Computers, Materials & Continua
DOI:10.32604/cmc.2021.014156
Article

Tech Science Press

SwCS: Section-Wise Content Similarity Approach to Exploit Scientific Big Data

Kashif Irshad[1], Muhammad Tanvir Afzal[2], Sanam Shahla Rizvi[3], Abdul Shahid[4], Rabia Riaz[5] and Tae-Sun Chung[6,*]

[1]Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan
[2]Department of Computer Science, NAMAL Institute, Mianwali, 42250, Pakistan
[3]Raptor Interactive (Pty) Ltd., Eco Boulevard, Witch Hazel Ave, Centurion, 0157, South Africa
[4]Institute of Computing, Kohat University of Science and Technology, Pakistan
[5]Department of CS&IT, University of Azad Jammu and Kashmir, Muzaffarabad, 13100, Pakistan
[6]Department of Artificial Intelligence, Ajou University, Korea
[*]Corresponding Author: Tae-Sun Chung. Email: tschung@ajou.ac.kr

Received: 02 September 2020; Accepted: 28 October 2020

Abstract: The growing collection of scientific data in various web repositories is referred to as Scientific Big Data, as it fulfills the four "V's" of Big Data—volume, variety, velocity, and veracity. This phenomenon has created new opportunities for startups; for instance, the extraction of pertinent research papers from enormous knowledge repositories using certain innovative methods has become an important task for researchers and entrepreneurs. Traditionally, the content of the papers are compared to list the relevant papers from a repository. The conventional method results in a long list of papers that is often impossible to interpret productively. Therefore, the need for a novel approach that intelligently utilizes the available data is imminent. Moreover, the primary element of the scientific knowledge base is a research article, which consists of various logical sections such as the Abstract, Introduction, Related Work, Methodology, Results, and Conclusion. Thus, this study utilizes these logical sections of research articles, because they hold significant potential in finding relevant papers. In this study, comprehensive experiments were performed to determine the role of the logical sections-based terms indexing method in improving the quality of results (i.e., retrieving relevant papers). Therefore, we proposed, implemented, and evaluated the logical sections-based content comparisons method to address the research objective with a standard method of indexing terms. The section-based approach outperformed the standard content-based approach in identifying relevant documents from all classified topics of computer science. Overall, the proposed approach extracted 14% more relevant results from the entire dataset. As the experimental results suggested that employing a finer content similarity technique improved the quality of results, the proposed approach has led the foundation of knowledge-based startups.

Keywords: Scientific big data; ACM classification; term indexing; content similarity; cosine similarity

1 Introduction

The myriad of scientific research publication over the web has been increasing over the past several years [1]. This knowledge base is generated by numerous researchers worldwide, and the scientific documents are published in different journals, conferences, workshops, etc. As this scientific data can be described in terms of the "four V's" of Big Data: volume (huge repositories are available), variety (different venues require their own format), velocity (increasing with rapid pace), and veracity (content extraction from PDF versions may be noisy). This expanding knowledge base can be referred to as Scientific Big Data. According to a recent analysis of Scientific Big Data, more than 50 million journal papers and billions of conference papers have been published; in addition, 1.3 billion books have been digitized by Google [2]. These documents are then indexed in various digital repositories such as Web of Science, SCOPUS, and PubMed. As of October 2017, PubMed contained 27.5 million records, representing approximately 7000 journals [3]. Further, SCOPUS indexes 75 million records [4] and Google Scholar indexes 389 million documents [5]. Thus, identifying pertinent research papers from such huge repositories is an immense challenge. Generally, thousands of papers are returned from these systems for a user query.

This significant amount of data on different web repositories hinders the process of retrieving relevant information in a concrete manner. Millions of generic hits and irrelevant documents are returned by contemporary indexers, posing a challenging task for researchers. Consequently, the problem has grabbed the attention of scholarly communities, and researchers are in the process of developing effective solutions. Subsequently, the solutions to such problems may lead to the foundation of new and emerging startups. Therefore, the community is seeking solutions through various perspectives such as by exploiting citation, metadata, collaborative filtering, and content-based approaches.

Citations are deemed to be a great source of information for recommending relevant papers. Researchers have proposed various citation-based techniques, including bibliographic coupling [3] and co-citation [4]. Kessler used a bibliographic connection as basic similarity metrics [3] to locate groups or clusters in technical and scientific literature. In particular, authors carefully screen the citations of their papers, which is advantageous because there is a high probability that most of the articles cited in the references list will be relevant to the topic of the citing paper. However, citation-based approaches do not yield proper results for articles that were not referred because authors cannot cite all the relevant papers.

In addition, metadata-based techniques, suggested by [5,6], uses various types of metadata such as paper title, authors, and venues for finding relevant documents. In this method, the discovery and use of documents is characterized by metadata that help in document discovery, network management, visibility, and organizational memory [7].

Another widely used approach to obtain relevant documents is collaborative filtering, which determines relevant documents by utilizing collaborative knowledge. These recommendations are based on user profiles and past preferences of the user's taste [8]. The collaborative filtering system provides better accuracy than the content-based approach, but suffers from the item cold-start problem. The authors have suggested using a unified Boltzmann machine that naturally combines content and collaborative features to produce better recommendations for both cold-start and non-cold-start items.

In content-based approaches, the content of two individual papers is analyzed to determine their relevance with respect to each other. For example, papers with contents similar to that of focused paper "A" will be considered more relevant for paper "A" [9]. More precisely, the content-based approach extracts important terms from the contents of two papers and compares them to find the relevance between the papers.

In context, IMRAD (Introduction, Method, Results, and Discussion) is a common structure for organizing a research article, which was introduced by Louis Pasteur [10]. Its adoption began in the 1940s, and it became the dominant format for preparing papers in the 1980s [11]. Subsequently, different detailed structures were developed. For example, Shotton proposed an ontology (discourse elements ontology) that conceptually describes the logical sections and other related information pertaining to scientific documents [12,13]. In addition, the technique developed and proposed in [14] maps the diversified section names over the logical sections of a research document. Their proposed approach does not depend on the regular flow of sections (Introduction, Related Work, Methodology, Results, Conclusion, and Summary), rather it uses paper template information (possible positioning of sections) and the dictionary terms of sections. Besides, every section in a paper has its own meaning. For example, the authors provide an overview of their research in the "Abstract" section, whereas the "Introduction" section contains a brief introduction about the focused research topic. An overview of related previous research as well as the problems and deficiencies in the existing techniques are generally described in the "Related work" section. Subsequently, the "Methodology" section contains the architecture of the proposed solution along with a detailed description of the proposed technique. Then, the obtained results are analyzed and discussed in the "Result" section of the document. Finally, the findings from the research are presented in the "Conclusion" section.

As discussed above, each logical section has its own importance and significance; however, the entire document is treated at the same level of importance in standard content-based approaches. For example, if a term in the "Abstract" section of paper "A" matches with a term in the conclusion section of paper "B," the standard content-based approaches will declare that paper "B" is relevant to paper "A," although these documents might not be actually relevant. On the contrary, if the importance of logical sections are defined and a term from the "Abstract" section of paper "A" matches to a term in the "Abstract" section of paper "B," then the probability of the two documents being relevant may increase.

