SwICS: Section-wise In-Text Citation Score

Ansar Mehmood Khan, Abdul Shahid, Muhammad Tanvir Afzal, Fakhra Nazar, Fahd S. Alotaibi, and Khaled H. Alyoubi

Abstract—Since past several years, finding relevant documents from plethora of web repositories has become prime attention of the scientific community. To find out relevant research articles, state-of-the-art techniques employ content, metadata, citations, and collaborative filtering based approaches. Among all of them, citation based approaches hold strong potential because most of the time, authors cite relevant papers. Bibliographic coupling is one of the well-known citation based approaches for recommending relevant papers. In this paper, we present an approach SwICS that harnesses number of common references between pair of documents as similarity measure whereas the distribution of in-text citations within the text are not analyzed. The proposed approach explores the in-text citation frequencies within contents of the paper and in-text citation patterns between different logical sections for bibliographically coupled papers. For evaluation, the employed data set contains 1150 research documents are obtained from a well-known autonomous citation index known as: CiteSeer. A comprehensive user study is conducted to build a gold standard for comparing the proposed approach. The approach is compared with the state-of-the-art bibliographic coupling and content similarity based techniques. The comparison results revealed that proposed approach significantly performs better than the contemporary approaches. The comparison result with gold standard yielded an average of 0.73. The average gain achieved by the proposed approach is 60% from state-of-the-art: bibliographic coupling. Whereas, the correlation between gold standard and content based approach remains 20%. The proposed approach can play a significant role for search engines and citation indexers in terms of improving the quality of their results.

Index Terms—Bibliographic Coupling, Citation Analysis; Collaborative Filtering; Content Analysis; Metadata analysis

1 INTRODUCTION

ACCESSING relevant information has become a crucial process due to rapid growth of scientific literature. For instance, according to [1], in 1950, the quantity of Journals was 60,000 which exceeded to 1 million journals in 2000 [1]. Contemplating into more statistics from individual organizations, British Library Lending Division (BLLD) indexed 43,000 journals in 1982 [2]. According to Ulrich’s International Database, around 250,000 journals were published in 2004 [3]. PLOS (Public Library of Science) was launched in 2006 which published 10,000 articles during the span of 4 years [4]. Similarly, in bio-medical domain, more than 1,800 new documents are published on daily basis [5] In this state of information explosion, it becomes very necessary to devise such techniques that can help researchers in finding closely relevant information. The scientific community has developed many solutions like CiteSeer, Google Scholar, Microsoft Academic Search, DBLP, and PubMed to assist the scientific community in finding relevant documents [6]. However, most of these tools are unable to deliver most relevant literature to the users. For example, CiteSeer is state-of-the-art autonomous system for citations indexing containing a large number of scientific documents [7] [8] [9]. In order to comprehend the quality of its results, a search was initiated by providing keywords e.g. "Bibliographic Coupling by M.M. Kessler". The provided query is about a well-known technique for scientific document clustering based on the common citations. In the search result, CiteSeer presented a list of 40,607 papers span across multiple pages. This list was manually browsed and it was found that a document holding strong relevance (as shown in below Figure 1) was lying at 42nd position. Further, CiteSeer only provides option for finding related documents in the list of co-cited documents [10]. Although, the Co-citation is a well-known technique for finding related documents, however its scope is limited to web documents [11]. Similarly, Google Scholar and other well-known citation indexers and search engines also present a large list of documents based on term matching [12] [13] [14]. To retrieve strongly relevant results against the query posed by a user, the scientific community has been proposing various approaches which may help in producing a very precise list of related documents. These approaches utilize different data sources such as metadata, content, collaborative filtering based similarity, web access logs, citations and citation networks etc. The most renowned approaches are co-citation and bibliographic coupling analysis. The pioneer study based on co-citation was coined by Small et al. [15]. According to the approach, if two research documents are co-cited in another document, then they are considered as relevant to each other. In co-citation technique, number of co-citations

Fig. 1: A closely relevant paper to the search query was found at 42nd position in the CiteSeer results

