IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 22, 2020, accepted July 3, 2020, date of publication July 13, 2020, date of current version July 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3009021

A Systematic Approach to Map the Research Articles' Sections to IMRAD

IBRAR AHMED AND MUHAMMAD TANVIR AFZAL
Department of Computer Science, Capital University of Science and Technology, Islamabad 44000, Pakistan

Corresponding author: Ibrar Ahmed (ibrar.ahmad@gmail.com)

ABSTRACT The amount of scientific publications is believed to get doubled every five-years. These publications are stored by citation indexes and digital libraries in the form of complete PDF or/and by extracting terms from these documents. This indexing behavior poses several challenges for the scientific community as well as for digital repositories in terms of handling the advanced requirements of a user. For instance, addressing queries like ''Give me those papers that contain the term ''Pagerank'' in their result section'' may not be answered unless the papers are indexed section-wise. This issue has been focused by researchers and international prestigious challenges by top venues in the world like Semantic Publishing Challenge in ESWC. One of the important metadata extraction from research papers is the section information such as IMRAD (Introduction, Methodology, Results, and Discussion). Researchers have presented different approaches to identify and map the section-headings to IMRAD sections. The existing studies have employed parameters like dictionary terms, the template of a paper, and in-text citation frequency to map section-headings onto logical sections. The critical analysis of state-of-the-art revealed that some immensely potential features have been ignored, which might result in accurate mapping. In this study, we propose a novel approach that employs new features along with previously well-known features to map sections-headings to IMRAD. The newly proposed features are: (1) variant of In-text Citation count (2) Figure counts, (3) Table counts, and (4) subheading implicit mapping. The employed data set contains 5000 research papers, collected from CiteSeer. The evaluation of the proposed approach and comparisons with state-of-the-art three approaches revealed an improvement of 18.96%, 21.77%, and 9.50% in average precision with Ding et al, Shahid et al, and Habib et al respectively. This research has significant implications for citation indexes and digital libraries.

INDEX TERMS IMRAD, relations database, section mapping, scientific document classification.

I. INTRODUCTION

Communication in science is realized through scientific publications. Due to the latest inventions in science, a tremendous increase has been reported in the amount of publications on WWW. The amount is believed to get doubled after every five-year [1]. The scientific plethora is diffused on the web through different means like publishing in different venues such as conferences, journals, and workshops. These venues publish their research corpora on the web which is indexed by digital repositories like CiteSeer, Google Scholar, and Scopus, etc. A user exploits information retrieval (IR) systems like search engines, citation indexes or digital libraries to extract desired information. A user poses a query with an intention that he/she will obtain the maximum relevant information. For instance, a research scholar finding papers to conduct a literature survey will always wish to retrieve a maximum amount of strongly relevant research papers against the topic posed in the query. However, the existing IR systems index the data in a semi-structured format. The PDF is one of the most widely employed semi-structured formats, which was developed under the Camelot Project to share documents that include text and images. Due to improper indexing of PDF files, the existing IR systems are unable to handle advanced queries of a user. The examples of advanced queries are: (1) find all the research papers containing the term ''Data Science'' in the Methodology section of a research paper or (2) find all those papers that have calculated ''F-measure'' in the ''Results'' section etc. The queries of such nature can only be addressed if research papers are section-wise indexed.

The associate editor coordinating the review of this manuscript and approving it for publication was Justin Zhang.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

The community has proposed solutions in the form of defining the semantic structure of PDF documents and performing their section-wise mapping [2].

Initially, research papers were written in the letter-style format. In the early 20th century, a standard format was presented in the form of IMRAD (Introduction, Methods, Results, And Discussion) [3].

The origin of the IMRAD is vague. However, according to Batmanabane [4], Louis Pasteur is the first person who used that format and later used by Sir Austin Bradford Hill. The structure states that a research paper should comprise logical sections like Introduction, Methods, Results, and Discussion. The current behavior from the majority of the scientific community in terms of preparing research papers favors the IMRAD structure. Identifying logical sections of a research paper has already been focused by international prestigious challenges by the top venues in the world like Semantic Publishing Challenge in ESWC [5] and the research community [2], [6]–[8].

It should be noted here that the names given by the authors of the papers as the section-headings usually are not identical to the names being used by IMRAD structure (i.e Introduction, Methods, Results, and Discussions). For example, according to the state-of-the-art [2], that performed experiments on 1,833 section-headings, have concluded that none of the methodology section was named as ''Methods/Methodology'' by the authors of respective papers. Only 1

During the course of many years, researchers presented different section identification techniques [2], [6], [8]. Ding *et al.* [8] performs a study to identify the distribution of in-text citations across sections. For this task, the researcher identified the section headings and mapped those headings on to logical sections. For this, the researcher used very extensive dictionary terms to identify the section and applied their technique on 866 full-text articles containing 6866 sections and achieved 81% accuracy. Shahid and Afzal [2] extended Ding *et al.* [8] technique with different dictionary terms along with research paper templates and layout to identify section headings and mapped them to IMRAD structure. The researcher applied the technique on 1200 papers containing 12,180 sections and got 0.78 precision and 0.79 recall. Habib and Afzal [6] used frequency of in-text citation to identify the section and applied that technique on 5000 bibliographically coupled papers and achieved 90% accuracy.

However, on the critical investigation, we identified two key areas. In some scenarios, the contemporary approaches are unable to differentiate between sections and subsections of a research paper. For example, a logical section ''1. Introduction'' contains a subsection, named ''1.2 Background''. The existing approaches treat both of them as independent headings and map them individually to the IMRAD structure, which may increase the chances of inaccurate mapping. Some studies also consider the subheading and rely on heading tags <h1…hn>, which can be failed in some cases, we have explained the issue in detail in section 2 of this paper. Other than subheadings, there exists a list of potential section identifier parameters that have been ignored by the existing state-of-the-art.

In this paper, we present a comprehensive approach that intelligently maps research articles to IMRAD. The proposed approach takes advantage of accurately identifying the subsections and mapping them to IMRAD headings based on their main section mapping to achieve better results. Furthermore, the proposed approach also exploits novel potential features which have great potential to improve the performance of IR systems. The features include: (1) In-Text Citations Count (2) Figures Count, and (3) Tables Count. These features have been chosen based on the assumption that their certain frequency may hint in determining the association of one typical heading to a specific logical section. The proposed methodology is evaluated on logical sections of 5000 research papers in PDF taken from CiteSeer having 39420 section headings.

We have compared the proposed approach with all of three section-mapping techniques [2], [6], [8] on the same dataset of 5000 papers. Our proposed technique outperformed all three. The proposed approach gained 18.96%, 21.77%, and 9.50% improvement in average precision of all sections from Ding *et al.* [8], Shahid and Afzal [2] and Habib and Afzal [6] respectively.

The rest of the paper is organized as follows: Section II presents a critical analysis on state-of-the-art systems. Section III highlights the lessons learned from empirical experiments on research papers. Section IV contains the information about the dataset followed by Section V which elaborates the methodology proposed to map sections onto the IMRAD structure. Section VI discusses the results and analysis. In the end, the performance evaluation has been demonstrated in section VII followed by the conclusion that is presented in section VIII.