Similarly, if two researchers have independently proposed two different algorithms to solve the same problem in papers "A" and "B," respectively, then there is a greater possibility of a higher number of matching terms between the methodology sections of both papers. However, a survey paper "C" covering the same problem area might have a higher number of matched terms with paper "A" in the entire content of the paper. In this case, standard content-based approaches will consider survey paper "C" to be more relevant to paper "A" than paper "B; " however, papers "A" and "B" are more related in reality. Thus, considering all of the aforementioned issues, we present a study that identifies *whether section-wise content similarity increases the chances of recommending relevant papers, as compared to the conventional content-based approach*.

Therefore, we performed a section-wise content comparison between research papers to address the objectives of the study. The vectors of each logical section were formed from each scientific paper. Furthermore, the corresponding vectors were compared in a standard manner using cosine similarity, as in the content-based approach. This indicated that the terms appearing in each logical section were more likely to be compared only with the terms occurring in the corresponding logical section of the other papers. The proposed approach was comprehensively evaluated by accounting data from each topic under the ACM classification hierarchy. The section-based approach outperformed content-based approach in identifying relevant documents in all topics of computer science. Further, the gain percentage varied from 36% for Topic-E "Data" and 2% for the Topic-D "Software. " Overall, the gain percentage of the proposed approach was 14% for all dataset.

The rest of the paper is organized as follows. The literature review is presented in Section 2 for the validation of the framed hypothesis. The proposed methodology of the study is presented in Section 3. The evaluation of the experiment is elaborated in Section 4 with the experimental setup and considerations, where a sample paper from the classified ACM topics was selected. The experimental results and its comparisons are presented in Section 5, and the results are discussed in Section 6. Finally, the contributions of this study are summarized and concluded in Section 7.

2 Related Work

In the above section, the extent of research publications and the estimated quantity of scientific documents were discussed along with commonly faced problems. Consequently, various approaches have been proposed in the literature to help the scientific community in this task. These contemporary approaches are divided into four major categories. The first approach identifies and recommends relevant papers by utilizing user collaborations; the second approach uses the metadata of the papers; the third approach uses citations to identify relatedness between documents; and the fourth approach exploits the content of papers to recommend relevant research papers. Certain hybrid systems utilize two or more of these techniques to enhance the recommendations for relevant papers based on different considerations. This section reviews the most important, recent, and classical methods related to these approaches.

2.1 Collaborative Filtering-Based

Collaborative filtering-based approaches list relevant documents by exploiting user profiles and past preferences of the users' choices. Recommender systems envisage a user's choices and preferences based on his/her accessed and rated items. Thus, collaborative filtering can be categorized into two approaches: model-based and memory-based. In model-based approaches, predictions are made based on the model, which contains information on items and users' interactions with each other; whereas memory-based approaches exploit the user's existing rating data for predicting their preference to other items. Therefore, collaborative filtering is considered an important approach because of its high performance and simple requirements, as highlighted by [15,16].

Currently, search engines used for academic purposes, such as Scienstein, have become influential and prominent hybrid systems recommending research.

2.2 Metadata-Based

Another approach to recommend relevant papers is that based on metadata, where the metadata of a paper, such as paper title, author names, publication date, and venues are used to extract relevant documents. Thus, the relevance and discovery of documents is characterized using metadata. Moreover, one of the core services provided to users is the creation and provision of metadata that support the functionality of various digital libraries. In particular, objects of interest and relevant information are accessed utilizing a metadata technique. As the documents and metadata are digital, the alternative implementation of the data can be made accessible. However, conventional metadata are less likely to exist in digital libraries. Metadata in digital libraries help in document discovery, network management, visibility, and organizational memory [17].

Thus, metadata are an important source for creating recommendations of relevant research papers. Moreover, recommendation systems based on metadata work efficiently, mostly because only a few terms need to be analyzed for recommending relevant research papers. Consequently, as the metadata are a set of small number of terms—generated from authors' keywords, title terms, and categories, the quality of recommendation is not highly accurate owing to the difficulty of the recommender system to make concrete decisions by analyzing a small number of terms. Therefore, several authors have developed hybrid approaches that use metadata as well as collaborative filtering, content, and citations to make accurate recommendations.

2.3 Citation-Based

Citations form a highly important dataset that use various techniques for recommending relevant research papers. The two common techniques pertaining to this field are bibliographic coupling [6] and co-citation [7]. In the former approach, two papers, A and B, are considered similar if they share a citation of paper C in their references. In the co-citation technique, the number of common citations received by two given papers, A and B, is used as an indicator of similarity between A and B. Therefore, co-citation between papers A and B indicate that they have been cited by paper C. Moreover, the citations techniques have been extended by various researchers in recent times [18,19].

Most of the citation-based techniques use citation network information. These techniques provide adequate and appropriate recommendations, because the citations are carefully hand-picked by the authors. However, these approaches are limited to working well within certain citation networks, because the authors cannot cite every relevant paper in their research. The relevant papers that have not been cited become a weak candidate to be discovered with relevance. Thus, it has considerably high chances of missing relevant papers.

2.4 Content-Based

In content-based approaches, the contents of two papers are analyzed to determine their relevance. For example, papers with a more similar content to that of focused paper A will be considered more relevant to paper A. Thus, content-based recommendation systems analyze the internal content of the documents to recommend relevant papers [20]. This method is used by the majority of literature reviews that compare the content of certain research papers to recommend relevant scientific documents. A given paper of any file format (e.g., pdf and doc) is transformed into text format having any one typographical case (e.g., lower case), and cleaned by removing stop words. In addition, known standard abbreviations are expanded using their full text, and term-frequency vectors using term frequency–inverse document frequency (TF–IDF) vectors are generated. Thereafter, the TF–IDF is used to determine the similarity between documents. As existing techniques do not use the internal logical sections of research articles, the current study conducted experiments to evaluate this novel approach on detecting the similarity and relevance between documents.

3 Proposed Methodology

This section comprehensively delineates the proposed methodology of the current study; the architecture of the proposed methodology is presented in Fig. 1. This methodology comprises various logical steps including (a) selection of a dataset of research articles, (b) section-wise and complete extraction of text from PDFs, (c) section-wise and complete indexing of terms using Apache Lucene, (d) computing similarity between documents using cosine similarity based on the terms indexed from both approaches, and (e) comparing both experimental results. Each step of the proposed methodology is elaborated in the following sections.

3.1 Comprehensive Dataset Selection

The dataset selection for the proposed approach included the criteria of a dataset covering a vast number of topics and being comprehensive enough to conclude the research. In addition, the dataset should allow us to access the logical sections of the papers for section-wise term extraction and matching. Based on these requirements, we selected the dataset of the *Journal of Universal Computer Science* (J. UCS), as papers from all topics of computer science are published in J. UCS. Moreover, it is one of the most comprehensive journals in the computer science domain, publishing research articles of authors from various fields and backgrounds [21,22]. Therefore, the selected dataset will play a consequential role in comprehensively investigating the proposed research. The J. UCS dataset is presented on the top row of Fig. 1.

3.2 PDF-to-Text Conversion

The text from the PDF files available at the J. UCS server was required to utilize its content. Thus, the PDF files were converted into XML format using a tool named PDFx with the approach adopted in [23]. Upon selecting the dataset and acquiring the XML files of every paper along with its logical sections, the content- and section-based approaches were performed to obtain highly ranked papers.

