Journal of Scientometric Research, 2023; 12(2):243-253.
https://www.jscires.org

Research Article

Discovering Search Space Using M-distance Clustering of Semantic Relatedness Based Weighted Network for the Content-based Recommender System

Mayur Makawana*, Rupa G. Mehta

Department of Computer Science and Engineering, Sardar Vallabhbhai National Institute of Technology, Surat, Gujarat, INDIA.

ABSTRACT

As part of the research process, relevant documents are identified to keep up with the latest advancements in the domain. Document recommendation systems are used by researchers as a means of accomplishing this goal. Textual content, collaborative filtering, and citation information-based approaches are among the proposed approaches for the recommendation systems. Content-based techniques take advantage of the entire text of papers and produce more promising results, but comparing input document text data to every document in the dataset is not practical for the content-based recommender system. This study looks into the possibility of using bibliographic data to reduce the number of comparisons. The proposed system is based on the assumption that two scientific papers are semantically connected if they are co-cited more frequently than by chance. The likelihood of co-citation, also known as semantic relatedness, can be used to quantify this connection. This work presents a new way to distribute the weight among connected scholarly documents based on a semantic relatedness score. Our proposed solution eliminates a substantial amount of needless text comparisons for the system by gathering scholarly document pairs with high likelihood values and using them as a search area for the content-based recommender system. By spreading the co-citation relationship out to certain distances, the proposed approach can find relevant documents that are not found by traditional co-citation searches. The results reveal that the system is capable of reducing computations by a significant margin and of detecting false positive situations in content comparison using Doc2vec.

Keywords: Semantic Relatedness, Content-based Recommended System, Graph Clustering, Co-citation network.

Correspondence:
Mayur Makwana
Department of Computer Science and Engineering, Sardar Vallabhbhai National Institute of Technology, Surat-395007, Gujarat, INDIA.
Email: ds18co003@coed.svnit.ac.in
ORCID: 0009-0002-1763-4918

Received: 06-06-2022;
Revised: 06-04-2023;
Accepted: 11-07-2023.

INTRODUCTION

Information in the world of science is continually growing. As a result of this rapid growth in science, digital publications like books, journals, and conference proceedings have exploded. It is important for researchers to stay updated on the latest research in their sector. Researchers can use a wide variety of databases to aid in their efforts to discover new information. Google Scholar, Science Direct, and IEEE Digital Library are among the most popular engineering databases. The growing number of digital research libraries has provided ample reference material. The overabundance of the information has produced a big challenge for the researchers to find the most related research paper for his/her ongoing research problem.

Most researchers also take the traditional strategy of following the list of references from the publications they already have.[1] Although this strategy may be effective in some cases, it does not guarantee complete coverage of recommended research papers and does not allow for the tracking of publications released after the current paper. The use of research paper recommender systems,[2,3] which automatically offer suitable papers to researchers based on some initial information provided by users, is an alternate strategy that has been proposed in the literature. Many researchers have devised paper recommender systems.[2] Content-based filtering, collaborative filtering,[1] co-citations, and bibliographic coupling are among the techniques used in recommendation systems. Content-based techniques look for similarities across articles by analyzing the contents of the documents. The Bag-of-Words (BOW) model is the most commonly utilized technique for this purpose. Because this model is based on the word-matching principle, it ignores real language ambiguities such as synonym, polysemy, and homonymy. Recommender systems take into account the users' contexts as well as possible contextual information from the ingested contents to

DOI: 10.5530/jscires.12.2.024

deliver more accurate and relevant recommendations.[4] Various researchers have also proposed a recommendation system using a list of citations[5] or a list of papers authored by an author.[6] In a recommendation system, the issue is not just to provide very rich recommendations at any time, any location, and in any form to the researchers, but also to provide them in a short amount of time and with the least amount of complexity. Content-based systems either employ exact matching, such as keyword- based approaches, or circumvent the limits of exact matching through the use of embedding- based techniques such as Word2vec and Doc2vec.[7,8] Due to the consideration of contextual information and synonyms, embedding-based approaches outperform exact matches. The use of embedding techniques has a number of disadvantages. Embedding techniques require good training to translate text into vector representations more accurately. The cost of storing all of the vectors and running every input query against the entire database is unsustainable when dealing with a massive volume of information. The major purpose of our suggested method is to reduce the number of text comparisons by removing superfluous documents or grouping relevant documents together before the final text comparison. To accomplish this, our suggested method relies on citation information, an essential feature of scholarly articles. Citation establishes a meaningful link between two research publications, and a network based on this connection has been extensively applied in numerous studies. The techniques of co- citation and bibliographic coupling[9] are commonly employed to determine the relevance of two documents. Co-citation looks at common incoming links and bibliographic coupling looks at common outbound links. For a paper, incoming links are citing papers, whereas outgoing links are cited papers. Because new articles don't get many citations, our proposed solution uses an undirected graph, which is a blend of bibliographic and co-citations.

In the citation network, distance is inversely proportional to relevance. In simple terms, when the distance between two papers is large, the relevance of the two papers is reduced. Our proposed approach aims to solve questions such as how much distance is sufficient to discover relevant papers in a citation network and how it might aid in reducing the number of comparisons for a content-based recommendation system. Our proposed approach also includes a method for allocating weights to citation connections based on semantic relatedness[10] rather than the number of common links. Due to the inclusion of citation information, the proposed approach can also minimize the number of false positive papers in the recommendation. Using a co-citation information system, review papers that receive a lot of citations could be given the greatest priority. In contrast, the suggested system makes use of semantic relatedness rather than basic co-citation to normalize the situation and concentrate on the importance of paper in the network. For final content comparisons and evaluation, this study employs the Doc2vec model, which is trained on our dataset.