II. LITERATURE REVIEW

Since 1665, the letter style format has been followed by the research community to prepare research documents [9]. In the earlier 20th century, a standard structure designed specifically to write research papers was presented in the form of IMRAD (Introduction, Methodology, Results, and Discussion). Gradually, the popularity of the structure increased and most of the research community adopted this for preparing research documents [9]. The community believed that if the structure of research papers is defined by mapping their logical sections to IMRAD, then the performance of IR systems can significantly be enhanced. In this regard, the scientific community has put various efforts from defining the structure of PDF files to mapping them to IMRAD. The studies have focused on adding semantic structure to the content of PDF files to retrieve information in an intelligent manner. A semantically defined structure of a document can be used for different applications pertaining to information retrieval, such as, to generate summaries or process advanced queries. The studies focusing on defining the

structure of a research paper's content are referred to as discourse analysis. The scientific community has presented research papers sections-based studies in two dimensions: some have employed logical sections of research papers to identify different aspects pertaining to bibliometric analysis and others have mapped logical sections to IMRAD structure.

A. TECHNIQUES USING LOGICAL SECTIONS

The study proposed by Teufel *et al.* [10], presented an Argumentative Zoning (AZ) system, using elements of scientific argumentation [10]. These scientific arguments are referred to as ''owns a work'', ''others work'' and ''contrast'' [11]. Another similar scheme, Core Scientific Concepts (CoreSC) extracts generic concepts like Hypothesis, Model, and Experiments, etc. from research papers [12]. Besides defining the structure, IMRAD has also proven useful in various other contexts, for instance, Teufel [13] stated that investigating a citation count with respect to its location in a logical section can produce good results to discern the sentiment aspect of the citing author. Similarly, another study [14] has performed citation analysis by exploiting in-text citations in different sections (introduction, methods, results, and discussion) of a research paper. The importance of logical sections has also been delineated by Shahid and Afzal [2], to discover the semantic relationships between research articles. Another approach [15], presented an ontology (DEo) to define logical structures of scientific documents. The semantic indexing of research documents holds great potential in identifying implicit knowledge. In the study [16] authors have exploited in-text citation frequencies and their patterns from the logical section of research papers. The evaluation has been done on the data set of scientific papers taken from CiteSeer. Kafkas *et al.* [17] presented a study wherein sections-based search functionality was provided for the research papers published in the Journal of Biomedical. The approach presented by Shahid and Afzal [2] is closest to Kafkas *et al.* [17] The difference is that the approach Kafkas *et al.* [17] manually extracts the sections by utilizing designed rules. The semantic indexing of research documents holds great potential in identifying implicit knowledge. The contemporary IR systems are unable to semantically index the structure of PDF files due to which advanced queries cannot be processed. The examples of advanced queries are: (1) find all the research papers containing the term ''Data Science'' in the Methodology section of a research paper or (2) find all those papers that have calculated ''F-measure'' in the ''Results'' section etc. Addressing such queries is a dire need of the current era especially when there exists a huge amount of research plethora on the web. However, defining a semantic structure that is able to address the aforementioned queries is a challenging process. This is due to the fact that people use different sets of vocabulary or semantic terms to name the same logical sections. For instance, some authors use the term ''Literature Review'' while some use ''Related work'' for a section to represent state-of-the-art studies. In such scenarios, it becomes difficult to semantically distinguish the terms. Such issues have also been reported by Shahid and Afzal [2] wherein 329 papers were manually assessed and it was revealed that none of the papers used term ''Methodology'' for the methodology section and only 1% of the papers used the term ''Result'' for the ''Result'' section. Besides these, some other researchers also used the logical section to identify important citations for example like Nazir *et al.* [18], Hassan *et al.* [19], and Pride and Knoth [20].

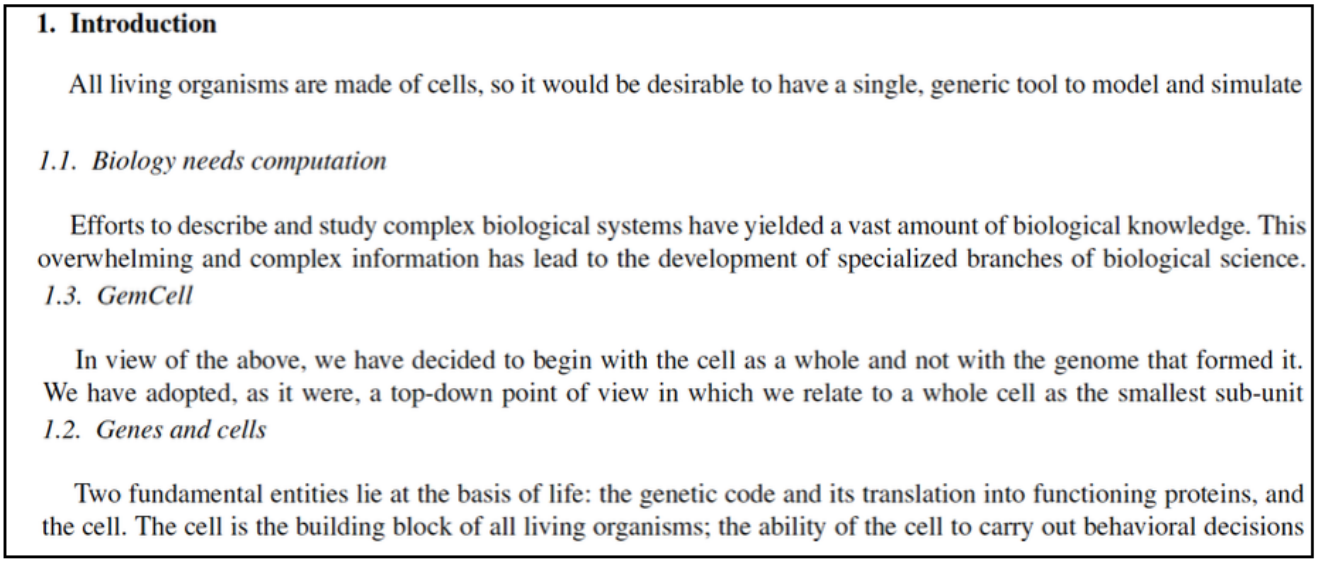
B. TECHNIQUES MAPPING LOGICAL SECTIONS

Ding *et al.* [8] used extensive dictionary terms to identify and map the logical sections to IMRAD. This technique was tested on 866 full-text articles containing 6866 sections and achieved 81% accuracy. Shahid and Afzal [2]. extended the study of Ding *et al.* [8]. and used different dictionary terms and templates of the papers to identify and map the section to IMRAD. The study of Shahid and Afzal [2] has mapped the research articles from the domain of Computer Science and automatically extracted the sections using DEo ontology. The study has then performed a rigorous analysis of comparison done with various ML techniques and unlike the study presented by Kafkas *et al.* [17], the approach [2] has formally presented the proposed algorithm. The approach Shahid and Afzal [2] have mapped the sections of research papers into six logical sections of IMRAD (''Introduction'', ''Related Work'', ''Methods'', ''Results'', ''Discussion'' and ''Conclusion''). The approach [21] has harnessed two main features, layout information of scientific documents and dictionary terms to form heuristics to section-wise map the research articles on IMRAD structure. The evaluation has been done on 329 papers from the domain of Computer Science published in the Journal of Universal Computer Science (J.UCS). Habib and Afzal [6] used in-text citation frequency to identify the sections. This technique achieved 90% accuracy when tested on 5000 papers.