3.3 Content-Based Approach

The content-based approach had been implemented because the proposed technique was extended from it. The standard implementation of the content-based approach was performed using the Apache Lucene API that is widely applied to identify content/word similarities [24,25], as it extracts rare terms (based on TF–IDF), and computes the cosine similarities between research papers. As shown on the right-hand side of Fig. 1, the following steps were performed to implement the content-based approach.

3.3.1 Extracting Important Terms

The first step of implementing the content-based approach involves the acquisition of important terms from a research paper, and inputting the text files to the Apache Lucene API, where the term extractor, TF–IDF, extracts terms from the entire document using Eq. (1); this process is repeated for all the documents in the dataset. The TF–IDF prioritizes the rare terms of a document. For example, if term "T1" frequently occurs in document "D1," but does not frequently occur in other documents "D2" to "Dn," then term "T1" is considered important for document "D1." Conversely, term "T" is not considered unimportant when it frequently occurs in all documents "D1" to "Dn."

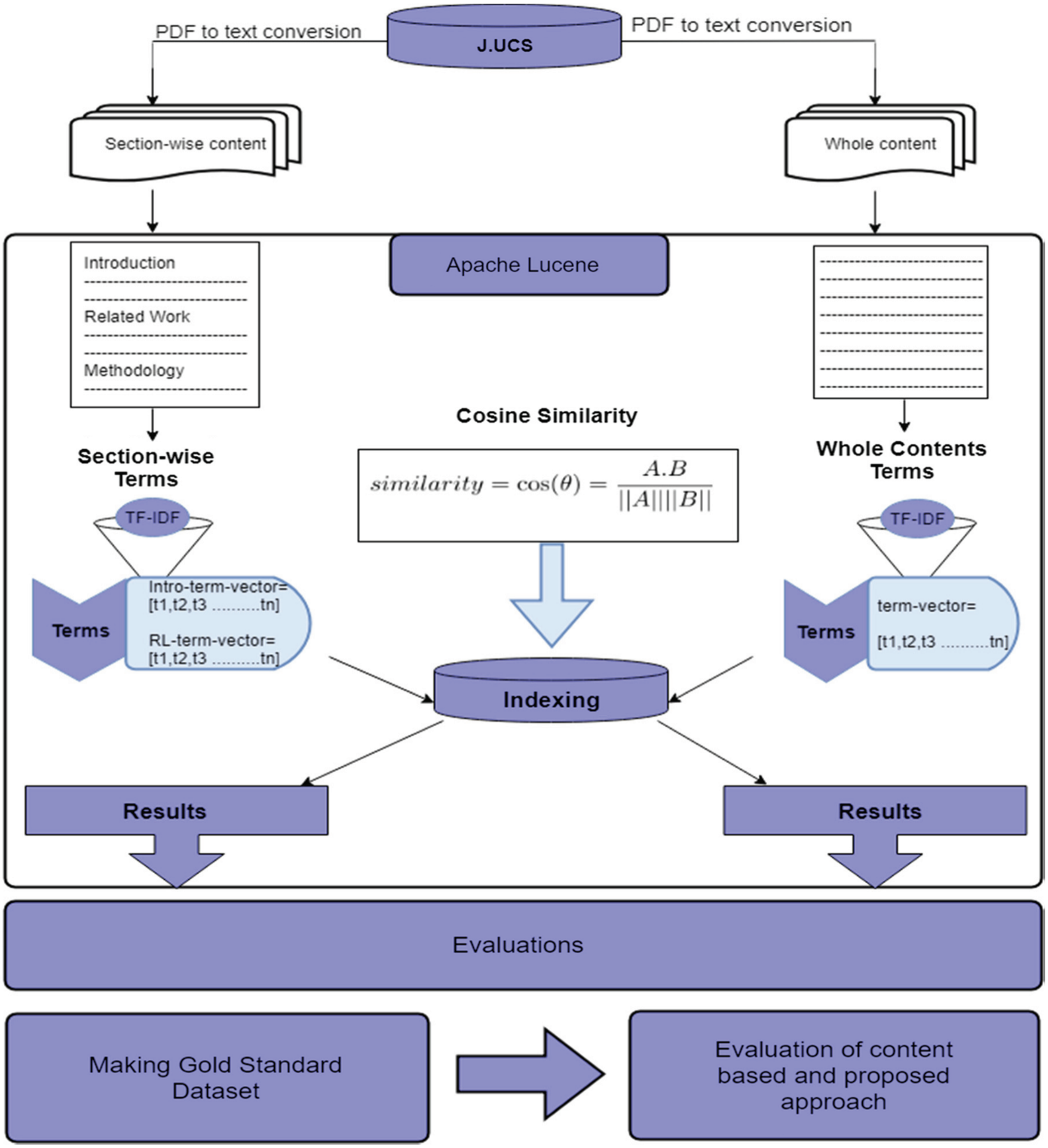

J.UCS
PDF to text conversion
PDF to text conversion
Section-wise content
Whole content
Apache Lucene
Introduction
Related Work
Methodology
Section-wise Terms
Cosine Similarity
$similarity = \cos(\theta) = \frac{A.B}{||A||||B||}$
Whole Contents Terms
TF-IDF
TF-IDF
Terms
Intro-term-vector=[t1,t2,t3 ....tn]
RL-term-vector=[t1,t2,t3 ....tn]
Terms
term-vector=[t1,t2,t3 ....tn]
Indexing
Results
Results
Evaluations
Making Gold Standard Dataset
Evaluation of content based and proposed approach

Figure 1: Overall process architecture for comparing the content- and section-based approaches

This measure has been extensively used by the scientific community to select important terms from text documents [26,27]. The extracted terms are indexed and stored in the database, as shown on the right-hand side of Fig. 1.

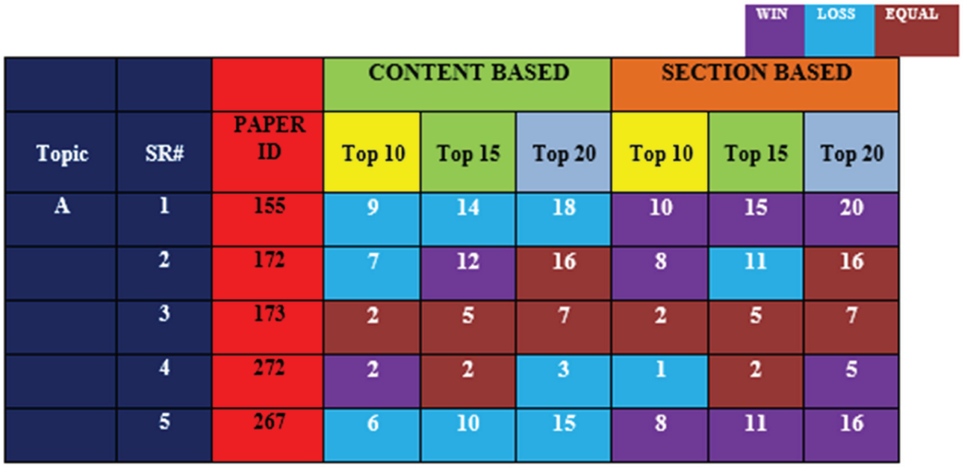
$$tfidf\,(t, d, D) = tf\,(t, d) * idf\,(t, d) \tag{1}$$