The rest of the paper is laid out as follows. Literature and related works are discussed in Section 2. Section 3 explains the suggested Semantic Relatedness-based Weightage Scheme for K-distance Network Clustering. Experimentation with the proposed method is shown in Section 4. Finally, Section 5 summarizes the presented research findings, including the problems and future aims. Literature Survey In the past decades, researchers have proposed several ways to lead research paper recommendations. As indicated by an author,[11] more than 200 different recommendation approaches have been proposed. These approaches can be classified as: (1) metadata-based approaches,[12] (2) citation-based approaches,[10,13-18] (3) content-based approaches,[19-22] (4) Collaborative Filtering (CF) based approaches[1,23-25] (5) user profile-based approaches[26,27] and hybrid approaches.

Citation-based Analysis

Citation analysis is a variety of methods developed in bibliographic studies to visualize and measure information in different subject areas. For example, when an author cites a specific paper, it may highlight ideas or other important resources or similar to the author's interaction with the cited text. When a group of authors cites a similar paper, these co-papers indicate that they may have similar techniques or thoughts. Therefore, exploring how papers are cited can help researchers to understand the essential endeavours in the field. Since free citations are available in a variety of digital libraries, Bibliographic coupling[28] and co-citation[29] became the most popular methods recommendation systems.[11]

Multiple use cases of co-citation analysis resulted in an improvement in accuracy of research paper recommendation i.e., the personalized approaches that uses the co-citation between authors and documents to create the user's profile and generate the appropriate recommendations based on user's approach,[23,24] the co-citation approaches that used the content analysis recommended the papers that were more relevant to the input paper, as compared to those produced by tradition co-citation analysis, may publications also provide full-length research articles freely, many researchers have also exploited the citation context information to improve the accuracy of recommendation.

Hebatallah *et al.*[23] proposed a personalized research recommendation system that recommends articles based on user feedback. Feedback can be explicit or implicit. They also analyzed user actions to improve user profiles. With a similar approach, Chen and Ban[24] used published papers, citation papers, and references to represent the user's interest. In order to learn the researcher's interests correctly from research contents, they used the concept of a penalty factor.

Each citation link may provide a different meaning to a paper. So, the necessity of assigning weights on citation links emerges with time. As a solution, Tanner *et al.*[20,30] proposed a weighted network using citation relation as citation weight on edges. They defined

citation relation based on the number of times the origin paper cited the reference paper and used this citation relation to measure the strength of the papers' relation. Earlier, it was challenging to parse citation location or context (i.e., texts surrounding citations in documents) due to limited machine-readable text data. Recently, machine-readable documents have become widely available, and therefore citation context-based techniques have become popular.

Masaki Eto[13] has attempted to distinguish citation into two categories, strong and weak, based on the location of citation in the document. For example, if a citation of two documents appears in the same paragraph, they are classified as "strong," but if they appear across two paragraphs, they are classified as "weak" citations. Larger weights are allocated to 'strong' co-citation links, and smaller weights are given to 'weak' ones in the citation network. Furthermore, Masaki Eto[14] extended their previous approach with the graph-based algorithms and co-citation network containing citation context. The proposed technique expands the search space of the co-citation technique to find more common documents that are not likely to be found with traditional methods. Precisely, this is a combination of graph techniques and citation context analysis to measure the similarity score between documents. This combination reduces the comparison between irrelevant documents. The limitation of this method is the cost of calculating similarity scores due to the involvement of citation context information.

Bela Gipp and Joeran Beel[2] proposed an approach called Citation Proximity Analysis (CPA). They used the proximity of citations in entire textual content to discover the strength of in-text co-citation between sets of citations. CPA considers a set of citations more fitting to one another when they come within the same sentence than when they come within the same section. In addition, Boyack *et al.*[31] presented techniques that use the distance between citations. However, as opposed to using the sentence structure to find the distance, they used character or byte offset to propose four different schemes for the same objective. In addition, Boyack *et al.*[32] presented techniques that use the distance between citations. However, as opposed to using the sentence structure to find the distance, they used character or byte offset to propose four different schemes (P1, P2, O, and B) for the same objective. B is based on simple calculations. Each co-citation pair has been assigned a weight of 1. Technique O represents weights based on byte positions of citations. If citations are within 375, 1500 and 6000 bytes, they will get the weights of 3, 2, and 1, respectively. In P1 and P2, they divided the text into 20 equal parts and gave the centiles-based weighted scheme to calculate the similarity between the two papers.

With citation proximity analysis, the accuracy of the paper recommendation system has improved, but another study[18,33] highlighted that some sections tend to have more in-text citations over others. According to this study, authors tend to distribute more citations in the introduction and related works section compared to the methodology and result section. Furthermore, some sections' citations tend to have more importance than other sections in papers. For example, citations within the literature review and Introduction usually indicate that the cited papers could be the supporting document. Nevertheless, the papers mentioned in the 'Methodology' and results sections seem most strongly related. Using this reference, Arjumand *et al.*[16] proposed a method that explores in-text citation frequencies in addition to in-text citation of co-cited papers within the various sections of cited-by papers. The papers were ranked by simply combining section weights with the frequencies of co-cited paperwork. Furthermore, they have examined the proposed method with the co-citation and citation proximity analysis methods. Most of the time, the proposed method outperformed state-of-the- art methods.

