. Mercer, "Towards an automated citation classifier," in *Conf. of the Canadian Society for Computational Studies of Intelligence*, Berlin, Heidelberg: Springer, pp. 337–346, 2000.

[26] F. Qayyum and M. T. Afzal, "Identification of important citations by exploiting research articles' metadata and cue-terms from content," *Scientometrics*, vol. 118, no. 1, pp. 21–43, 2019.

[27] R. Ahmad, M. T. Afzal and M. A. Qadir, "Information extraction from PDF sources based on rule-based system using integrated formats," in *Semantic Web Evaluation Challenges*. H. Sack, S. Dietze, A. Tordai and C. Lange (eds.), Cham: Springer, pp.293–308, 2016.

[28] S. Nazir, M. Asif, S. Ahmad, F. Bukhari, M. T. Afzal *et al.,* "Important citation identification by exploiting content and section-wise in-text citation count," *PLoS One*, vol. 15, no. 3, pp. e0228885, 2020.

[29] A. Shahid and M. T. Afzal, "Section-wise indexing and retrieval of research articles," *Cluster Computing*, vol. 21, no. 1, pp. 481–492, 2018.

[30] J. Beel, B. Gipp, S. Langer and C. Breitinger, "Paper recommender systems: A literature survey," *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, 2016.

[31] C. L. Giles, K. D. Bollacker and S. Lawrence, "CiteSeer: An automatic citation indexing system," in *Proc. of the Third ACM Conf. on Digital libraries*, Pittsburgh Pennsylvania USA, pp. 89–98, 1998.

[32] S. Lawrence, C. L. Giles and K. Bollacker, "Digital libraries and autonomous citation indexing," *Computer*, vol. 32, no. 6, pp. 67–71, 1999.

[33] S. Bonzi, "Characteristics of a literature as predictors of relatedness between cited and citing works," *Journal of the American Society for Information Science*, vol. 33, no. 4, pp. 208–216, 2007.

[34] B. A. Lipetz, "Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators," *American Documentation*, vol. 16, no. 2, pp. 81–90, 1965.

[35] C. Oppenheim and S. P. Renn, "Highly cited old papers and the reasons why they continue to be cited," *Journal of the American Society for Information Science*, vol. 29, no. 5, pp. 225–231, 1978.

[36] S. B. Pham and A. Hoffmann, "A new approach for scientific citation classification using cue phrases," in *Australasian Joint Conf. on Artificial Intelligence*, Berlin, Heidelberg: Springer, vol. 2903, pp. 759–771, 2003.

[37] Z. Gu, Y. Cai, S. Wang, M. Li, J. Qiu *et al.,* "Adversarial attacks on content-based filtering journal recommender systems," *Computers, Materials & Continua*, vol. 64, no. 3, pp. 1755–1770, 2020.

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

[39] D. Tkaczyk and Ł. Bolikowski, "Extracting contextual information from scientific literature using CERMINE system," in *Semantic Web Evaluation Challenges*. F. Gandon, E. Cabrio, M. Stankovic and A. Zimmermann (eds.), Cham: Springer, pp. 93–104, 2015.

[40] P. Lopez, "GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications," in *Int. Conf. on Theory and Practice of Digital libraries*, Berlin, Heidelberg: Springer, pp. 473–474, 2009.