---

Manuscript received XXX;
received by any pair increases their relevancy. However, later investigations revealed that co-citation produces better results for web documents whereas Bibliographic coupling performs better for Digital Libraries [11]. Bibliographic Coupling is also citation based approach proposed by Kessler [16]. In bibliographic coupling, relationship between scientific documents is defined on the basis of their common references. The number of common references between two papers are referred as bibliographic coupling unit. The larger value of bibliographic coupling unit corresponds to close relationship between the papers. This research focuses on solving the aforementioned problem by devising a method for finding most relevant research documents. In this study, literature review of state-of-the-art tool and technique is presented. Based on in-depth critical analysis, a new approach is designed, implemented, tested and evaluated which is an extended version of bibliographic coupling approach denoted as SwICS: Section wise In-text Citation Score. The approach is implemented and evaluated on the data set comprising of research papers from the domain of Computer Science. The proposed approach explores the in-text citation counts within contents of the paper and in-text citation patterns between different logical sections for bibliographically coupled papers. To evaluate this approach, a data set containing 1150 research documents is obtained from a well-known autonomous citation index known as: CiteSeer. A comprehensive user study is conducted to build a gold standard data set to evaluate the proposed study. The outcomes of the study is compared with Bibliographic Coupling and Content-based similarity based state-of-the-art techniques. The results revealed that the proposed approach significantly performs better than the contemporary approaches. The average correlation co-efficient between proposed and gold standard is 0.73 as compared to 0.45 between Bibliographic coupling. The average gain achieved by the proposed approach is 60% as compared to the bibliographic coupling. Similarly, the approach is also compared with content based approach whereof the results indicated 20% correlation between gold standard and Content based approach.

2 RELATED WORK
Finding relevant literature from large corpus is a challenging task and in the past, a variety of techniques have been incorporated in Digital Libraries (e.g. CiteSeer, ACM DL, CiteULike, PubMed Central etc.). These proposed solutions by the scientific community can be placed into following broader categories based on their underlying similarity method’s input parameters: metadata, citations, content, citation-content, citation network, content-citation network, collaborative filtering (CF), User profile based. These methods provide a personalized list of relevant literature to the users.

2.1 Metadata based approaches
In metadata based approaches, similarity between scientific articles is computed by comparing their metadata. For example, ACM Digital Library maintains following information as metadata of their articles: author, title, journal, issue_date, volume, number, month, year, ISSN, pages, artcleno, numpages, URL, doi, acmid, publisher, address and keywords. These elements can be manually extracted from articles by publishers, as being done by CiteSeer [10] [17] [17]. Different approaches based on metadata analysis have been proposed to find relationship between papers [18] [19]. Although metadata is helpful in retrieving and managing digital contents, however, it fails to determine close relationship between articles. For instance, same authors do not always mean to work on the same topic. There exist many authors who have contributed in more than one research problems throughout their research career. Similarly, it is not always necessary to match titles of two papers for identifying relatedness. Many research titles evolve over time and are modified into new titles.

2.2 Citations based approaches
References of the paper are manually chosen by the authors of a paper. Authors most choose closely related references for a given paper. Thus, citations contain valuable information and have been used to measure different Scientometrics. For instance, an important technique was developed by Garfield in [21]. In this technique, author introduced a method for indexing scientific articles on the basis of their citation analysis. This technique used the citation links of a given paper to define its worth or contribution in the contemporary knowledge. Citation analysis has also been widely used for measuring the popularity and impact of journals, articles and authors. [20] [15] [16]. Citation based approaches mainly focuses on the references section of scientific papers and measure the similarity between papers considering their references in different manners.

An important investigation of collective utilization of bibliographic coupling and co-citation was carried out by Bichetler et al. in [22]. The aim of this study was to gauge improvement in results by combining both similarity measures. However, these techniques have just considered the references of a paper. Later on, different extensions of these techniques have also been presented in the literature. An extension of bibliographic technique was proposed by Habib and Tanvir [23]. They considered citation proximity of the bibliographically coupled papers. Recently, they have improved their proposed technique by exploiting generic sections details that exist in a research paper [24].

Arjumand et al. extended co-citation technique by incorporating the in-text citation details at sections of a paper [25]. The sections of a paper refer to the generic sections such as Introduction, Literature Review, Methodology, Result etc. They have performed experiments on CiteSeer dataset, and conducted user studies to evaluate their proposed approach. How the logical sections of a paper can be found? For this, Shahid and Tanvir proposed a system based on Sections terms and Paper template details to map actual sections of a paper over logical sections of the article [26].

Shahid et al. argued that all the references given in article do not always indicate close relevance to the concept of a citing paper. Sometime, they are just referred to mention the origin of the idea or only to acknowledge author’s previous
work [27]. Due to these issues with citations, mere citation analysis without considering the content of papers could not be effectively used for finding strong relationship between scientific papers.