All these techniques use citation information directly or indirectly to generate the recommendations and by combining it with other techniques it is possible to generate better accuracy. In many cases authors have shown that citation-based techniques do generate better results compared to content-based. Authors began utilizing citations in the year 2000 to discover semantic connections among their works. Oscar *et al.*[10] presented a model to calculate the likelihood of two scientific papers being mentioned more frequently than they would be if they were cited randomly. This relationship is measured by the probability of co- citation, which comes from a null model that calculates what it thinks is pure chance. To be more specific, it employs a null model in which citations are dispersed randomly and independently across the collection documents to assign weight to the co-occurrence of two citations. The algorithm proposes a ranking of articles in response to a particular article by looking for article pairings that reduce the pure chance probability. The key contributions of this study are (1) an algorithm capable of capturing associations other than those based on apparent similarity of content, and (2) a method for selecting relevant co-occurring instances. MacRoberts and Barbara[35] discussed the discussed the issues with using the number of times a paper is cited to judge how good it is. The authors say that just counting citations is not a good way to measure the quality. They argue that it is highly unlikely that citations can be used as quality indicators. So, it is important to include the text of documents when comparing them.

Content-based Analysis

For content-based research, there are several ways for finding semantic similarities in content-based research. After the surge of machine learning techniques, text comparison became more prominent. Authors started using different word embedding techniques to find the semantic relationships between text documents. Techniques like Word2vec, Doc2vec and Bert are gaining much popularity in this domain. Soumyajit Ganguly and Vikram Pudi[20] developed Paper2vec, unique neural network

embedding-based approach for constructing scientific paper representations that include both textual and graph-based data. Individual nodes in an academic citation network can be seen as graphs, with each node containing rich textual information. They proposed Paper2vec, a method that combines data from both modalities and produces a rich representation for scientific research literature. They sought to discover the k closest neighbours for each node and connect those using "artificial text-based edges" in the citation network. They argued that linguistic similarity between the two publications is crucial, and that their goal is to push two papers closer together that have comparable text content but no direct citation edge. Despite the fact that content is useful information, content-based analysis is more difficult than citation-based analysis. It's difficult to perform text comparisons with each article, especially in large datasets.

Many ways have been presented by researchers to address this issue. The recommendation was proposed by Titipat *et al.*[19] combining the Rocchio Algorithm[34] with a large-scale approximate closest neighbour search using ball trees,[39] Another approach Content-based Node2Vec proposed by B. Kazemi *et al.*[21] used scientific text in a lower- dimensional format. The content and citation network of the article are both employed to create a distributed and universal lower-dimensional representation. This lower-dimensional representation is beneficial in a recommendation system when using the doc2vec method to compare documents.

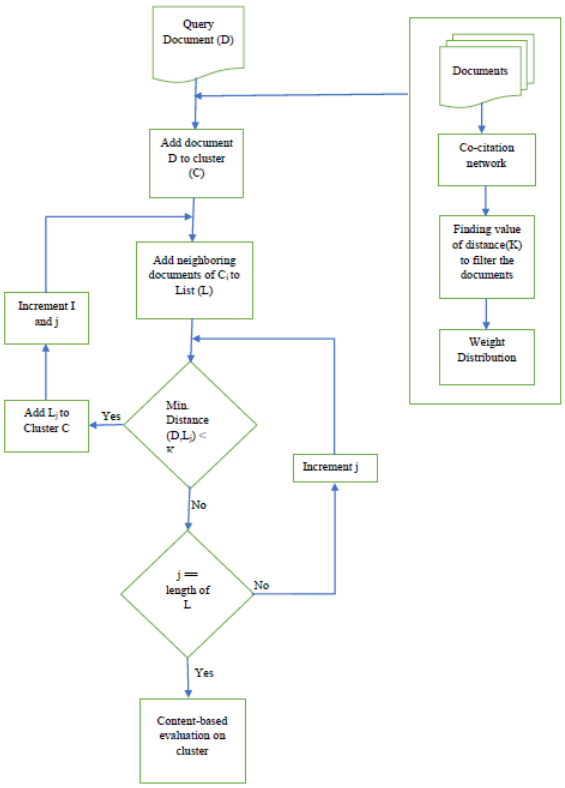


Figure 1: Flowchart of the Proposed Approach.

Proposed Approach

Compiling similarities between the texts is a computationally costly task. While using a content-based recommendation system, whether it is an exact match technique or an embedding-based vector comparison, it is costly to compare input documents to the whole database. To construct an effective search area for content-based comparison, all which is required is a cluster of interconnected, related documents. For clustering techniques like k- mean, Dbscan, and spectral clustering, which need a predefined number of clusters and may not work in broad areas with many subfields, such as Computer Science, for a text-based comparison of a user's input document, we show a new way to get around this problem by using the citation network to create a search region or cluster of similar documents for each user query.

Proposed Semantic Relatedness-based Weightage Scheme for M- distance Network Clustering

Figure 1 shows the flowchart of the proposed approach. The important modules for this system are Network Creation, Finding value of M (Maximum Cluster Distance) and Weight Distribution, Cluster detection and Similarity Score Measuring. The first module, Citation Network, aims to create an undirected citation network, with nodes representing documents and links determining connectivity via citation.

The second step is concerned with determining the search area's maximum distance from the input document to locate related documents. As a means of further enhancement, rather than simply assigning a weight of "1" to each link, Weight distribution describes a method for determining a link's weight based on the semantic relatedness of two connected documents. The final module demonstrates the procedure for determining the difference between our technique and the base approach. These modules are described in depth in the following subsections.