After a critical analysis of the studies presented above illustrates that contemporary state-of-the-art has proposed several methods to semantically index research documents according to some predefined structure. A few efforts have been made to define the structure of research documents from the domain of Computer Science. The approach proposed by Shahid and Afzal [2], is the most recent approach which performs section-wise mapping of research articles from the domain of Computer Science by utilizing heuristics formed using layout information and content information. However, the approach holds various deficiencies which can adversely impact the performance of IR systems. We have critically scrutinized the approach by manual investigation and identified existing gaps which are the focus of the proposed study. The following section ''Lessons Learned'' presents an in-depth analysis of the identified issues with the help of examples taken from a real data set.

III. LESSONS LEARNED

As explained earlier, our work is closest to the work presented by Shahid and Afzal [2]. The approach presented by



1. Introduction

All living organisms are made of cells, so it would be desirable to have a single, generic tool to model and simulate

1.1. Biology needs computation

Efforts to describe and study complex biological systems have yielded a vast amount of biological knowledge. This overwhelming and complex information has lead to the development of specialized branches of biological science.

1.3. GemCell

In view of the above, we have decided to begin with the cell as a whole and not with the genome that formed it. We have adopted, as it were, a top-down point of view in which we relate to a whole cell as the smallest sub-unit

1.2. Genes and cells

Two fundamental entities lie at the basis of life: the genetic code and its translation into functioning proteins, and the cell. The cell is the building block of all living organisms; the ability of the cell to carry out behavioral decisions

FIGURE 1. Subheading example in form of PDF file.

```
<section class="deo:Introduction">
<h1 class="DoCO:SectionTitle" id="9" page="2" column="1">
1. Introduction </h1>

<h2 class="DoCO:SectionTitle" id="20" page="3" column="1">
1.1. Biology needs computation </h2>

<h2 class="DoCO:SectionTitle" id="40" page="3" column="1">
1.2. Genes and cells</h2>

<h2 class="DoCO:SectionTitle" id="44" page="3" column="1">
1.3. GemCell </h2>
```

FIGURE 2. Subheading Example 1 (XML).

Shahid and Afzal [2] maps logical sections of research papers to IMRAD by using template information and dictionary-based rules. We have implemented the approach of Shahid and Afzal [2] and discovered some patterns that led us to formulate the proposed research questions. Let us look into the patterns with the help of real examples taken from the PDF files of the employed data set.

A. SUBHEADINGS MAPPING

In the PDF file shown in Figure 1, the Introduction section of a research paper "1. Introduction" contains three subsections, "1.1. Biology needs computation", "1.2. Genes and cells" and "1.3. GemCell". The XML file of this PDF shown in Figure 2 the main heading (i.e., 1. Introduction) is represented with tag <h1> and all the sub-headings are represented with the tag <h2>. The content inside the opening and closing bracket of the heading tag is considered as the name of an independent logical section. All the remaining subheadings are also represented with the same tag <h1>. This manual inspection revealed the fact that the approach [2] is unable to differentiate between the main section and subsection of a paper, rather it treats all of them as independent headings. This deficiency causes another major issue, explained below. Ding *et al.* [8] already used the heading <h1> and <h2>. However, if we see in Figure 3, there is a sub-section named "1.3 Related Work". Now, as per the IMRAD structure, the section will be considered as an independent section and will get mapped to the "Related Work" section of IMRAD. However, in reality, the section does not belong to the literature review section of IMRAD. Such issues result in false mapping,

```
<section class="DoCO:Section">
<h1 class="DoCO:SectionTitle" id="5" page="1" column="1">
1 Introduction 1.1 Skyline Queries </h1>

<h1 class="DoCO:SectionTitle" id="13" confidence="possible"
page="2" column="1"> 1.2 Online Skyline Computation </h1>

<h1 class="DoCO:SectionTitle" id="18" confidence="possible"
page="2" column="2"> 1.3 Related Work </h1>
```

FIGURE 3. Subheading Example 2 (XML).

```
<region class="DoCO:FigureBox" id="F1">
<caption class="deo:Caption" id="9" page="1" column="2">
Figure 1: Skyline of hotels in Nassau (Bahamas) </caption>
```

FIGURE 4. Figures in research publications.

further compromising the performance of IR systems. On careful examination of the heading content, we identified that most of the headings start with a bullet number. For instance, the main heading starts with "1…" followed by the sub-headings with "1.1..", "1.2..", "1.3.." and "1.4..". We argue that the inability of differentiating the logical structure by XML can be addressed with the help of regular expressions that are able to intelligently differentiate between the sections on the basis of the mentioned patterns.

To recapitulate, the examples discussed above depict that the scope of the contemporary state-of-the-art on logical sections mapping fails to intelligently map the subsections. Treating all the sub-sections as independent sections and explicitly mapping them to the IMRAD structure can have an adverse impact on the overall precision of IR systems. Our study overcomes all the mentioned issues by utilizing regex to implicitly map the subsection to the same section which is its main section. As explained earlier, the proposed study employs some potential features such as In-Text Citations count, Figures count, and Tables count to determine the association of a section to specific logical sections of IMRAD. The justification for harnessing these parameters is given below.

B. FIGURES AND TABLES COUNT

A scientific article illustrates results in the form of a figure or table. There may be a high probability that the frequency of figures or tables hints towards the association of the section to a particular logical section of IMRAD. For instance, a "Methodology" or "Result" section contains a comparatively higher "Figure count" or "Table count" than other sections. To the best of our knowledge, the contemporary approach [2] has not given the due importance to the potential parameters "Figure count" and "Tables count". In this study, we consider both parameters "figure count" and "table count" based on an assumption that the count of figures links to the specific section of the IMRAD structure. In the XML files, figures and tables are represented in the form of objects, as shown in Figures 4 and Figure 5.

C. IN-TEXT CITATIONS FREQUENCY

Similar to the aforementioned assumption followed for "Figure Count" and "Table Count" parameters, the "In-Text Citation count" can also serve as an important indicator in determining association to a specific logical section. Ding *et al.* [8] have employed the frequency of in-text citation in all the logical sections. We have manually investigated research papers of the employed data set and identified that the number of in-text citation counts is not equal in all the sections. For instance, the "Literature Review" section contains the highest amount of in-text citations than other sections. Considering this aspect, our approach maps a logical section having the highest number of in-text citations to the Literature review section. Habib and Afzal [6] have considered this parameter for mapping logical sections to the IMRAD structure using the Ding *et al.* [8] study. Similar to figures and tables, citations are also represented in the form of objects in XML files, as shown in Figure 6.