### 2.3 Content based approaches

Besides the metadata and citations, content of a paper has also been explored to find similarity between scientific documents [28] [29]. One of the most important techniques was developed by Ratprasartporn et al. [28]. It places papers into a specific context using ontology based scheme. They used gene ontology in their work and then computes similarity between context using text-based, structure based and count based approaches. Papers which lie in more similar contexts are considered as relevant to each other. However, such an ontology hierarchy is hard to obtain for all research topics.

Association rule mining (ARM) is another technique for finding relatedness between scientific articles. ARM has been developed to find meaningful relations among articles [31]. In this technique, the text of a paper has been filtered to gain pure scientific terms. After the filtering, uni-gram and bi-grams have been employed to embed semantic relationship between words into terms. These terms are collected into a candidate set. Similarly, candidate sets of all papers under consideration are generated and further used to find relevant documents. Another well-known system, CiteSeer uses different similarity measures to present related publications to their users. TFIDF is one of them [10] [32]. In this method, vectors of words frequencies are normalized by their inverse document frequencies. Then these vectors are used to yield similarity scores according to the scheme given in [33]. CiteSeer uses another metric which is based on citation analysis.

In this method, vectors of normalized frequencies of words are computed. Then these vectors are used to yield similarity scores according to the scheme given in [10].

Although text based similarity algorithms provide good coverage but have different limitations. For example, one of these limitations is unavailability of content of all research papers. Most of the time, full text is available based on subscriptions. Furthermore, another important consideration is the computation cost. Thus, text based approaches are not feasible for large document corpus. Content composition is another issue associated with content based similarity metrics. When an author writes a paper, three types of elements are blended to form contents. These elements are author’s skill level, his domain concerned vocabulary and the actual semantic of the issue being discussed. So, to segregate semantic of the paper is not fully possible by using only term based approaches.

### 2.4 Citations-Content based approaches

The aforementioned techniques only focus on references, metadata, and text similarity etc. to find similarity between papers. contents and citations based techniques also exist in the literature to compute similarity between papers. In these techniques, an attempt is made to filter out citation of a paper having weak relationship.

Citation Proximity Analysis has been proposed by Gipp et al. [34] to identify related documents. An extension to CPA has been proposed by Gipp et al. in [35]. In this method, user can input an entire document as a query to identify related documents. The process then uses all different possible contextual information such as the citation analysis (Co-citation, bibliographic coupling, cited by and reference list), authors, and sources, implicit and explicit ratings to find related citations. Additionally, the authors have also employed the Distance Similarity Index (DSI) and In-text Impact Factor (ITIF). DSI is the citation proximity analysis which is similar as being used in [36]. Another extension to CPA has been proposed by Boyack et al. in paper [37]. Although this technique also contemplates closeness of citation tags, however, the method of finding proximity has been reformatted by the authors. Shikha et al. proposed another approach based on the hybridization of content and citation [38]. In this technique, an algorithm denoted as C3 has been developed to compute similarity between papers. Similarly, Strohman et al. proposed another approach of ranking related documents [39]. In this approach, authors combined the text based similarity with citation and feature based similarity. This approach uses six features, publication year, text similarity, co-citation coupling, same author, Katz and citation count.

All of these techniques are based on the combined use of contents and references to identify documents relatedness. As mentioned earlier that content and bibliographic references has their own limitations. For example, two papers on different topics from same domain can contain overlapping term sets. Conversely, if two papers are on the same topic then there is also a possibility that their term vectors are different due to inherited ambiguity in English language. Based on this overlapping effect, content based approaches are not suitable to find related documents with high accuracy.

### 2.5 Citation Network based approaches

When references are given in an article, links between citing and cited by papers are created which lead towards a citations network. Many techniques exist in the literature for finding relationship between web documents like Google’s famous PageRank algorithm. Thus, due to the analogy between scientific articles and web documents, the author has extended this technique for finding relevant documents.

Waleed et al. [40] proposed a technique based on citations and author network information for recommending relevant papers. The AMiner dataset has been employed for evaluation of the proposed approach [41]. They have used four papers as query paper for evaluation of their proposed approach. Similarly, Krapivion et al. defined an extended technique based on PageRank to discover relevant scientific [43]. In this category, Nassiri I. et al. proposed a method to prioritize the citations of a target paper. The developed technique has been named as Normalized Similarity Index (NSI) which quantifies the similarity between the source and target of a citation. Nassiri et al. proposed a method to prioritize the citations of a target paper [42]. In this work,
a technique has been developed as Normalized Similarity Index (NSI) to quantify the similarity between the source and target of a citation. SimCC is another most current technique which considers the similarity between two papers as contribution score of cited paper into citing paper [44]. Authors believed that number of citations received by papers is not true indicator of its relevance, rather it is more important to measure the contextual enrichment made by cited papers [44].