Network Creation

To establish a network of papers, the system treats each document as a node and each citation as a link. Numerous studies (10, 13, 14, 16, 17, 29, 35) describe the DAG (Directed Acyclic Graph) as representing the citation network for a scientific network. In this case, we disregard the direction of the graph and consider the original graph to be an undirected graph, with each edge representing a bi-directional relationship between a specific pair of documents. Despite the fact that the new assumption obscures citation information by omitting the direction, it allows for the enrichment of training data points because each connection in the graph corresponds to a context-dependent relationship. Furthermore, recent papers may not get many citations in a short amount of time because new studies may quote older works, but the reverse is not possible. The relationship between articles can be captured more effectively by switching from a directed to an undirected graph. Furthermore, it can resolve the cold-start issue

when finding the co-citation value. Because even for recently released papers, an undirected graph can capture numerous linked publications.

Finding value of M and Weight Distribution

This process can be divided into three sub-processes.

1) Finding a value of Maximum Depth.

2) Weight Selection.

3) Weight Distribution.

Finding a value of Maximum Depth

Parameter M is the cluster's radius, or the maximum distance from its center, where the center is the user's input document. M is equivalent to the maximum depth in a network from the input document if the weight of all links in the network is one. Because we are applying different weights to the links, we must first determine the maximum depth value to fix the value of parameter M. The value of M will be the same for all users' inputs.

The computational approach for calculating the maximum depth value is shown in Figure 2. An undirected citation network serves as the experiment's input. This procedure is followed for the four arbitrary papers. Using the Co-citation and Doc2vec embedding methods, the system gathers the top-100 documents for each publication. The proposed system uses the Doc2vec model trained on the scholarly literature dataset. The procedure of finding coverage of these Top-100 documents for different depths from the input document is shown in steps 3–7. The system collects the documents at depth one and counts the common documents among the top 100. Next, it checks for the depth of 2 and repeats this process until we locate at least 90% of the top-100 documents. It repeats this process four times for different documents. The average maximum depth is utilized to determine the value of M. Figure 6 shows how many documents have been covered for paper-id 13591 using depths from 1 to 5.

Proposed Weight Selection Strategy

Three separate weights are being used here, and each of these weights acts as a distance between two documents. The smallest feasible distance signifies the greatest chance of finding related documents. So, if the weight is low, the probability of relatedness is high, and vice versa. The main idea is to take the path with the most possibilities and look for the longest run with a high chance of success. To ease the weight distribution procedure, we used 2x the maximum depth for the Maximum distance (M), and we followed specific principles for choosing weights, such as the system can traverse up to the maximum distance M, and the shortest distance is 2. The minimum depth is 2 because all documents up to 2 are either directly or indirectly connected to the input document. If there are three weights, W1, W2, and W3, in increasing order, W1 equals M/depth, W3 equals M/2, and W2 equals the middle element of W1 and W3. The varying weights for the various depth values are shown in Table 1. The procedure from Figure 2 resulted in 5 as the maximum depth for the dataset under experiment. We used the Maximum distance (M) of 10, W1 as 2, W2 as 3, and W2 as 4 for experiments.

Proposed Weight Distribution Procedure

The distribution of weights is determined by the semantic relationship between the linked documents. The co-citation value can be used for weight distribution instead of semantic relatedness, however semantic relatedness has an advantage over basic co-citation value.[10]

Let us consider real cases corresponding to papers from the dataset used for evaluation.

Case 1: Let be paper A the one with identifier P-01 in the corpus, and let be B the one with identifier P02. In this case nA (papers citing only A) is 5 and nB (papers citing only B) is 10, being r, the number of co-citations between A and B, 20.

Case 2: A different case happens for the papers P03 (C) and P04 (B), two of the most cited papers in the collection. In this case nA is 105, nB is 212 and r is 25.

Although the number of co-citations in this second case, 25, is larger than in the first case, 20, it is small with respect to the number of cites of A and B, and thus paper A and B should be considered as strongly connected compared to paper C and D. As a result, proposed system calculates the semantic relatedness likelihood instead of direct co-citation count as weight. It is possible to calculate the degree of similarity between two objects using equation 1. Here, two papers, A and B, are directly linked. R here refers to some connections that papers A and B have in common. nA refers to a reference in paper A that isn't present in R. same as nA, except for paper B, nB is the same.

$$P = \frac{R}{nA+nB} \qquad \text{Eq. 1}$$

The procedure from Figure 3 is used to apply weights to each link in the network after assessing the probability of each link. In the shown procedure, we executed indexes 2 to 6 for every document in our corpus. Let us have a look at this example. Six

```
Algorithm 1: Find a value of Maximum Depth
     Input: Document citation network
     Output: Cluster distance
1    Initialization of variables: Assign zero to variable I, M and Coverage
2    While (I < 5 number of test case)
3        Doc <- Select random document
4        Top 100 <- Select documents using Co-citation and Content-based technique
5        While (Coverage < 90)
6            M <- M+1
7            Find coverage of Top 100 documents for M depth
8    M <- average of M for test cases
```

Figure 2: Algorithm for finding value of Maximum Depth.

Table 1: Weight Selection.