The important aspect overlooked by contemporary studies is that they have not given adequate importance to the potential features like "In-Text Citation count", "Figure count" and "Table count". Although, as explained above, these parameters could be the potential contributors for section identification. In this study, our focus is to overcome the stated deficiencies to improve the performance of IR systems to a great extent.

```
<region class="DoCO:TableBox" id="Tx100">
<content>
    <table class="DoCO:Table" number="1" page="9">
    <thead class="table"/>
        <tbody>
            <tr class="table">
                <td class="table"></td>
                <td class="table"> 100,000 Points</td>
                <td class="table"> 1,000,000 Points</td>
            </tr>
        </tbody>
    </table>
</content>
```

FIGURE 5. Tables in research publications.

```
<xref ref-type="bibr" rid="R1" id="21" class="deo:Reference">
1 </xref>,
<xref ref-type="bibr" rid="R2" hidden="1" id="22" class =
"deo:Reference" > 2
</xref>,
<xref ref-type="bibr" rid="R3" hidden="1" id="23" class =
"deo:Reference" > 3
</xref>,
```

FIGURE 6. Example of a figure caption.

IV. DATA COLLECTION

Contemplating the fact that an appropriate data set plays a crucial role in determining the significance of a proposed study, we have collected the data set in such a way that ensures the validity of the proposed study on the papers published in diverse domains.

For the verification of our approach, we require a comprehensive dataset from diversifying domains, covering different authors, and different journals. We found a dataset that has the characteristic which we require from Habib and Afzal [6]. This data set is freely available. We used 17 different queries mentioned in the Table 1 which are adapted from Habib and Afzal [6] to collect the data from CiteSeer. CiteSeer indexes a vast amount of research papers in diversified disciplines of Computer Science. The employed dataset contains 5000 papers containing 39420 sections of different journals.

TABLE 1. Queries used to collect dataset.

Queries used to collect dataset [6]	
Number	Query
1	Social network
2	Information retrieval
3	Bayesian networks
4	Feature selection
5	Collaborative recommendation
6	Recommendation system
7	Content based filtering
8	Black box testing
9	Automatic generation
10	Regression testing
11	Query processing
12	Sensor networks
13	Wireless communications
14	Opinion mining
15	Subjectivity analysis
16	Online marketing
17	Graph theory

V. METHODOLOGY

This section encompasses details about the proposed methodology. It works in four modules to map the logical sections of research articles to the IMRAD structure. The modules include Schema Generation Engine (SGE), (2) Data Extraction Engine (DEE), (3) Data Mapping Engine (DME), and Mapping View Engine (MVE). First of all, the data set containing PDF files of research papers are collected from a digital library named CiteSeer. The PDF files are converted

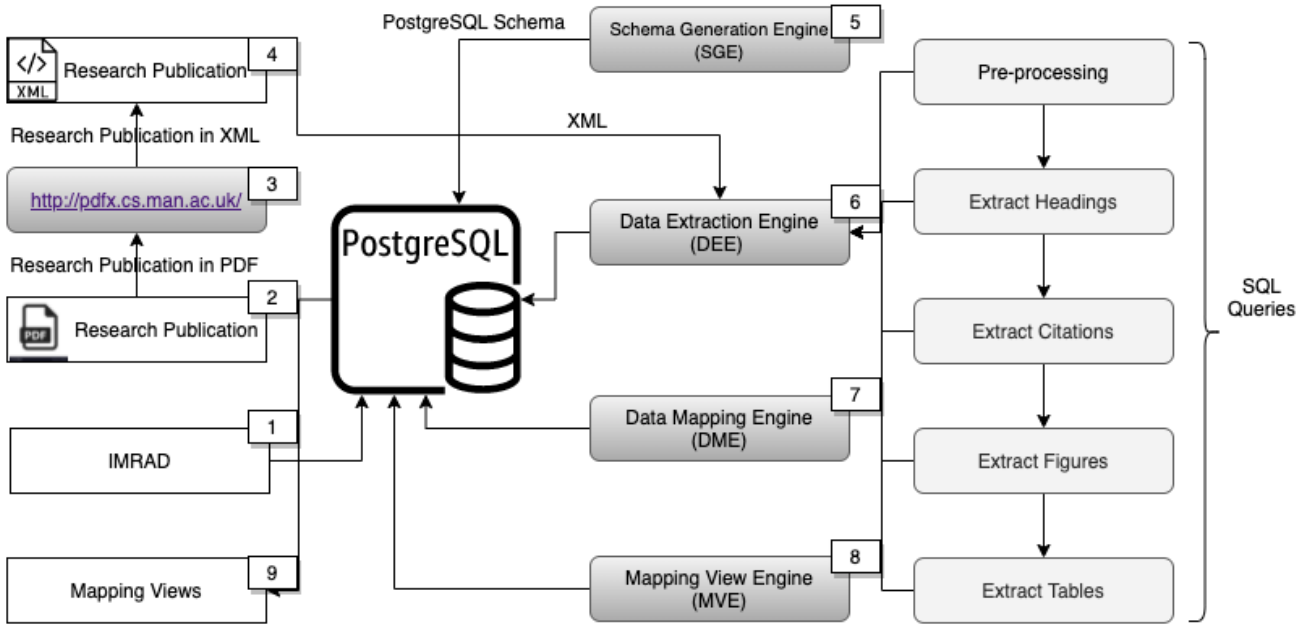


FIGURE 7. System design and component.

into XML using the PDFX [22] tool. The Schema Generation Engine (SGE) is used to generate schema of the XML files, which is maintained in PostgreSQL to parse and insert the XML data. Thereafter, Data Extraction Engine (DEE) is used to extract headings and subheadings from sections of research papers and other objects like citations, figures, and tables. The Mapping SQL Engine (MSE) maps the extracted headings and subheadings to IMRAD with the help of a devised algorithm 1. The last module, Mapping View Engine (MVE) is used to visualize the resultant mapping using XPath/XQuery expressions. In the end, the mapped sections are evaluated by using the benchmark data set that contains section annotations formed with the help of a user study. The overall structure of the proposed methodology is shown in Figure 7 and Figure 8. The proposed algorithm is formally represented bellow in form or different modules 1, 2 4, 5 6, 7. The detailed explanation of all the modules implemented in the proposed methodology is delineated in the following sections.

A. SCHEMA GENERATION ENGINE (SGE)

As explained earlier, research papers in the employed data set were in the form of PDF, which was converted into XML format. The requisite parameters from XML files are required to be stored in some meaningful format so that extensive queries could be processed on it. In this regard, we developed Schema Generation Engine (SGE) wherein we stored that information in different tables of the database and created links between them. The XML publications were inserted into the PostgreSQL, which is a renowned relational database formed using the generated schema. The schema is the true mapping of research publications to the relational databases. Since PostgreSQL is a very powerful XML engine used to manipulate XML data, therefore, we picked it to parse and insert the XML data.

B. DATA EXTRACTION ENGINE(DEE)

The Data Extraction Engine (DEE) is used to pre-process the data. While preprocessing, all the noise from the data is removed. By the term ''noise'', we mean those papers that do not follow the paragraph and heading syntax of research papers. In XML files, the tags of sub-headings are not properly nested within the tags of main headings, rather all are represented as independent headings. We have extracted all the tags from XML using DEE to determine their position (i.e, main heading, or subheading) in the paper. The DEE extracted the headings and subheadings using the XML heading structures and populated the database accordingly. Afterward, DEE identified citations and objects like figures, tables, etc. from the XML files. The algorithms 2 shows the process of data extraction from XML.