The citation networks are created using reference section of papers. However, as mentioned earlier, all the references given in a paper do not always hold a strong relevance to the content of a paper. Thus, parts of citation network which are created by loosely related references introduce a type of noise into the results.

2.6 Content-Citation Network based approaches
Citation networks are derived by references of articles, but as it has been discussed that all references are not always related to the content of original papers, therefore problems associated with references are crept into the citations network and make infeasible to find relationship between papers only based on mere use of citation network. One of important techniques of finding related research papers has been developed by Gori et al. [45]. The approach is based on a random walk approach and the citation graph Random walk approach which is actually a link prediction approach initially developed by Karl Pearson 4. Pruitikanee S. et al. proposed another approach to find similar research papers based on fuzziness [46]. The textual content of paper has ambiguousness inherited from the language and contexts. There are many such terms which are used differently in different contexts for instance partially ordered set or poset are often called directed acyclic graph or DAG. Conversely, two different concepts may have same name in different domains e.g hybridization which is commonly used for software hybridization, hardware hybridization, or algorithmic hybridization [47]. These two problems may introduce irrelevant results as citation graph which is formed on the reference list of papers may create paths between relevant and irrelevant papers [48].

2.7 Collaborative filtering based approaches
Collaborative filtering has been extensively used for recommending items in many different domains. For example in e-commerce, items are recommended to users by exploiting their social networks [49]. A Collaborative filtering based technique of finding related papers has been proposed by McNee et al. [50]. Citation graph of papers is considered as a social network between papers. In this approach, papers-citation relationship is mapped into users-item framework. Although using collaborative filtering for finding related papers is a novel approach, however CF has some limitations such as cold start problem. Collaborative Filtering based approaches may suffer from scalability problem because of ever increasing number of scientific documents.

2.8 User Profile based approach
Most of the Digital Libraries provide access to only their subscribed users. These users perform different activities such as searching, downloading articles and publishing their own articles etc. These activities implicitly build their profile that is further used for presenting adapted and personalized content to users. User profiles contain information which represents the user’s research interests. Different researchers have also exploited user profiles to find relationship between papers exclusively or in combination with other techniques.

One of the important studies based on user profile for recommending research articles has been presented by Lee et al. [52]. In this technique, content-based approach has been combined with user’s previous published work for extracting personalized recommendations. Another technique proposed in [53] uses the user’s recent research interests in order to find related papers. This approach mainly focuses on the user’s profile while generating recommendations. Recommending related papers based on digital libraries access record has also been proposed [15]. In this paper, authors identify two problems in citation data; citation extraction and delayed citation generation.

Summarizing the literature review, the strengths and weaknesses of these approaches have been discussed. The analysis of above mentioned approaches revealed that finding similarity between scientific documents is a special case of information retrieval (IR) process. In this case, there are four types of informative elements embodied in the documents. These are metadata, contents of the paper, organizational structure in the shape of different sections and placement of in-text citations in these sections. Therefore, these elements can be exploited to find degree of similarity. Some of these elements either in isolation or in combination have already been used to find relationship. Placement of in-text citations in different sections is useful information that can assist in finding similarity between documents.

3 Methodology
The in-depth analysis of contemporary state-of-the-art depicts that features such as in-text citations and logical sections hold strong potential in terms of improving the quality of citation indexers and search engines. Therefore, this research extends the standard bibliographic coupling technique by using the in-text citation frequencies and in-text citation patterns in different logical sections of papers. Normally, when an author writes a scientific document, he has specific reasons to place citations in different sections. For example, introduction and related work sections often contain in-text citations of those papers which are topically related. However, these sections also have room for placing citations into less semantically related documents that is the reason behind a large constellation of citations in these sections. Whereas, methodology, results and conclusion sections often contain citations to most relevant papers. Thus, citations in these sections might have more strong relationship with the cited by paper. The proposed technique investigates in-text citations frequencies and in-text citation patterns by extending the traditional bibliographic coupling technique.
This proposed study has drawn the following hypothesis based on discussion presented above.