3.3.2 Ranking Papers on Content Similarity

The cosine similarity measure is widely applied to compute content similarity between research documents [28,29]. The terms of documents "D1" and "D2," respectively, are represented as vectors "A" and "B" in Eq. (2).

$$\textit{Document similarity} = \cos\theta = \frac{A \cdot B}{||A||\,||B||} \tag{2}$$

This similarity measure is available in Apache Lucene. The cosine similarities of each document with all other documents of the dataset were computed. Moreover, the ranked list of other similar research documents was retrieved based on the descending scores for each document, as shown on the right-hand side of Fig. 2. The cosine similarity is a fundamental measure that computes the similarity between two datasets in form of vectors. In this study, the research documents were represented as input data vectors in form of key terms. The representative terms of the documents were extracted using TF–IDF, which is considered as a benchmark technique for extracting key terms from documents. Besides, there are certain other techniques for key-term extraction, such as KEA, Yahoo Key-Term Extractor, and Alchemy API that can be used for similar purposes. However, applying all or some of them may lead to a different research question that involves evaluating the effect of various key-term extractors on the quality of results. Overall, TF–IDF and cosine similarity are considered as default mechanisms for evaluating content-based similarity. Therefore, the focus of this study was to evaluate the standard content-based and section-wise content similarities to determine the role of sections in finding relevant papers.



🟪 WIN	🟦 LOSS	🟫 EQUAL

			CONTENT BASED			SECTION BASED		
Topic	SR#	PAPER ID	Top 10	Top 15	Top 20	Top 10	Top 15	Top 20
A	1	155	🟦 9	🟦 14	🟦 18	🟪 10	🟪 15	🟪 20
	2	172	🟦 7	🟪 12	🟫 16	🟪 8	🟦 11	🟫 16
	3	173	🟫 2	🟫 5	🟫 7	🟫 2	🟫 5	🟫 7
	4	272	🟪 2	🟫 2	🟦 3	🟦 1	🟫 2	🟪 5
	5	267	🟦 6	🟦 10	🟦 15	🟪 8	🟪 11	🟪 16

Figure 2: Top-10, top-15, and top-20 results for topic-A

3.4 Section-Wise Content-Based Approach to Find Similar Documents

The same dataset was converted into logical sections to apply a section-wise content-based approach in finding relevant documents. The steps followed in this approach included certain additional tasks, which are shown on the left-hand side of Fig. 2 and described as follows.

3.4.1 Extracting Sections of Research Papers

In the current study, all the text files were converted into six logical sections: “Abstract,” “Introduction,” “Related work,” “Methodology,” “Results,” and “Conclusion.” The section headings appearing in the research papers were converted into these logical sections using the approach proposed in [14]. Although a paper consists of various logical sections, such as the “Abstract,” “Introduction,” “Literature Review,” “Proposed Work,” “Results,” and “Conclusion,” these sections are not explicitly mentioned in any article. Therefore, the proposed approach extracts the sections instanced in the papers and maps them over the logical sections. This task is achieved using the template information pertaining to the paper and terms dictionary. The template portrays sequential information regarding the article, such as the “Introduction” is the first section of the paper, and the “Results” section occurs before the “Conclusion” section. Similarly, the terms dictionary refers to the various terms that are commonly used for representing a certain section, e.g., literature review, related work, and experimental setup. The left-hand side of Fig. 2 depicts the accuracy of the stated approach at 78%, which was improved up to nearly 100% by manually checking the content of the logical sections from each source article. Although the accuracy of the stated approach was low, it still performed various prerequisite tasks such as extraction of section instances (i.e., actual section labels) from a research article and conducted its mapping over the predefined logical sections. As the objective of this study was to evaluate the performance of section-wise similarity results with a trivial approach, we preferred to adopt the technique proposed by [14] for convenience.

3.4.2 Extracting Section-Wise Important Terms

The process of section-wise term extraction is similar to that discussed in Section 2.3.1. However, the content of each section (“Abstract,” “Introduction,” etc.) of a paper was separately marked for Apache Lucene. The proposed approach separately required the important terms extracted by Apache Lucene from each section of each research paper. The extracted section-wise terms were then indexed in a database for future usage, as illustrated on the left-hand side of Fig. 1.

3.4.3 Ranking Papers Based on Section-Wise Content Comparisons

The cosine similarity was applied upon the indexed terms to compute the similarity score between the source and target papers, where each paper was represented as a six-term vector. The corresponding term vectors of each paper were compared, i.e., six similarity scores were obtained for each paper based on the matching of the corresponding term vectors (“Abstract” with “Abstract,” “Introduction” with “Introduction,” “Related work” with “Related work,” “Methodology” with “Methodology,” “Results” with “Results,” and “Conclusion” with “Conclusion”) of every other paper. The final similarity score of each paper was evaluated by averaging all the scores obtained in each ACM topic with respect to all other papers. In this way, a ranked list of relevant papers was acquired based on the similarity score of each paper arranged in descending order. The similarity score was computed based on the cosine similarity between the papers, as shown on the left-hand side of Fig. 1.

4 Evaluation Setup

Based on the similarity scores computed above, two separate ranked lists were formed: one for content-based similarity and another for section-wise content similarity of the papers. These lists are required to be evaluated based on certain benchmarks; however, to the best of our knowledge, there is no standard benchmark that can be employed to evaluate the results of the proposed study. Therefore, we constructed a benchmark to compare and evaluate both approaches, as shown in the lower part of Fig. 1.

The development of a gold standard dataset was a crucial task, because there was no available benchmark that could be used for evaluating the proposed approach in the domain of relevant paper recommendations. Normally, authors logically define such standards or prefer user studies. In this study, the documents belonging to the same ACM topic were defined as relevant documents. Besides, authors manually select suitable ACM topics to represent their research during publishing in J. UCS. Therefore, it could be presumed that the authors chose the best topic(s) representing their papers. Thus, we decided to consider the topic information of every research paper available in J. UCS as the gold standard (benchmark), and both the approaches (content and section-wise content) were evaluated against this benchmark. The evaluation process examined the number of top recommendations extracted using both the approaches from a list of 200 documents belonging to the topic(s) of the query paper. The complete list of topics developed by ACM—hierarchically classified as topics A to K is illustrated in Tab. 1.

Table 1: Complete list of topics in the 1998 ACM computing classification system

I	II
Topic-A: General literature	Topic-G: Mathematics of computing
Topic-B: Hardware	Topic-H: Information systems
Topic-C: Computer systems organization	Topic-I: Computing methodologies
Topic-D: Software	Topic-J: Computer applications
Topic-E: Data	Topic-K: Computing milieux
Topic-F: Theory of computation	

Comparing the top recommendations of 200 documents with the benchmark would require exhaustive manual effort. Therefore, five papers from each of the topics ("A" to "K") were selected for evaluation. Moreover, both the approaches were employed to produce results for each of the 200 documents, to comprehensively judge the working of both the approaches. The topics of each selected paper were compared with the topics of top recommendations (top 10, top 15, and top 20) that resulted from using both the techniques. In particular, the topics of paper A (source paper) were compared with the topics of the recommended papers and ranked in a list to find the total number of matched topics. The number of recommended papers, having the same topics as the source paper, was noted, and the detailed results are discussed below.