Depth	Maximum Distance (M)	W1	W2	W3
3	6	2	2.5	3
4	8	2	3	4
5	10	2	3/4	5
6	12	2	4	6

Algorithm 3: Weight Distribution

***Input**: Citation network (C), Probability Matrix(PM), W1, W2, W3*
***Output**: weighted citation network*
For each Document(D) in C
    L<- Collect the neighbouring pairs with probability for D
    B1, B2, B3 <- Equal Width Binning on L for Bin-size 3
    Assign W1 to B3 (Edges with maximum probabilities)
    Assign W2 to B2
    Assign W3 to B1 (Edges with minimum probabilities)

Figure 3: Algorithm for Weight Distribution.

nodes surround Paper-P: A, B, C, D, E, and F. As indicated in Table 2, a probability is assigned to each connection between P and its neighbours. For bin size 3, we used equal width binning on the probability values. After collecting the bins, we distributed the edges based on the bin's probability range. Table shows that we put the least amount of weight in the bin with the highest probability and the most amount of weight in the bin with the lowest probability.

Cluster Detection

Cluster Detection is the final phase of our proposed approach. By this phase the system is equipped with a weighted citation network and the network's maximum distance. For the cluster to be created, system accepts the document as an input, which must be part of the dataset. When a user issues an input paper, the document system attempts to locate the document in the network. Once the document has been located, the document system traverses the network and collects all documents within a reachable distance, where reachable distance is defined as the maximum cluster distance.

Figure 4 shows an example of how the network clustering process works in its entirety. In Figure (A), P denotes the user's input for which the recommendation is required. For this example, the maximum depth is 3; using Table , we get a Maximum distance (M) of 6, and weights W1, W2, and W3 are 2, 2.5, and 3, respectively. Figure 4 (B) depicts the network after the weight distribution process has been completed. In the final phase, the system traverses a maximum distance of 6 from the user's input P, seeking a cluster of similar documents. Figure 4 (C) depicts the final collection of documents using red colour.

Similarity Score Measuring

In the vector space paradigm, it is simple to compare the text of two publications and determine how similar they are. This phase

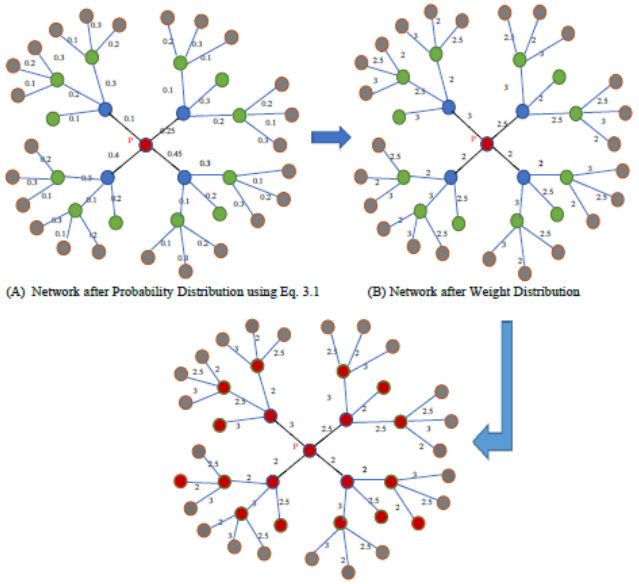


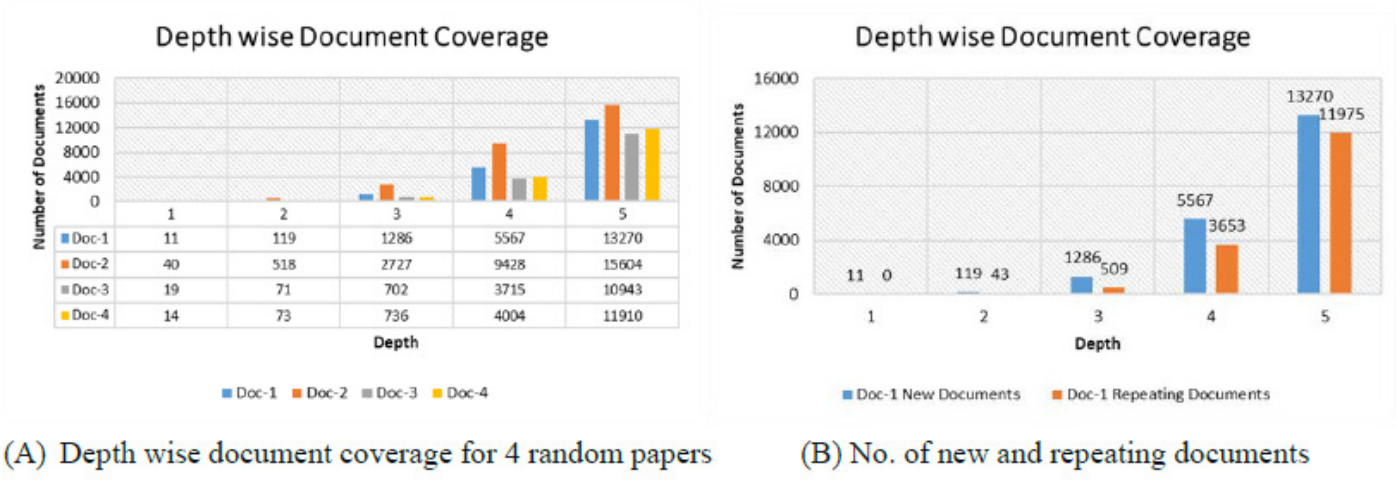
Figure 4: Cluster Detection for Document P with a maximum distance (M) of 6.

aims to transform the text into a low-dimensional, real-valued, and semantically rich embedding (i.e., vector) space. Any quantity of text can be converted to a semantically rich vector form using the well-known Doc2Vec[8] document embedding method. Doc2Vec is an advanced version of Word2Vec. Doc2Vec takes vocabulary words and turns them into vectors using a process known as "word embedding." It is a neural network-based word embedding solution that learns semantically rich word embeddings by analyzing the context of the surrounding words. Standard similarity measures such as the cosine (cos), generalized Jaccard (g.jacc), extended Jaccard (e.jacc), and Dice similarity can be used to compare the vectors d1 and d2 belonging to distinct paper.