**Bibliographic coupling based ranking of relevant papers can be improved by exploiting section wise in-text citation frequencies and in-text citations patterns between bibliographically coupled documents.**

To evaluate the above hypothesis, an appropriate research methodology is adopted. Figure 2 shows block diagram of the overall working of the proposed approach. It consists of various sub components e.g. Data Acquisition module that extracts papers from CiteSeer. The details of each component is given below.

### 3.1 Data Acquisition Module

A large corpus is downloaded from CiteSeer which is then converted into XML format using an online tool PDFx [54] to automatically extract the information. Useful information from XML documents is saved into database. The overall architecture of the Data Acquisition Module is presented in Figure 3. This module consists of multiple sub-parts and here the function of each sub-system is presented.

#### 3.1.1 Web crawler

A web crawler which automatically poses different search keywords of computer science domain into CiteSeer is implemented using .Net technology. Search result received by the Crawler consists of DOIs (Digital Object Identifier) for every related paper. These DOIs are scraped and stored in database as shown in the Figure 4. DOIs are further used by crawler for URLs construction to visit summary pages.

Fig. 4: Snapshot of database containing metadata papers stored in database

#### 3.1.2 Metadata Extraction

CiteSeer’s summary page consists of many useful metadata elements e.g. title, authors, year of publishing, number of received citations, abstract, references, downloading link of the paper, links of summary pages for cited references, active bibliography. These metadata elements are extracted and stored in database as shown in Figure 4.

#### 3.1.3 Checking Bibliographic Coupling

Bibliographic Coupling is checked between each pair of papers using their citation information from database. Matching of common citations is based on CID value assigned by CiteSeer to each unique citation. A pair of paper having number of common citations is inserted into a separate table of database along with its number of bibliographic units as shown in Figure 5. Papers are first converted into XML as shown in Figure 6. The reason for converting PDF files into XML is to make it convenient for computations. An example of this process is shown in Figure 5. In this
Fig. 6: XML version of a paper in PDF file. The content of the paper has been represented by various XML tags.

For example, there are two papers, Paper A and Paper B. Some of the common citations between paper A and paper B are highlighted in the figure. Similarly, all pairs of papers which are bibliographically coupled are found and stored in the database as shown in Figure 7.

Fig. 7: The computed Bibliographic Coupling score between Paper A and Paper B where papers (i.e. A & B) are represented by DOIs.

Fig. 8: Linking of xref and ref tags: Ref tag is used to represent a reference in bibliography of the paper whereas xref tag is used in body text of the paper.

3.1.4 Downloading PDF documents

A list of unique papers having bibliographic coupling between them is prepared from database. Then PDF document of each paper is automatically downloaded by the system.

3.1.5 Format conversion using PDFx [54]

Set of downloaded documents are passed to conversion tool PDFx for converting PDF to XML format. XML file contains information enclosed in special tags. Section tags with class attribute “Bibliography” contain a XML representation of all references through a list of nested tags <ref> · · · </ref>. Every Ref tag has an attribute name “rid” whose value is unique throughout entire list of references as encircled in Figure 8. In-text citation tag is represented by <xref> · · · </xref> tag and is also circled in Figure 9. Both of these tags are interconnected through attribute “rid”. Value of this attribute is very important for module 3 which counts the sections-wise in-text citation frequencies.

3.1.6 Citation Synchronizer

After the conversion of PDF files into XML documents, two sets of references for each article are obtained as shown in Figure 9. One set is obtained from CiteSeer’s summary page stored in database and other set is in the form of bibliography section present in XML file. Both of these are used to execute the experiment. To update RID value, a utility is developed and named as Citation Synchronizer.

Input:

string c1=Title of citation stored in database
string c2=Citation in XML file (inner html of ref tag)
List L1=List of English stop words.
List L2=List of special characters.
integer match_score=0
integer index=0

c1=RemoveSpecialChars(c1, L2)
c2=RemoveStopWords(c2, L1)

c2=RemoveStopWords(c2, L1)
if c2.contain(c1) then
  return true

else {
  foreach (word w in c1) {
    if c2.contain(word, index) then //Start from index
      index=c2.indexOf(w) + length(w)
      match_score=match_score+number of word in c1*100
      if score >= 80 then
        return true
      else
        return false
  }
}

Fig. 10: Citation Matching Algorithm
The main part of this utility is citation matching algorithm. Figure 10 shows the complete algorithm. The basic task of this citation matcher is to find degree of similarity between title of stored citation and a reference given in XML file. If a sufficient degree of similarity is found, then corresponding “rid” value from XML file is updated in database. However, matching needs some preprocessing due to variations in both strings. This includes removal of hyphens and other special/accented characters like ˆa, œ, , #, , “et.al” etc to obtain better matching results. Matching is also performed in two phases; first direct string containment is tested, if it fails then partial words containment (considering order of words) is tested with a threshold of 80% [56].