C. DATA MAPPING ENGINE (DME)

The previous module Data Extraction Engine extracts the data from XML and after some processing, it inserts data into a relational format in the database. The relation database format of the dataset provides ease to manipulate the data to make information extraction easy, but there is still a need to find a section and map these sections to IMRAD. The Data Mapping Engine(DME) uses XQuery/XPath/SQL queries to extract the section and mapped that section to IMRAD. By carefully examining the structure of the XML files, we have designed some rules to map the sections to IMRAD. The algorithm 5, 6 and 7 shows the process of mapping sections to IMRAD.

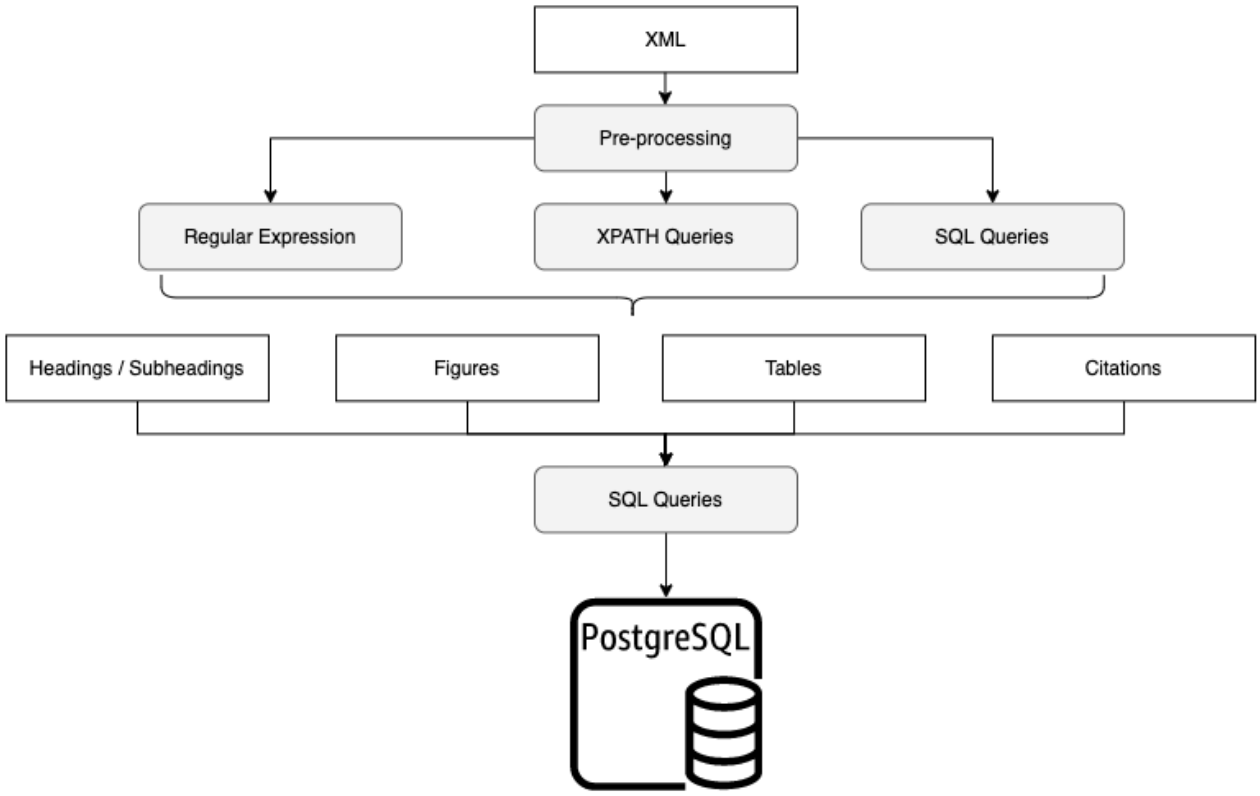


FIGURE 8. System flow.

1) SUBHEADINGS MAPPINGS

The research publications consist of heading and subheadings. It is really important to segregate the heading and subheadings. While mapping the heading to sections, some studies keep subheading into consideration, but some studies totally ignore to handle the subheadings. For instance, if a heading <h1> contains two sub-headings <h2> and <h3>, then all of three tags are considered as independent sections by the existing studies. Due to this negligence, IMRAD heading can be mistakenly mapped to the subheading.

Our proposed approach distinguishes between the main heading and subheading with the help of regular expressions 3. Unlike other studies, our approach does not map sub-heading to the IMRAD structure. The sub-headings are mapped to that section of IMRAD which is the parent heading of that sub-heading in the hierarchy. The proposed approach validates the following two aspects that have been overlooked by the scientific community.

1 - Section subheading is properly distinguishable in XML format by XML heading tags <h2><h3> etc.

2 - Section subheadings are not distinguishable in XML format, and need to run some regular expression 3 to distinguish heading by bullets.

The following regular expression is used to detect the headings and subheadings, in the scenarios wherein separate tags for headings were not present in XML.

We observed that in some cases headings and sub-headings are not explicitly defined in XML. In such scenarios, it becomes difficult to detect the starting and ending of the main headings<h1>. We have critically analyzed such XML files in the data set and found that XML's element ''region class=''DoCO:TextChunk'' can be used wherein headers are missing in XML files.

2) SECTIONS SEQUENCES

In recent years, the scholarly community follows IMRAD based rules while preparing research documents. They follow the rule of sequence. If the rules are properly followed then it becomes much feasible for IR systems to correctly identify the sections. However, we have critically analyzed PDF files from the employed data set and observed that some of the research papers have not followed the rule of sequence. The alternate methods are required to identify the sections from those papers that have not been structured according to the rules.

3) SECTIONS KNOWN NAMES

Typically, the scholarly community harnesses the below-mentioned names for the sections of research papers. The presence of these names makes it easier to identify the name of logical sections. The names include:

- Abstract
- Keywords
- Introduction
- Literature Review / History /Related work
- Methodology
- Results / Discussion
- Conclusions and future work
- References

In this study, we have also identified the synonyms of these section names. Although the scientific community follows the IMRAD structure but does not strictly adhere to use the same names as mentioned above. Usually, a researcher picks synonyms of the names as per his/her choice. For instance, as identified in PDF files of the employed data set, some authors have named ''Evaluation'' while some have named ''Analysis'' to the ''Results/Discussion'' section. We have extracted synonyms from all the 5000 research articles. Wherever the exact section name is not found, the proposed methodology matches the synonyms for section mapping.

4) CITATION COUNT

Citation is one of the powerful bibliometric indices, used to reveal facts regarding various aspects of scientific literature. A citation count is the number of occurrences a citation appears in the body of a document, also known as in-text citations. We have considered ''citation count'' as a measure to detect logical sections. This measure has been chosen based on an assumption that a certain count of a citation identified in a logical section may serve as a strong relevance clue. For instance, typically, the ''Literature Review'' section contains comparatively a greater number of in-text citations than other sections. Such a section can be mapped to the ''Literature Review/Related Work'' section.

The XML files contain elements like Xref, ref, section, etc. The Xref along with ref-type = 'bibr' exhibits in-text citation. This can be linked to the 'ref' tags via an attribute. After extracting all these tags, we computed the count of citations. Thereafter, the tag anchors were detected using the CAD [23].