Related documents discovered using Doc2vec were employed in two phases of our proposed system, the first to determine the maximum cluster distance and the second to determine the coverage of our formed cluster. In order to assess the effectiveness of the proposed approach, we look at the percentage of the top 100 related documents in the generated cluster and the number of comparisons our system can eliminate.

Table 2: Weight Distribution.

Edges	Probability	Equal Width Binning	Edges	Weight
P->A	0.1	B1 <- (0.1 to 0.2)	P->A, P->E	3
P->B	0.5			
P->C	0.3	B2 <- (0.2 to 0.4)	P->C, P->E	2
P->D	0.6			
P->E	0.2	B3 <- (0.4 to 0.6)	P->B, P->D	1
P->F	0.4			



(A) Depth wise document coverage for 4 random papers (B) No. of new and repeating documents

Figure 5: Depth wise Document Coverage in Network.

Experiment Analysis

This section is about experiments performed on a scholarly literature dataset like data sampling, network building, word embedding and clustering.

Dataset

For the purposes of testing relevant scientific publication recommendation algorithms, data is essential. Hung N. *et al.*[36] have been conducting an extensive study into a possible method for overcoming data limitations in scientific publishing recommendations. They made a vast dataset of academic publications available online, complete with author, title, abstract, and citation information. That dataset is being used here for testing purposes. The dataset contains 702,643 publications published between 1965 and 2009. Each document is managed by eliminating the noisy words and retrieving the abstract, title, and citation information. There are approximately 7,654,677 citations in this dataset. It is challenging to conduct tests with minimal resources due to its enormous size.

A small dataset with strong citation connectedness was obtained by data sampling. Only those cases with strong citation connectedness are selected for selective sampling because the suggested method heavily depends on the citation network. We went through a straightforward procedure. Each citation connection is assigned a value of 1, and we establish a network up to the depth of 8 by selecting one random document. In the end, we had 37,122 publications with 455464 citations after deleting duplicates and papers with fewer than five citations. For all of our experiments, the system used this sample dataset.

Finding value of M and Weight distribution

In the graph, M represents the greatest possible distance in the search region from the input document. We examined the dataset to determine M's value, and the results are shown in Figure 5(A). Four papers were chosen at random and their depth of coverage was plotted on a graph. As the depth of the graph increases, the number of papers in the graph grows at an exponential rate. If we look at Figure 5(B), there are only 11 papers at a depth of '1.' This increases to 131 at a depth of 2, 1367 at a depth of 3, 5860 at a depth 4, and 12136 at a depth 5 due to an increase in depth. In addition, it has been found that the number of redundant documents is also increasing at a similar rate, hence a meaningful depth or the value of M must be chosen. As a possible solution we used co-citation and doc2vec to locate 100 papers that are similar to our input paper, and we examined at what depth we can retrieve utmost similar documents. Doc2vec[37] provides a pre-trained model that can be used for vector generation, but in search of better results, the model has been trained on our sample corpus. In total, the corpus contains 52153 distinct keywords and 37122 documents for training. For paper-id 13591, the top 50 papers from Doc2vec and the top 50 papers from the co-citation

Table 3: **Weights after applying Equal Width Binning.**

Document-ID	Document Relatedness	Weight
5635	0.289	3
5552	0.1	5
5553	0.411	2
5546	0.33	2
14320	0.275	5
14319	0.290	3
5543	0.525	2
8228	0.1	5
5560	0.360	2
5545	0.308	3
19911	0.0	5

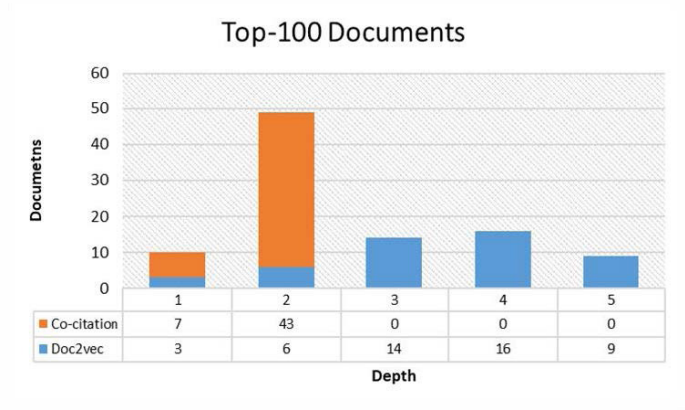
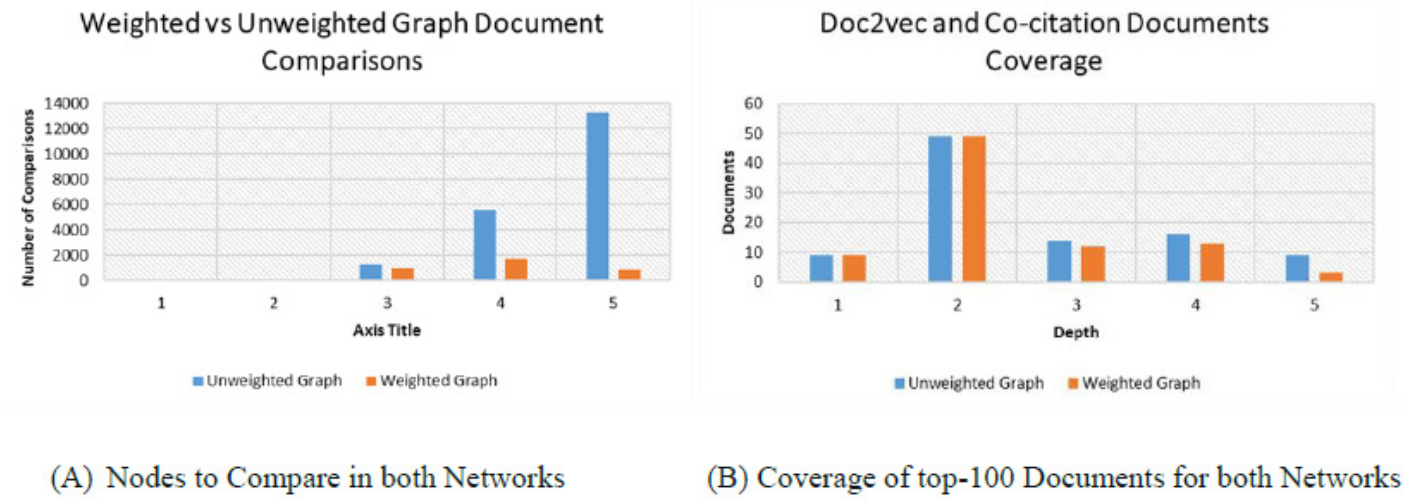