By using Data Acquisition Module, information of 1150 documents is collected, downloaded as PDFs and converted into XML files. This set of corpus and information is stored in database which will be utilized in the experiment.

Fig. 11: Sections are extracted from XML files and then those extracted sections are mapped over logical sections.

### 3.2 Section Mapping

In the second step, section titles of each paper of selected corpus are extracted and mapped over the proposed generic sections. The overall architecture of this module is shown in Figure 11. XML documents are passed on to a utility Section Extractor to generate a list of all sections properly identified by PDFx.

**Table 1: List of generic sections**

<table>
<thead>
<tr>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
</tr>
<tr>
<td>Related Work</td>
</tr>
<tr>
<td>Architecture/Methodology</td>
</tr>
<tr>
<td>Results/Comparison</td>
</tr>
<tr>
<td>Conclusion/Future work</td>
</tr>
</tbody>
</table>

In the next step, mapping between paper specific sections and the generic sections is generated. The mapping algorithm is shown in Figure 12. This technique exploits the structure of scientific documents to generate this mapping.

Scientific articles are divided into different organizational sections and often have a general layout pattern [57] [58]. A typical list of generic sections is given in Table 1. To automatically place context specific section’s heading into their generic counter parts, layout pattern combined with term matching can be used. Figure ?? shows a sample of mapping results. This mapping is used in next module for counting in-text citations.

Now, the approach finds the locations of these logical sections. Similarly mapping is generated for each document in corpus and results are persisted in database.
Results

Methodology

Introduction

Discussion/Future work

Conclusion/Future work

Fig. 13: A snapshot of database having details of papers’ sections mapping over logical sections

3.3 Section-wise In-text Citations

In third step, the section wise in-text citation frequency of every common reference between two papers is counted.

Fig. 14: In-text citation distribution of references in various section of the paper is shown in this diagram

Figure 14 shows an example of section wise in-text citations in a document. In this figure, a list of generic sections (mapped sections heading are marked with ‘→’ symbol) is presented as a row heading, and RID values are shown as column headings. Cell value shows the in-text citations of a specific section-rid combination.

3.4 Measuring Similarity Score

In the fourth step, the similarity between papers is calculated using section wise in-text citation score. In this research work, bibliographic coupling has been combined with section-wise in-text citation count and denoted as SwICC.

3.4.1 Section-wise Citation weighting scheme

If a document is cited in related work/Introduction section then it might be a supporting document. In contrast, if a document in cited in methodology/results section then this document may have close relationship with citing document [59] [60]. Similarly, it is already believed that related work section has less importance [57]. Therefore, more weights are assigned to the citations in methodology and results. Based on these intuitions, the following equation (Eq 1) is constructed in this research.

\[ \omega_{\text{Meth}}/R_{es} > \omega_{\text{intro}} > \omega_{\text{rel}} \] (1)

The Eq 1, \( \omega_{\text{meth}} \) and \( \omega_{\text{res}} \) represents the weights of methodology and results sections respectively. Similarly, \( \omega_{\text{intro}} \) and \( \omega_{\text{rel}} \) represents the weights of introduction and related work sections respectively. The weight of methodology/results section was 3, introduction section was 2 and related work section is fixed at 1. These weights are presented in Table 2.

TABLE 2: Relative importance weights for three possible combinations with introduction section

<table>
<thead>
<tr>
<th>Section of Paper A where common citation appeared</th>
<th>Section of paper B where common citation appeared</th>
<th>Relative importance weight (( \omega ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>Related Work</td>
<td>0.5</td>
</tr>
<tr>
<td>Introduction</td>
<td>Methodology</td>
<td>0.24</td>
</tr>
<tr>
<td>Introduction</td>
<td>Results</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Generally important references can be cited in more than one section. Possible combinations and their corresponding cross sectional weights are shown in Table 3 to 5. It has been observed that Conclusion/Future work section often have very small proportion of total citations. Therefore, in this investigation these sections are not being considered.