5) THE OCCURRENCE OF OBJECTS

Most of the research articles contain multiple figures and tables. The XML files represent these figures and tables in the form of objects. We have picked these objects as a measure to identify logical sections based on the same notion of harnessing citation count measure. The results/discussion section contains comparatively a greater number of objects than other sections. These objects are detected using XML tags like ''TableBox'', and ''FigureBox'', as shown in Figure 4 and Figure 5.

D. MAPPING VIEW ENGINE (MVE)

The Mapping View Engine (MVE) is designed to visualize the resultant mapping for analysis. As discussed earlier, we mapped all the sections with IMRAD; then we need some views to get that information from the database. To get that data, we need to run some SQL/XPATH queries using regular expressions which is not always an easy way; therefore MVE provides different views of that results.

Algorithm 1 Section Mapping to IMRAD

Input: XML Publications Dataset
Result: IMRAD Mapped Sections

```
1:
2: for each xml ∈ XML do
3:     lh = exctractHeaders(xml_i) 2
4:     ml = mapSectionUsingDisctionary(lh, ml)
5:     ml = mapSectionUsingTemplate(lh, ml) A. Shahid
       and M. T. Afzal [2]
6:     ml = mapSectionUsingCitationFreq(lh, ml)
7:     ml = mapSectionUsingFigureFreq(lh, ml)
8:     ml = mapSectionUsingTableFreq(lh, ml)
9: end for
```

VI. RESULTS AND EVALUATION

Once all the modules of the proposed methodology have been implemented, the next steps involve the evaluation of data using the benchmark data set. The outcomes are evaluated in two phases: (1) We have identified a one-layer hierarchy of sub-headings and then mapped the sections to the IMRAD (2) Afterwards, we have determined collective contribution of all the parameters in a hybrid manner by combining the parameters ''citation count'', ''figures count'' and ''tables count''. The standard evaluation measures, including precision, recall, and f-measure have been employed. We calculated the precision/recall / f-measure of each section separately for both the aforementioned phases. Figure 12 shows the combined average score of precision, recall, and F-measure for all the sections, and figures 9, 10, and 11 represents the individual scores for each section. The formal description of the formulas of precision, recall, and f-measure is shown below.

$$\begin{aligned} P_x &= TP_x/(TP_x + FP_x) \\ R_x &= TP_x/(TP_x + FN_x) \\ \Rightarrow x_0 &= Introduction \\ \Rightarrow x_1 &= Methods \\ \Rightarrow x_2 &= Results \\ \Rightarrow x_3 &= Discussion \end{aligned} \tag{1}$$

$$P_{avg} = \sum_{x=0}^{3} P_x/4 \Rightarrow P_{avg} = Average\ Precision$$

$$R_{avg} = \sum_{x=0}^{3} R_x/4 \Rightarrow R_{avg} = Average\ Recall \tag{2}$$

Figure 9 shows the comparison of the precision of our approach with state-of-the-art approaches. It is quite evident from the graph that all three approaches mapped the introduction with high accuracy but failed to accurately map the other sections. On the other hand, our approach performed

Algorithm 2 extractHeader

Input: XML File: *xml*
Output: Header/Section List *hl*

```
T ← xml // extract text from xml

for each heading ∈ T do
    hl <= xPathQuery(getH1) // get Heading <h1>
    hl <= xPathQuery(getH2) // get Heading <h2>
    hl <= xPathQuery(getH3) // get Heading <h3>
    bl <= extBullets(hl) // get Bullets
    hl <= getSectionClass"DoCO : SectionTitle"
end for
```

Algorithm 3 exctractBullet

Input: Header/Section List *hl*
Output: Bullet List: *bl*

```
\?:^|
(?<=\s))\d\.?(?:\d+)?(?=\s)|
\*(?=\s)

return bl // Bullet List
```

Algorithm 4 mapSectionUsingDisctionary

Input: XML file:*xml*, Header List:*lh*
Result: Section Mapping List:*sml*, Header List:*lh*

```
for each l ∈ "lh" do

    for each imrad ∈ IMRAD do

        if l_j! = imrad_i then
            return sml // Section Mapping List
        end if
    end for
end for
```

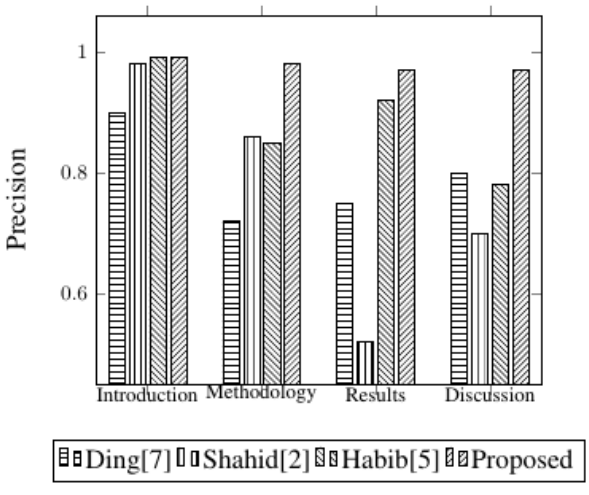


FIGURE 9. Section wise comparison of precision.

better and achieved almost similar precision in the mapping of all the sections. Further, the other approaches were unable to identify the Result or Discussion section. Similarly, Figure 10 shows the comparison of recall of our proposed approach with all three state-of-the-art approaches.

Since the proposed approach has an application in information retrieval (IR) systems is specifically concerned about returning most of the correct responses against a query,

Algorithm 5 mapSectionUsingCitationFrequncy

Input: XML file:*xml*, Section Mapping List: *sml*
Result: Section Mapping List:*sml*

```
for each l ∈ "lh" do
    (CT_i, CTP_i) = String-Citaion_Anchor_Detection(t) CAD [23]
    (CN_i, CNP_i) = Numeric-Citaion_Anchor_Detection(t) CAD [23]

    for each imrad ∈ IMRAD do

        if l_j! = imrad_i then

            if CF_i ≥ 50% then
                sml_discussion <= true; // A DISCUSSION Section
                return sml // Section Mapping List
            end if

        else

            if CF_i ≥ 30% then
                sml_intro <= true; // A INTRODUCTION Section
                return sml // Section Mapping List
            end if
        end if
    end for
end for
```

Algorithm 6 mapSectionUsingFigureFreq

Input: XML file:*xml*, Section Mapping List: *sml*
Result: Section Mapping List:*sml*

```
for each l ∈ "lh" do
    FP_i = getRegionClass"DCO : FigureBOX"

    for each imrad ∈ IMRAD do

        if l_j! = imrad_i then

            if FP_i ≥ 60% then
                sml_results <= true; // A RESULT section
                return sml // Section Mapping List
            end if

        else

            if FP_i ≥ 30% then
                sml_meth <= true; // A METHODOLOGY Section
                return sml // Section Mapping List
            end if
        end if
    end for
end for
```

Algorithm 7 mapSectionUsingTableFrequncy

Input: XML file:*xml*, Section Mapping List: *sml*
Result: Section Mapping List:*sml*

```
for each l ∈ "lh" do
    TP_i = getRegionClass"DCO : TableBox"
    for each imrad ∈ ℐMRAD do
        if l_j! = imrad_i then
            if TP_i ≥ 70% then
                sml_results <= true; // A RESULT section
                return sml // Section Mapping List
            end if
        else
            if FP_i ≥ 20% then
                sml_meth <= true; // A METHODOLOGY Section
                return sml // Section Mapping List
            end if
        end if
    end for
end for
```