Figure 6: Doc2vec vs Co-citation.



(A) Nodes to Compare in both Networks (B) Coverage of top-100 Documents for both Networks

Figure 7: Unweighted and Weighted Network Comparison.

matrix were chosen. As a subset of co-citation and Doc2vec, the system got 15 papers.

Figure 6 shows the distribution of the top-100 papers irrespective to the depth from 1 to 5. As can be seen in the figure, co-citation only produces output from depths 1 and 2. Due to a connection break after the depth of 2, it is not possible to compute the co-citation similarity beyond the depth of 2. We may deduce from the Doc2vec output that there are a number of papers that are similar to the input paper beyond the depth of two. The network almost completely covers the top 100 documents at a depth of 5. This procedure was carried out with five input documents to choose the maximum depth for the remaining experiments. After analyzing the result, the value of depth was set to 5. Based on Table , parameters M, W1, W2, and W3 values have been set at 10, 2, 3, and 5 correspondingly for a maximum depth of 5.

Weight Distribution

The next step is to apply weights to each link in the network after discovering relatedness using equation 1. Equal width binning was used to construct three bins and assign values based on the edges' placement in the bins. Table 3 displays the values of relatedness and the weights allocated to the nodes that are adjacent to paper-id 13,519. After applying equal width binning to the relatedness value, the system obtained three bins: 0-1.75, 1.75-0.35, and 0.35-0.525 for the paper-id 13,159. In reference to Table , a minimum weight of "2" is selected for bin-3 and a maximum weight of "5" is selected for bin-1.

RESULTS AND DISCUSSION

For the purposes of this evaluation, the same top 100 documents were employed as before. Our goal was to see if the cluster of documents had all of the top-100 documents or if it was missing any. After assigning weights for the maximum distance of 10,