5 Topic-Wise Results of Content and Section-Based Approaches

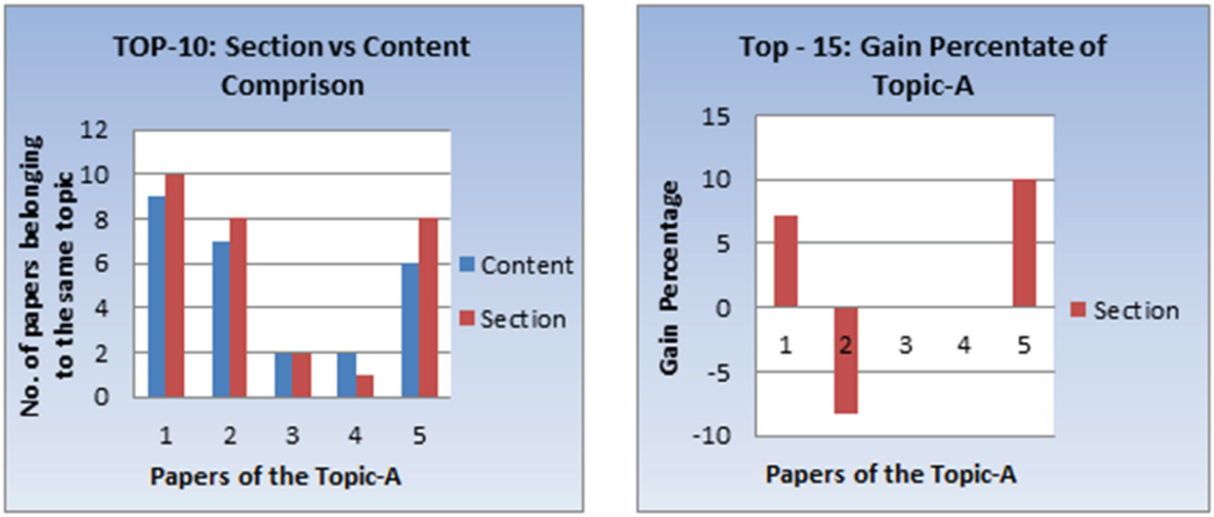
The topic-wise results of the content and section-based approaches on Topic-A are presented in this section; the rest of the topic results have not been discussed owing to their identical nature.

5.1 Topic-A Results

Topic-A in the ACM topic classification represents papers published under the "General Literature" category. Among the 200 papers present in the employed dataset, 14 papers belonged to Topic-A, whereas 186 papers were related to various other topics, and therefore, considered as noise.

In Fig. 2, the header row, i.e., “Topic,” “SR#,” and “PAPER ID” represent the topic, serial number, and IDs of the source papers, respectively. The sub-columns (top 10, top 15, and top 20) under the content-based column contain the matched frequencies reflecting the number of times a topic of the top recommendations were identical to that of the source paper. Similarly, the “Section-based” column contains the same sub-columns representing the number of topics matched in the ranked papers for the section-based approach. For instance, in the top 10 recommendation for the source paper ID = 1, 9/10 papers belonging to Topic-A were extracted by the content-based approach, whereas 10/10 papers belonging to the same topic were extracted by the section-based approach. For better understanding through representation in Fig. 2, distinctive colors—purple (Win), sky blue (Loss), and brown (Equal), respectively, indicate the comparative Win, Loss, and Equal status of the approach based on the number of relevant papers recommended in the top 10, top 15, and top 20 results. In the top 10 recommendations, the comparative scores of the content-based approach was Win = 1, Loss = 3, and Equal = 1, whereas the scores for the section-based approach was Win = 3, Loss = 1, and Equal = 1. Similarly, the section-based approach performed better in providing the top 15 recommendations with Win = 2, Loss = 1, and Equal = 2, in comparison to the content-based approach that had a score of Win = 1, Loss = 2, and Equal = 2. The scores were more indicative in the top 20 recommendations, with the section-based approach clearing Win = 3, Loss = 0, and Equal = 2 against the content-based approach at Win = 0, Loss = 3, Equal = 2.

The comparative results for the top-10 recommendations under Topic-A are portrayed in Fig. 3a, where the X-axis represents the source papers (1–5) for Topic-A and the Y-axis charts the number of relevant papers. Note that the top-10 results were acquired from a list of 200 papers related to Topic-A, where 14 papers belonged to the same topic and the remaining 186 belonged to other topics. This indicates that the similarity scores of the papers represented on the X-axis were computed with respect to each of the 200 papers, and the top 10 recommendations include papers that have the highest similarity scores among the entire list.



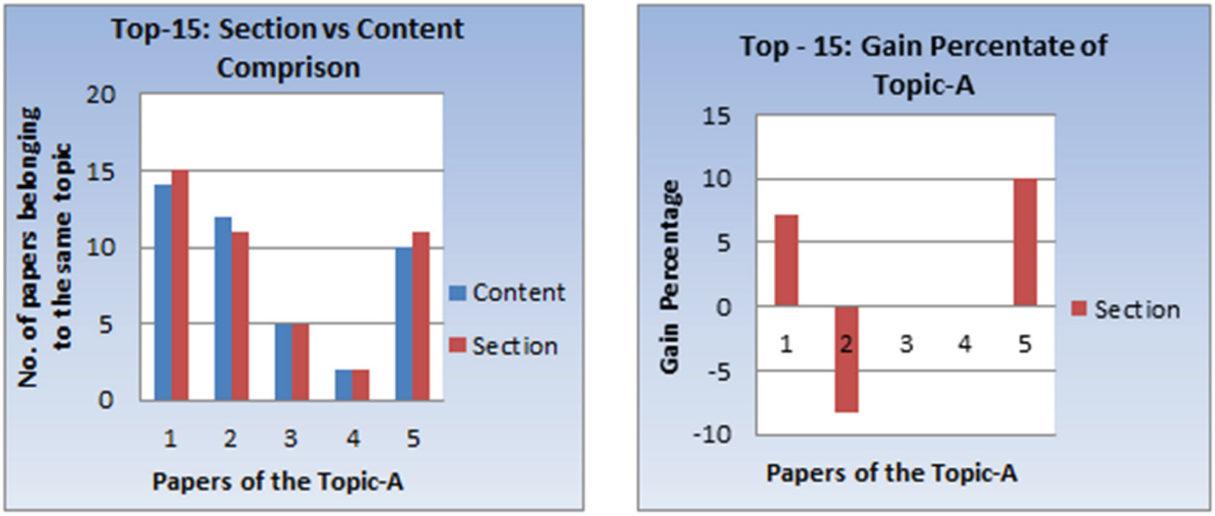
(a) (b)

Figure 3: Comparison of top-20 results for section- and content-based techniques with gain in the section-based results. (a) Top-10 comparison for Topic-A. (b) Top-10 gain percentages for Topic-A

Fig. 3b depicts the gain/loss percentage of the section-based approach over the content-based approach, where the former delivered better results for source paper IDs = 1, 2, and 5, and performed equally at position 3. However, the content-based approach outperformed the section-based approach at position 4.

The section-based approach achieved gain percentages of 11%, 14%, and 33% for the paper IDs = 1, 2, and 5, respectively; however, the loss percentage was 100% for paper ID = 4. Overall, the gain percentage of the proposed approach remained at 12% for the 5 papers in the top-10 recommendations for Topic-A.

In Fig. 4a, the comparisons of the two approaches for providing top-15 recommendations of Topic-A are visualized. At positions 1 and 5, the section-based approach produced better results than the content-based approach, whereas both the approaches performed equally at positions 3 and 4. The content-based approach performed better than the section-based approach at position 2.