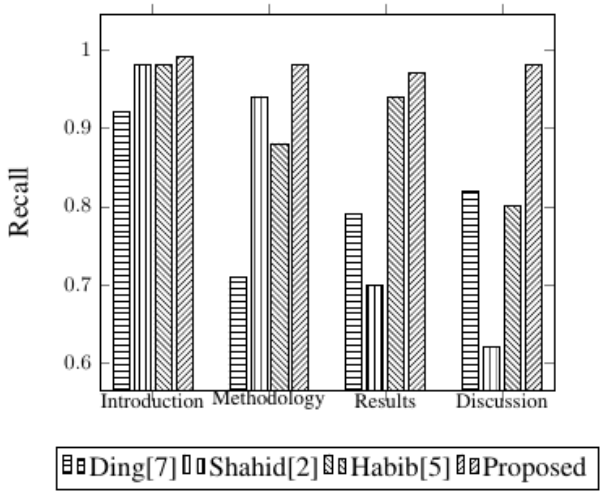


FIGURE 10. Section wise comparison of recall.

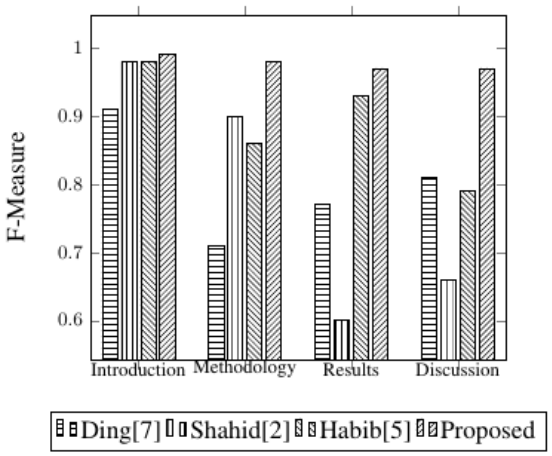


FIGURE 11. Section wise comparison of F-Measure.

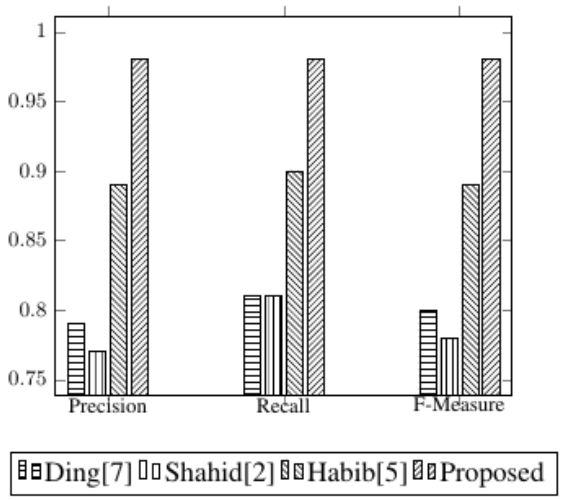


FIGURE 12. Comparison of average precision of combined sections.

therefore, we have drawn comparisons here on the basis of the precision score. The evaluation results of our approach illustrate the increased value of precision, and when all the features were assessed in a hybrid manner then the proposed approach tremendously outperformed the approaches [2], [6], [8]. We believe that this improvement is due to the consideration of objects in XML files.

The analysis discussed above revealed that our study has outperformed all of the three existing state-of-the-art studies on section mapping by achieving 8.96%, 21.77%, and 9.50% improvement from Ding *et al.* [8], Habib and Afzal [6] and Shahid and Afzal [2], respectively. This signifies the potential of all the novel parameters like In-Text Citation count, Figure count, and Table count. We argue that the worth of XML objects must not be overlooked while section detection. The collective contribution of all the employed parameters and potential of designed regular expressions can adequately cope up to overcome the deficiencies in the existing state-of-the-art in section mapping.

VII. PERFORMANCE EVALUATION COMPARISON

We have evaluated the performance of our approach with state-of-the-art techniques. All experiments were performed in the same environment on AWS medium instances. We have performed experiments in multiple iterations and reported the outcome in median values. The table 2 shows the number in seconds. The approach of Ding *et al.* [8] consumed comparatively less time than all other approaches. This is due to the fact that Ding *et al.* [8] have only used dictionary terms to identify sections and mapped them to IMRAD. Shahid and Afzal [2] took a bit longer than Ding *et al.* [8] because Shahid's approach also uses templates along with the dictionary terms. The Habib and Afzal [6] took longer than the previous two approaches because counting the citations and parsing the whole text take a lot of time. Most of the time

TABLE 2. Performance comparison with start-of-the-art techniques.

Techniques	Time in Seconds
Ding et al [8]	10
A. Shahid and M. T. Afzal [2]	14
Habi et al [6]	21
Proposed	34

is consumed by our proposed approach because it applies multiple techniques for section identification and mapping. Moreover, it is philosophically true because our approach uses all the features employed by the state-of-the-art techniques [2], [6], [8] along with the novel proposed features. The whole process was done offline, and we have maintained all the information in the database. Although the proposed technique takes more time than state-of-the-art techniques; however, this is not important because all of this processing will be done on the backend and all of these results will be precompiled before they are made available to be used in the online system. This indicates that when a user performs such a query then there is no such difference between the time taken by any of the approaches. Instead, all of the approaches will service the user based on the processed data.

VIII. CONCLUSION

This paper has presented a novel method to map logical sections of PDF formatted research papers to the IMRAD structure. Unlike contemporary state-of-the-art, the approach has been designed after a critical manual investigation of one of the most recent approaches of logical sections mapping. Our study has identified a set of novel features and justified their potential by evaluating them on a benchmark data set. These features include: (1) In-Text Citation count (2) Figure count and (4) Tables count. The study generated XML files from PDF files using PDFX [22] and exploited XML headers and objects with the help of regular expressions to detect logical sections, sub-sections, and mapping them to their corresponding section of IMRAD. The study has outperformed contemporary studies by attaining 0.97 precision (i.e., 0.85 vs. 0.97) and recall 0.98 (i.e. 0.86 to 0.98) This improved precision and recall is due to the incorporation of the features that have been overlooked by the contemporary state-of-the-art. The tool PDFX was unable to process approximately 10% of the PDF files into XML files. In the future, attention may be paid to initiate methods of section identification in the scenarios wherein PDFX [22] fails.

ACKNOWLEDGMENT

The authors acknowledge Ms. Faiza Qayyum for helping them in linguistics review of this article and giving feedback to make it readily understandable.

REFERENCES

[1] P. O. Larsen and M. von Ins, "The rate of growth in scientific publication and the decline in coverage provided by science citation index," *Scientometrics*, vol. 84, no. 3, pp. 575–603, Mar. 2010, doi: 10.1007/s11192-010-0202-z.

[2] A. Shahid and M. T. Afzal, "Section-wise indexing and retrieval of research articles," *Cluster Comput.*, vol. 21, no. 1, pp. 481–492, May 2017, doi: 10.1007/s10586-017-0914-4.