3.4.2 SwICS: Section wise In-text Citations Score Formulation

Assume there are two bibliographically coupled papers p and q. In the Eq 2, \( R_p \) and \( R_q \) are the set of references in paper p and q respectively, \( R_c \) is the set of common references and \( S \) is the set of all generic sections.

\[ R_c = R_p \cap R_q \] (2)

\[ Swics_{(p,q)} = \sum \Phi(r) \] (3)

The formula of SwICS is shown in equation Eq 3. SwICS score is computed between paper p and q by summing up the contribution score of all common references in \( R_c \).

\[ \Phi(r) = \sum_{x \in S} \sum_{y \in S} Min\{ICS_{px}(r), ICS_{qy}(r)\} * \omega_{xy} \] (4)
In Eq 4, function $\Phi(r)$ is the contribution score made by a common reference $r \in R_c$. The definition of function $\Phi(r)$ is given in Eq 4. In this function $ICS_{px}(r)$ denotes the in-text citation of common reference $r$ in generic section $x$ of paper $p$. Similarly, $ICS_{qy}(r)$ represents the in-text citation of reference $r$ in section $y$ of paper $q$. The minimum value of in-text count is multiplied with the relative importance weight ($\omega$) of corresponding section combination as shown in Tables from 2 to 5. The reason behind choosing minimum value is simple. The numbers of citation of those references which are cited in both papers are being boosted. If a paper has zero in-text citation of reference $r$ in corresponding section, then zero will be the minimum value and it will bring the whole score of $r$ to zero for that particular section combination.

4 RESULTS

In this approach, corpus of 1150 documents are submitted to PDFx, wherein 827 documents are properly converted into XML format to check bibliographic coupling between documents, 703 documents are having their coupling units to each other. From 703 documents, it is found that there are 8333 bibliographic coupling pairs. Mapping is manually evaluated by observing the actual sections. For this purpose, a sample of 300 out of 703 documents is selected from the corpus. It is noted that 300 documents are having 1971 sections (excluding subsections). The proposed mapping scheme correctly mapped 1914 section to the standard logical sections which is 97.11%. These statistics are graphically shown in Figure 15. Further, in-text citation frequencies computed for all 703 documents of each reference using its citation tags are calculated within different mapped sections of bibliographically coupled documents.

These sample documents are analyzed for two objectives; correctness and distribution of citations within logical sections. The complete distributions of citations are shown in Figure 16. Further, in-text citation frequencies are computed for all 703 documents of each cited reference. In final step, the SwICS: Section wise in-text Citation Score is calculated between all bibliographically coupled pairs.

4.1 System Evaluation

For the evaluation of this proposed approach, a gold standard ranking is developed using a detailed user study. Then ranking of this proposed approach, state-of-the-art approach, and the ranking using content based approach are compared with the gold standard ranking.

4.1.1 Building the Gold Standard

To gauge the performance of document recommender systems, a user study is usually performed [42] [61] [52]. Therefore, gold standard ranking is obtained using a comprehensive user study. A sample of 22 users (PhD scholars and post graduate students) is invited to give their opinion about the ranking of related papers through an online system. The details of the users are shown in Table 6. The evaluation setup consists of documents from various domains and users were free to select any document of their choice. A list of papers requiring user opinion is presented to the user as shown in Figure 17. User can select any paper of his/her interest. The last column in figure shows the
number of evaluations received by any paper. User may select the paper which has been evaluated previously. The reason of this is to identify their mutual consent and to measure the inter rater agreement. This means that if both users are looking at the paper in the same way then it can be considered as standard opinion. When user selects a paper, a new page is displayed as is shown in Figure 18.

For evaluation, only ten related papers (selected randomly) are provided; so that users can easily give their opinion. User can enter rank number of corresponding related document in the corresponding textbox. Rank numbers are simple integer from 1 to 10. In case a user thinks that a document is not related to the focused paper, he may leave its rank number. By using the above explained setup for evaluation, the user opinion for 32 paper having 320 bibliographically coupled pairs are obtained. Furthermore, it is pertinent to mention that each paper is evaluated by two distinct users. Inter rater agreement was calculated using Spearman’s correlation co-efficient [62]. The formula of spearman correlation coefficient is shown in Eq 5. Here, d represents the difference between two ranking values and n is the number of total values.

\[
r' = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}
\]

The ranking of papers which is agreed upon by both raters are considered as gold standard ranking. Figure 19 depicts the inter rater agreement between both users for 32 documents. Finally, the Bibliographic coupling and proposed approach is compared with the gold standard.