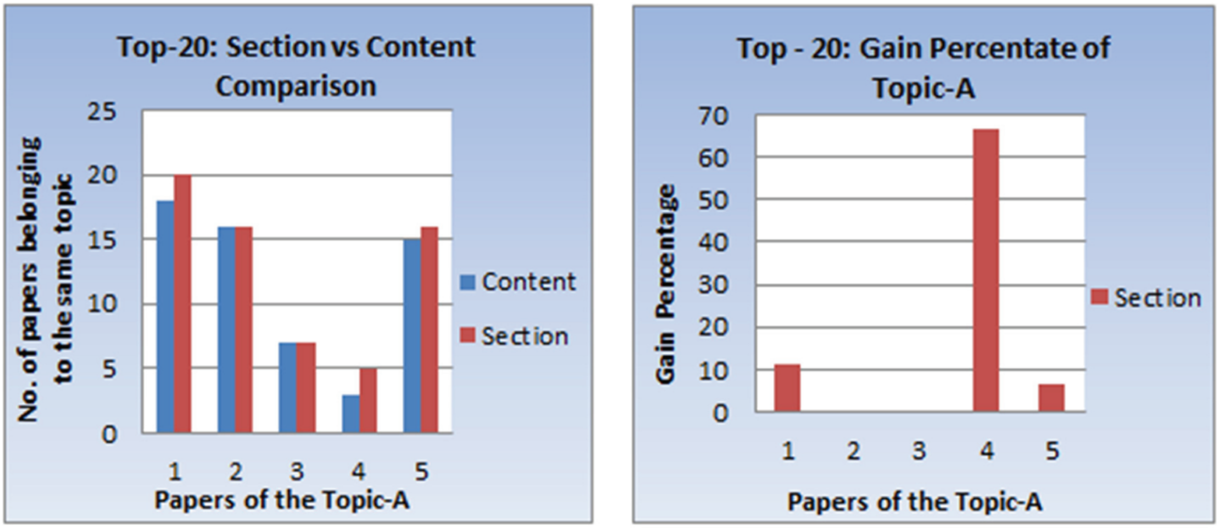
(a) **(b)**

Figure 4: Comparison of top-15 results for section- and content-based techniques with gain in the section-based results. (a) Top-15 comparison for Topic-A. (b) Top-15 gain percentages for Topic-A

Fig. 4b portrays the gain/loss percentages of the section-based approach over the content-based approach. The section-based approach gained 7% and 10% for paper IDs = 1 and 5, respectively, whereas it exhibited an 8% loss for paper ID = 2. However, the gain/loss percentages between the approaches remained at 0% for paper IDs = 3 and 4. Overall, the proposed approach delivered a gain percentage of 2% for all the 5 papers in the top-15 recommendations for Topic-A.

The top-20 recommendations of Topic-A extracted by the two approaches are compared in Fig. 5a, which shows the section-based approach performed better than the content-based approach at positions 1, 4, and 5 and produced an equal number of papers at positions 2 and 3.

Subsequently, the gain/loss percentages of the section-based approach are visualized in Fig. 5b. The section-based approach gained 11%, 67%, and 7% for paper IDs = 1, 4, and 5, respectively. However, both the approaches produced equal results (0% gain/loss) for paper IDs = 2 and 3. Overall, the gain percentage of the proposed approach was 9% for all the 5 papers within the top-20 recommendations for Topic-A. In summary, the average gain of the proposed approach was 7% for all the recommendations—top 10, top 15, and top 20.



(a)

(b)

Figure 5: Comparison of top-20 results for section- and content-based techniques with gain in the section-based results. (a) Top-15 comparison for Topic-A. (b) Top-15 gain percentages for Topic-A

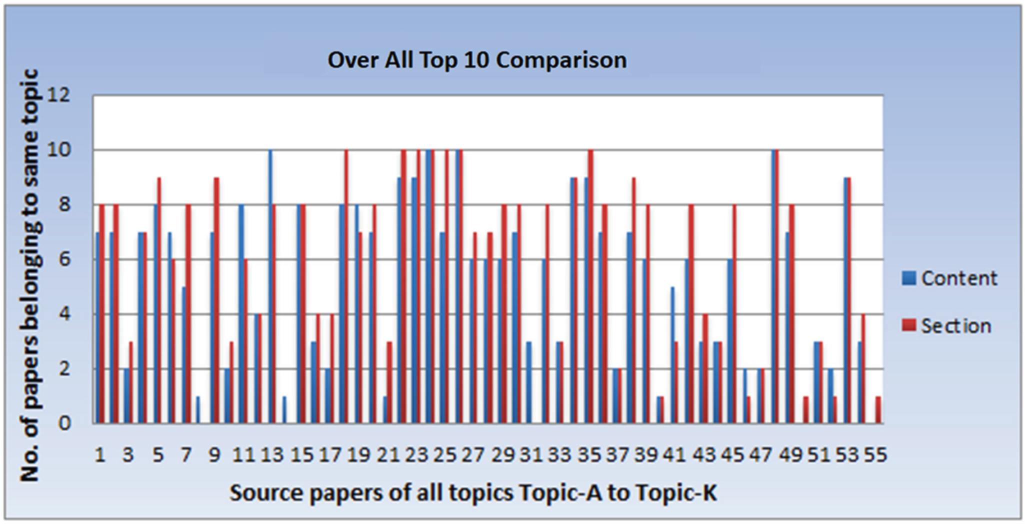


Figure 6: Overall comparison of top-10 results for content- and section-based approaches

5.2 Comparative Results

This section presents the comparative results for all the 55 papers selected for the study. Figs. 6–8 show the results of the top-10, top-15, and top-20 recommendations for each paper, respectively. Although the content-based approach performed better than the section-based approach on 10 occasions for providing the top-10 recommendations, the section-based approach outperformed the content-based approach on 31 instances. Similarly, for the top-15 recommendations, the content-based approach performed better on 9 occasions and produced equal results 13 times; however, the section-based approach outperformed the content-based approach 33 out of 55 times. Furthermore, the proposed section-based approach yielded more number of top-20 matching results on 34 instances and produced equal results 11 times—in comparison to the content-based approach that outperformed the section-based approach only on 10 out of 55 instances.