[3] L. B. Sollaci and M. G. Pereira, "The introduction, methods, results, and discussion (IMRAD) structure: A fifty-year survey," *J. Med. Library Assoc.*, vol. 92, no. 3, pp. 364–367, Jul. 2004.

[4] G. Batmanabane, "The IMRAD structure," in *Reporting and Publishing Research in the Biomedical Sciences*. Singapore: Springer, Dec. 2017, pp. 1–4, doi: 10.1007/978-981-10-7062-4_1.

[5] C. Lange and A. Di Iorio, "Semantic publishing challenge—Assessing the quality of scientific output," in *Semantic Web Evaluation Challenge*. Cham, Switzerland: Springer, 2014, pp. 61–76.

[6] R. Habib and M. T. Afzal, "Sections-based bibliographic coupling for research paper recommendation," *Scientometrics*, vol. 119, no. 2, pp. 643–656, Mar. 2019, doi: 10.1007/s11192-019-03053-8.

[7] A. Y. Khan, A. Shahid, and M. T. Afzal, "Extending co-citation using sections of research articles," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 26, no. 6, pp. 3346–3356, Nov. 2018.

[8] Y. Ding, X. Liu, C. Guo, and B. Cronin, "The distribution of references across texts: Some implications for citation analysis," *J. Informetrics*, vol. 7, no. 3, pp. 583–592, Jul. 2013, doi: 10.1016/j.joi.2013.03.003.

[9] B. Fecher and S. Friesike, *Open Science: One Term, Five Schools of Thought*. Cham, Switzerland: Springer, 2014, pp. 17–47.

[10] S. Teufel, A. Siddharthan, and C. Batchelor, "Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, vol. 3. Singapore: Association Computational Linguistics, 2009, pp. 1493–1502. [Online]. Available: https://www.aclweb.org/anthology/D09-1155

[11] S. Teufel and M. Moens, "Summarizing scientific articles: Experiments with relevance and rhetorical status," *Comput. Linguistics*, vol. 28, no. 4, pp. 409–445, Dec. 2002, doi: 10.1162/089120102762671936.

[12] M. Liakata, S. Teufel, A. Siddharthan, and C. Batchelor, "Corpora for the conceptualisation and zoning of scientific papers," in *Proc. 7th Int. Conf. Lang. Resource Eval. (LREC)*. Valletta, Malta: European Language Resources Association (ELRA), May 2010, pp. 1–9.

[13] S. Teufel, "Citations and sentiment," in *Proc. Workshop Text Mining Scholarly Commun. Repositories*. Univ. Manchester: Manchester, U.K., 2009, pp. 1–56.

[14] S. Maričić, J. Spaventi, L. Pavičić, and G. Pifat-Mrzljak, "Citation context versus the frequency counts of citation histories," *J. Amer. Soc. Inf. Sci.*, vol. 49, no. 6, pp. 530–540, 1998, doi: 10.1002/(sici)1097-4571(19980501)49:6<530::aid-asi5>3.0.co;2-8.

[15] S. Peroni, D. Shotton, and F. Vitali, "Faceted documents," in *Proc. ACM Symp. Document Eng. (DocEng)*. New York, NY, USA: ACM Press, 2012, pp. 191–194, doi: 10.1145/2361354.2361396.

[16] Y. Guo, A. Korhonen, M. Liakata, I. Silins, J. Hogberg, and U. Stenius, "A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment," *BMC Bioinf.*, vol. 12, no. 1, p. 69, 2011.

[17] Ş. Kafkas, X. Pi, N. Marinos, F. Talo, A. Morrison, and J. R. McEntyre, "Section level search functionality in Europe PMC," *J. Biomed. Semantics*, vol. 6, no. 1, p. 7, 2015, doi: 10.1186/s13326-015-0003-7.

[18] S. Nazir, M. Asif, S. Ahmad, F. Bukhari, M. T. Afzal, and H. Aljuaid, "Important citation identification by exploiting content and section-wise in-text citation count," *PLoS ONE*, vol. 15, no. 3, Mar. 2020, Art. no. e0228885, doi: 10.1371/journal.pone.0228885.

[19] S.-U. Hassan, M. Imran, S. Iqbal, N. R. Aljohani, and R. Nawaz, "Deep context of citations using machine-learning models in scholarly full-text articles," *Scientometrics*, vol. 117, no. 3, pp. 1645–1662, Oct. 2018, doi: 10.1007/s11192-018-2944-y.

[20] D. Pride and P. Knoth, "Incidental or influential?—Challenges in automatically detecting citation importance using publication full texts," in *Research and Advanced Technology for Digital Libraries*. Springer, 2017, pp. 572–578. [Online]. Available: https://doi.org/10.1007/978-3-319-67008-9_48

[21] J. Beel and B. Gipp, "Google scholar's ranking algorithm: An introductory overview," in *Proc. 12th Int. Conf. Scientometrics Informetrics (ISSI)*, Rio de Janeiro, Brazil, Jul. 2009, pp. 230–241.

[22] A. Constantin, S. Pettifer, and A. Voronkov, "PDFX," in *Proc. ACM Symp. Document Eng. (DocEng)*. New York, NY, USA: ACM Press, 2013, pp. 177–180, doi: 10.1145/2494266.2494271.

[23] R. Ahmad and M. T. Afzal, "CAD: An algorithm for citation-anchors detection in research papers," *Scientometrics*, vol. 117, no. 3, pp. 1405–1423, Sep. 2018, doi: 10.1007/s11192-018-2920-6.

IBRAR AHMED received the master's degree in computer science from International Islamic University Islamabad and the M.S. degree in computer engineering from the University of Science and Technology, Taxila. He is currently pursuing the Ph.D. degree in computer science. He is working as a Database Specialist at the Research and Development Organization. He has authored multiple books in the database field. His research interests include databases, digital libraries, and data science.

MUHAMMAD TANVIR AFZAL received the M.Sc. degree in computer science from Quaid-i-Azam University, Islamabad, Pakistan, and the Ph.D. degree (Hons.) in computer science from the Graz University of Technology, Austria.

He has been associated with academia and industry at various levels for the last 20 years. He is currently a Professor with the Department of Computer Science, Capital University of Science and Technology, Islamabad. He has authored more than 100 research papers in leading peer-reviewed journals and conferences in the field of data science, information retrieval and visualization, semantics, digital libraries, and scientometrics. He has authored two books and has edited two books in computer science. His cumulative impact factor is more than 60, with citations over 500. He played a pivotal role in making collaborations between MAJU-JUCS, MAJU-IICM, and TUG-UNIMAS. He received the Gold Medal for his M.Sc. He has served as the Ph.D. Symposium Chair, the Session Chair, the Finance Chair, a Committee Member, and an Editor of several IEEE, ACM, Springer, Elsevier international conferences, and journals. He is also serving as the Editor-In-Chief of a reputed impact factor journal known as the *Journal of Universal Computer Science*. He conducted more than 100 curricular, co-curricular, and extra-curricular activities in the last five years, including seminars, workshops, and national competitions (ExcITeCup), and invited international and national speakers from Google, Oracle, IICM, IFIS, SEGA Europe, and so on. Under his supervision, more than 60 postgraduate students (M.S. and Ph.D.) have defended their research theses successfully and a number of Ph.D. and M.S. students are pursuing their research with him.