4.1.2 SwICS: Section wise in-text Citations Score vs. Bibliographic Coupling

In this section, the findings of comparison performed between the proposed approach and bibliographic coupling approach is explained. A separate correlation between SwICS and Bibliographic Coupling with the gold standard ranking is calculated using the Spearman’s correlation coefficient. Then these correlations are compared and results are shown in Figure 20.

From Figure 20, it can be observed that agreement between users’ opinion and the proposed approach is better as compared to bibliographic coupling for majority of the documents out of 32 papers. There is only one document where proposed and Bibliographic coupling approach performed equally. However, there are four cases where bibliographic coupling approach performed better than the proposed approach. The average correlation of the proposed approach with the gold standard ranking is 0.73, whereas the average correlation of the bibliographic coupling with the gold standard ranking is computed as 0.45. Thus, average correlation has been improved by 60% in the proposed approach as compared to the state-of-the-art bibliographic coupling.

4.1.3 SwICS: Section wise in-text Citations Score vs. Content based similarity

To compare proposed approach with content base similarity, first cosine similarity measure is implemented on Lucene indexer [63]. Then for finding the correlation between ranking of documents by gold standard and content based similarity again the Spearman’s correlation co-efficient is used. Comparison between SwICS and content base similarity is shown in Figure 21.

Fig. 18: Selected paper by the user for giving opinion

Fig. 20: The spearman correlation values are computed between user ranking with proposed approach and Bibliographic Coupling

Fig. 19: Inter rater Agreement between users are computed using spearman correlation.

Fig. 21: Comparison between SwICS and Content-based similarity approach. The spearman correlation values are computed for each document.
In Figure 21, the blue color represents correlation between proposed and gold standard ranking and similarly the dark brown color represents correlation between content-based similarity and gold standard ranking. Furthermore, it can be observed that agreement between users’ opinion and the proposed approach is better as compared to content similarity for majority of the documents. There is only one document where proposed approach and content similarity approach performed equally. Similarly, there is only one document where content-based similarity performed better than proposed approach. The average correlation of the proposed approach with the gold standard ranking is recorded as 0.73, whereas the average correlation of the content based similarity with the gold standard ranking remained 0.20. The overall improvement achieved by proposed approach as compared to bibliographic coupling and content similarity is depicted in Figure 22.

These results proved the hypothesis that in-text citation frequencies and in-text citation patterns can improve the ranking between bibliographically coupled documents. Although, proposed approach works fine on the small size of corpus and specific domain of Computer Science; however, the generalization of such approach is still questionable for other sub-domains of Computer Science, and for the domains other than Computer Science. The performance of this approach largely depends on other tools and technique e.g. the correct identification of logical sections, the accurate in-text citation counts etc [64]. Thus documents having XML conversion issues can drastically decrease the performance of this approach.

5 CONCLUSION AND FUTURE WORK

In this paper, a section-wise in-text citation pattern based approach has been presented to find the relevant documents. The current state-of-the-art harnesses different data sources to recommend relevant research papers. These sources include metadata, content, citations, citation networks, collaborative filtering, user profile and access logs etc. A hypothesis has been constructed that the Bibliographic Coupling based ranking of related documents can be improved using in-text citation frequencies and patterns. The proposed approach identifies in-text citation pattern (occurrences of in-text citations of common references in different logical sections of the citing paper). Based on these in-text citation details, the proposed system calculates the relevance score and then ranks the documents according to obtained relevance score. The results show that the proposed system yielded an average correlation of 0.73 with the gold standard ranking; whereas the correlation of Bibliographic Coupling with the gold standard is 0.45 and the correlation of content similarity is 0.20. Although, the proposed approach performed significantly well, however, the study has only been evaluated on the corpus of Computer Science related documents. To generalize the approach, it must be evaluated on diversified disciplines. In the future, we intend to evaluate the proposed technique on other sub-topics of Computer Science. This technique may also be tested for scientific domains other than Computer Science. If similar or better results are obtained on other domains then a web service could be designed on top of existing state-of-the-art systems (like CiteSeer, IEEE digital library, ACM digital library, Google Scholar etc.) to help researchers and academicians.

6 ACKNOWLEDGEMENT

This work was supported by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under grant No. (DF-149-611-1441). The authors, therefore, gratefully acknowledge DSR technical and financial support.

References