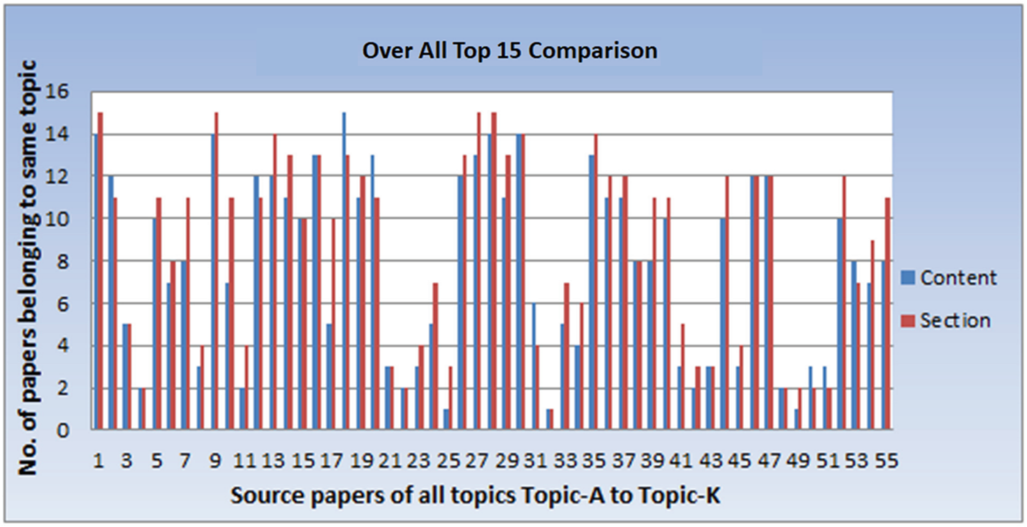


Figure 7: Overall comparison of top-15 results for content- and section-based approaches

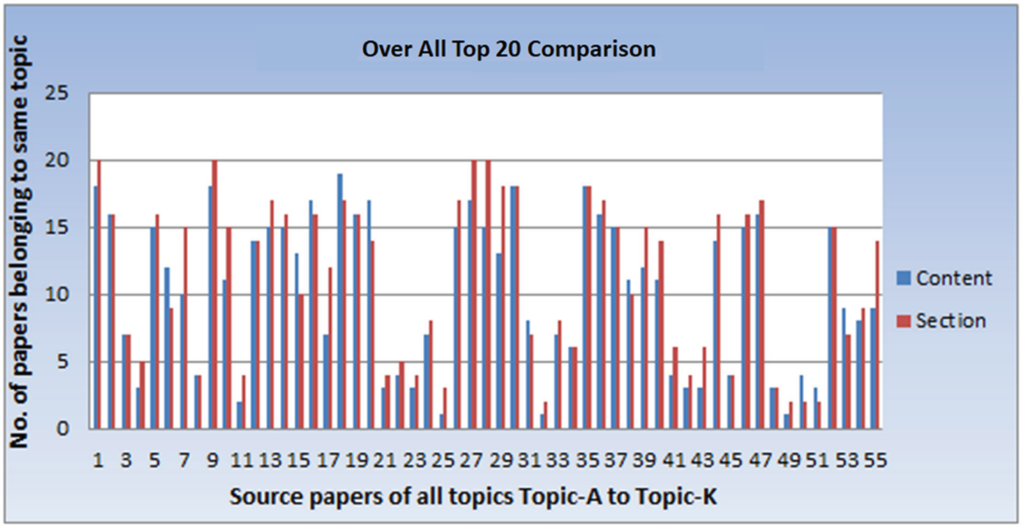


Figure 8: Overall comparison of top-20 results for content- and section-based approaches

6 Discussion

In this study, we performed multiple experiments to evaluate the effectiveness of using terms from various logical sections rather than extracting terms from a paper. Thus, a detailed comparison between the section- and content-based approaches was presented. From the ACM classification hierarchy, five papers were considered from each root-level topic (Topic A–K). In addition, there was enough noise (irrelevant papers) under each topic to validate the working efficiency of the sections- and content-based approaches. Moreover, topic-wise comparisons and statistics of every paper from the selected list of papers were highlighted in detail. Furthermore, the gain/loss percentages of the section-based approach for the top-10, top-15, and top-20 recommendations were evaluated.

The overall topic-wise gain percentages are shown in Fig. 9. The findings of the comprehensive analysis are as follows:

(1) For all topics, except Topic-D in the top-20 recommendations, the section-based approach outperformed the content-based approach.
(2) The highest accuracy of the proposed approach was observed for Topic-E and Topic-I, whereas that for Topic-D remained on low.
(3) The gain percentages for the section-based approach remained consistently higher than the content-based approach for the top-10 and top-15 results; however, the corresponding gain percentages in the top-20 list were not as high as those in the top-10 and top-15 lists, except for Topic-F. Nonetheless, the top-20 recommendations remained in close competition with the top-10 and top-15 results for Topic-A, Topic-E, Topic-F, and Topic-I.
(4) The overall gain percentages of the section-based approach were 15% for top-10, 14% for top-15, and 12% for the top-20 recommendations. The average gain percentage across all the topics was 13.72%.

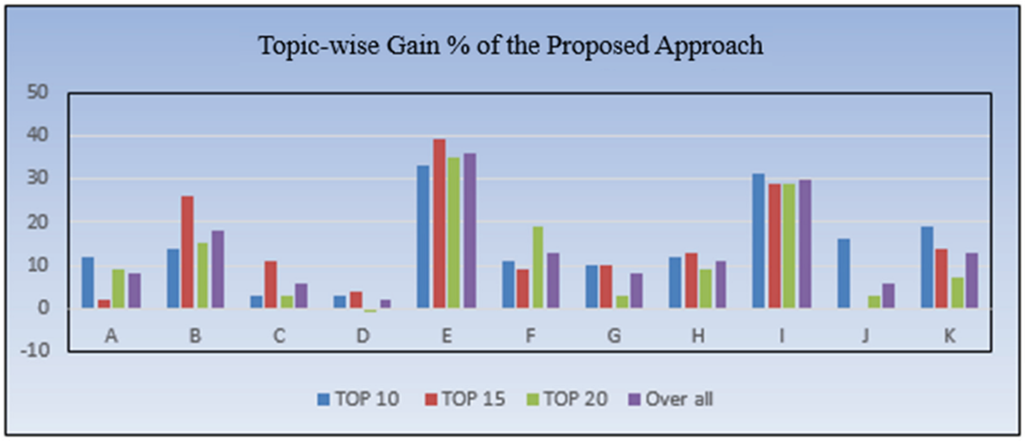


Figure 9: Overall gain percentages of section-based approach

Tab. 2 lists the results of the comparative analysis based on gain percentage of the section-based approach over the content-based approach for obtaining the top-10, top-15, and top-20 recommendations of topics A–K under ACM classification. As can be observed, the proposed approach yielded superior results in most of the cases, except for the top-20 results of Topic-D and the equal top-15 results of Topic-J. In summary, the overall gain percentage of the section-based approach was 13.72%. Thus, the section-wise comparison of terms identified a greater number of relevant papers than the existing methods.

Table 2: Gain percentages of the section-based approach for top-10, top-15, and top-20 results

Topic	Gain percentage of section-based approach (%)			
	Top 10	Top 15	Top 20	Overall
Topic-A: General literature	12	2	9	8
Topic-B: Hardware	14	26	15	18
Topic-C: Computer systems organization	3	11	3	6
Topic-D: Software	3	4	−1	2
Topic-E: Data	33	39	35	36
Topic-F: Theory of computation	11	9	19	13
Topic-G: Mathematics of computing	10	10	3	8
Topic-H: Information systems	14	26	15	18
Topic-I: Computing methodologies	31	29	29	30
Topic-J: Computer applications	16	0	3	6
Topic-K: Computing milieux	19	14	7	13

7 Conclusion

The body of scientific knowledge is rapidly expanding with more than 2 million research papers being published annually. These papers are accessed through various search systems such as general search engines, citation indices, and digital libraries; however, identifying pertinent research papers from these huge repositories is a challenge. On a general query, thousands of papers are returned from these systems, thus making it difficult for end users to find relevant documents. This phenomenon has attracted the attention of researchers to devise state-of-the-art approaches that could assist the scientific community in identifying relevant papers rapidly. These techniques can be categorized into four major categories: (1) content-based, (2) citation-based, (3) metadata-based, and (4) collaborative filtering approaches. Although the content-based approaches have a better recall than other methods, there are limitations that result in a long list of recommendations against user queries. Therefore, researchers have attempted to improve the quality of results by devising more intelligent techniques [30]. In this study, we presented one such idea that may lead the foundation of innovative research-based entrepreneurship. The section-wise content similarity approach intelligently processes the data components of research articles to produce enhanced results. The stated method compares the terms occurring in the corresponding logical sections of each research paper present in the repository. For example, the terms, occurring in the "Result" section of a research article, are compared with the terms occurring in the "Result" section of other research articles. Thus, identifying relevant state-of-the-art approaches is important for finding relevant articles in a short time. The findings of this research are as follows:

(1) The proposed approach outperformed the traditional content-based approach in identifying relevant documents in all topics of computer science classified under ACM hierarchy.

(2) The gain percentage varied from 36% for Topic-E (Data) to 2% for Topic-D (Software). The overall gain percentage of the proposed approach was evaluated at 14% for all topics.

The contribution of our study is related to the vision of applying Scientific Big Data techniques to facilitate the delivery of sophisticated next-generation library services. The retrieval of copious relevant knowledge with the adoption of sophisticated techniques is not just a necessity, but also a decisive action of the global scientific community toward the management of collective wisdom, prosperity, development, and sustainability. The proposed approach poses great potential for further exploration in future. Although certain previous studies [31,32] have highlighted the importance of article sections, such as the "Results" and "Methodology" being more important than the "Introduction" and "Related work," we did not assign any weight on the research article sections based on their importance. Thus, a weighted model may be built in future based on the importance of various sections. Furthermore, a mechanism based on artificial intelligence may be employed for identifying logical sections to extract relevant content.

Funding Statement: This work was partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2020-0-01592) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2019R1F1A1058548).

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

References

[1] M. T. Afzal, N. Kulathuramaiyer, H. Maurer and W. T. Balke, "Creating links into the future," *Journal of Universal Computer Science*, vol. 13, no. 9, pp. 1234–1245, 2007.

[2] A. E. Jinha, "Article 50 million: An estimate of the number of scholarly articles in existence," *Learned Publishing*, vol. 23, no. 3, pp. 258–263, 2010.

[3] K. Funk, R. Stanger, J. Eannarino, L. Topper and K. Majewski, "PubMed journal selection and the changing landscape of scholarly communication," National Library of Medicine, 2017. [Online]. Available https://www.nlm.nih.gov/bsd/disted/video/selection.html.

[4] Scopus Data, 2020. [Online]. Available https://www.elsevier.com/__data/assets/pdf_file/0017/114533/Scopus_GlobalResearch_Factsheet2019_FINAL_WEB.pdf.

[5] M. Gusenbauer, "Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases," *Scientometrics*, vol. 118, no. 1, pp. 177–214, 2019.

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

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

[8] M. T. Afzal and M. Abulaish, "Ontological representation for links into the future," in *Proc. ICCIT*, Gyeongju, South Korea, pp. 1832–1837, 2007.

[9] M. T. Afzal, H. Maurer, W. T. Balke and N. Kulathuramaiyer, "Rule based autonomous citation mining with tierl," *Journal of Digital Information Management*, vol. 8, no. 3, pp. 196–204, 2010.

[10] R. A. Day, "The origins of the scientific paper: The IMRAD format," *Journal of American Medical Writers Association*, vol. 4, no. 2, pp. 16–18, 1989.

[11] J. Wu, "Improving the writing of research papers: IMRAD and beyond," *Landscape Ecology*, vol. 26, no. 1, pp. 1345–1349, 2011.

[12] D. Shotton, K. Portwin, G. Klyne and A. Miles, "Adventures in semantic publishing: Exemplar semantic enhancements of a research article," *PLoS Computing Biology*, vol. 5, no. 4, e1000361, 2009.

[13] S. Peroni, D. Shotton and F. Vitali, "Faceted documents: Describing document characteristics using semantic lenses," in *Proc. DocEng*, Paris, France, pp. 191–194, 2012.

[14] A. Shahid and M. T. Afzal, "Section-wise indexing and retrieval of research articles," *Cluster Computing*, vol. 21, no. 1, pp. 1–12, 2017.

[15] Y. Koren and R. Bell, "Advances in collaborative filtering," in *Recommender Systems Handbook Springer*, 1st ed. vol. 1. Boston, MA, USA: Springer, pp. 145–186, 2011.

[16] Y. Cai, H. Leung, Q. Li, H. Min, J. Tang *et al.,* "Typicality-based collaborative filtering recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 766–779, 2014.

[17] L. D. Murphy, "Digital document metadata in organizations: Roles, analytical approaches, and future research directions," in *Proc. ICSS*, Kohala Coast, HI, USA, pp. 267–276, 1998.

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

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

[20] L. Pasquale, G. D. Marco and S. Giovanni, "Content-based recommender systems: State of the art and trends, " in *Recommender Systems Handbook Springer*, 1st ed. vol. 1. Boston, MA, USA: Springer, pp. 73–105, 2011.

[21] A. Hanan, R. Iftikhar, S. Ahmad, M. Asif and M. T. Afzal, "Important citation identification using sentiment analysis of in-text citations," *Telematics and Informatics*, 2020, 101492.

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

[23] A. Constantin, S. Pettifer and A. Voronkov, "PDFX: Fully automated pdf-to-xml conversion of scientific literature," in *Proc. DocEng*, Florence, Italy, pp. 177–180, 2013.

[24] M. Borg, P. Runeson, J. Johansson and M. V. Mantyla, "A replicated study on duplicate detection: Using apache lucene to search among android defects," in *Proc. ESEM*, Torino, Italy, pp. 1–4, 2014.

[25] Y. Zhou, X. Wu and R. Wang, "A semantic similarity retrieval model based on Lucene," in *Proc. ICSESS*, Beijing, China, pp. 854–858, 2014.

[26] S. Pascal and M. W. Guy, "Beyond TF-IDF weighting for text categorization in the vector space model," in *Proc. IJCAI*, Edinburgh, Scotland, pp. 1130–1135, 2005.

[27] L. H. Patil and M. Atique, "A novel approach for feature selection method TF-IDF in document clustering," in *Proc. IACC*, Ghaziabad, India, pp. 858–862, 2013.

[28] X. Bai, M. Wang, I. Lee, Z. Yang, X. Kong *et al.,* "Scientific paper recommendation: A survey," *IEEE Access*, vol. 7, no. 1, pp. 9324–9339, 2019.

[29] M. Qamar, M. A. Qadir and M. T. Afzal, "Application of cores to compute research papers similarity," *IEEE Access*, vol. 5, no. 1, pp. 26124–26134, 2017.

[30] M. D. Lytras, V. Raghavan and E. Damiani, "Big data and data analytics research: From metaphors to value space for collective wisdom in human decision making and smart machines," *International Journal on Semantic Web and Information Systems*, vol. 13, no. 1, pp. 1–10, 2017.

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

[32] K. Sugiyama and M. Y. Kan, "Exploiting potential citation papers in scholarly paper recommendation," in *Proc. JCDL*, Indianapolis, USA, pp. 153–162, 2013.