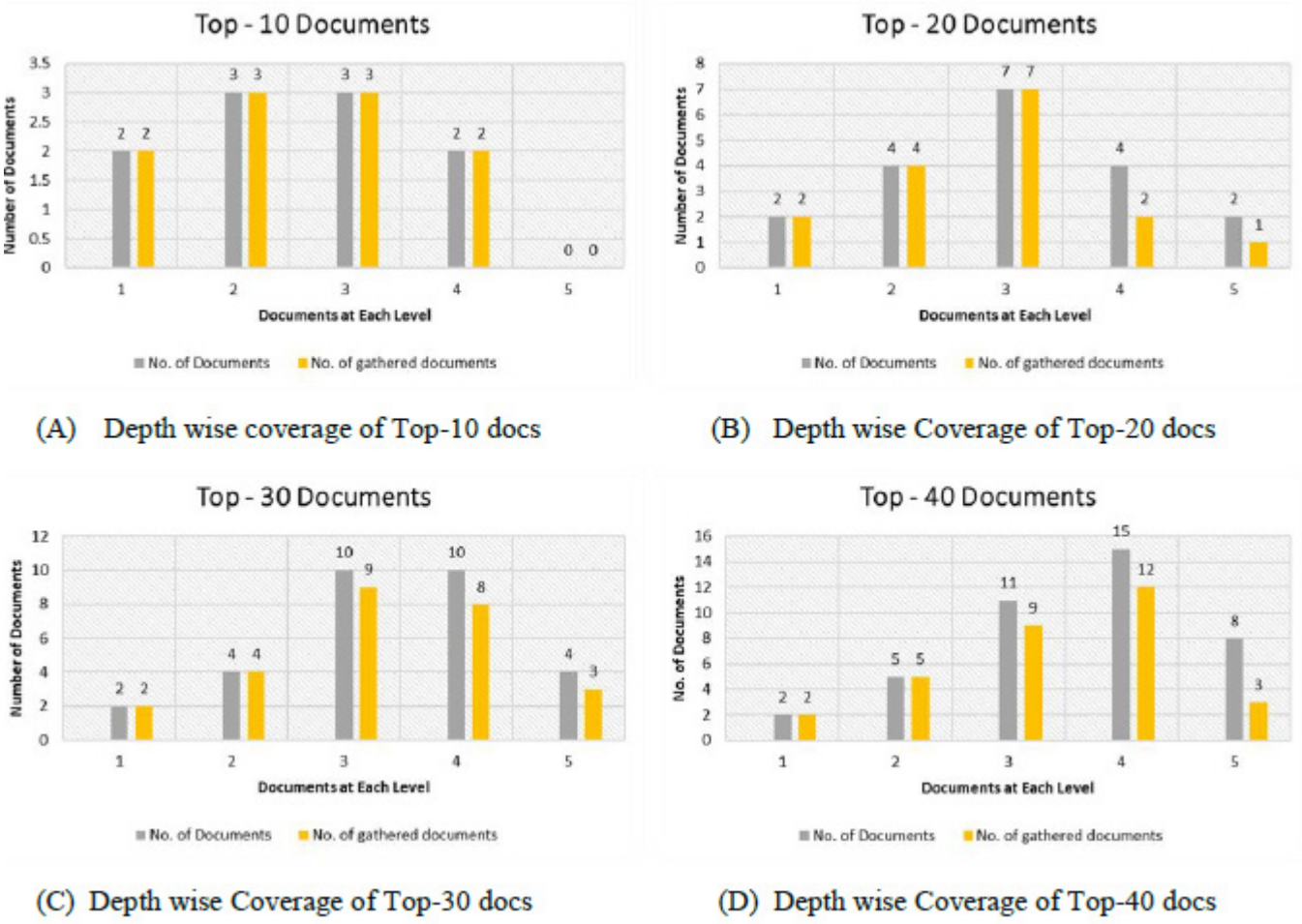


Figure 8: Top Documents with Depth wise Coverage.

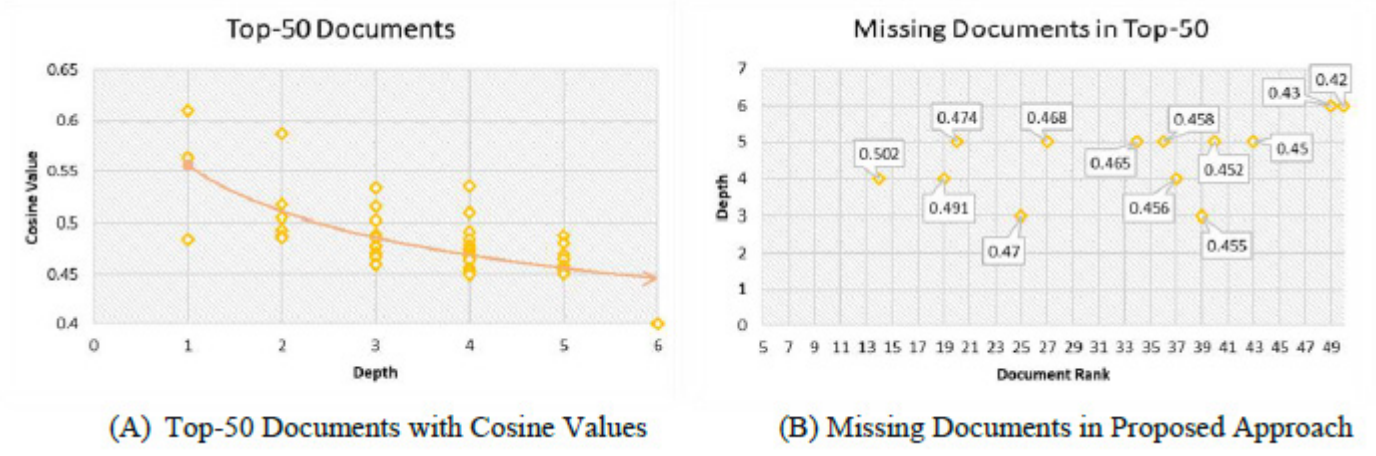


Figure 9: Top-50 Documents with Cosine values and Missing Documents.

we discovered a total of 3076 documents compared to 17,342 documents for the paper-id 13,591. Figure 7 (B) illustrates the top 100 documents' coverage in terms of depth, as seen in the graph. We examined for four input documents and consistently received 75–80% coverage.

The suggested method relies on a number of variables, including the maximum distance in the cluster and the link weights. Various datasets may have different parameters and different outputs. So, we may need to run this method on various datasets to get a clearer picture of the outcomes. A few of the Top-100 Doc2vec papers were missing from the proposed strategy, although manual

observation revealed that several of these missing documents were false-positive situations. It is possible for Doc2vec to offer a high cosine similarity for two completely different documents. We wish to verify this output for other embedding techniques like Bert. There are a few other challenges we may face with streaming data. In the future, we would like to discuss the change that occurs in the network and how it can be adjusted.

To validate the commitments in the earlier section, Figure 8 (A-E) shows how many documents Doc2vec has collected at each level of depth and how many of those documents our method has covered. We calculated the proportion of documents that our algorithm covered for the top- 10, 20, 30, 40, and 50 documents. As we progressed from the top 10 to the top 50, we saw an increase in the number of missing documents.

To locate related documents, the Doc2vec model was trained with a vector size of 190, an epoch value of 45, and a minimum word value of 5. As part of the effort to improve the doc2vec model, superfluous terms have been omitted from the texts, and only the most relevant words have been used. We examined the doc2vec output with various vector sizes, epochs, and minimum word values. The data has been filtered as well in order to obtain a satisfactory Doc2vec output. The present model has been trained to yield a maximum similarity score of 0.69. All of these graphs are for document ID 25087.

To evaluate how well the proposed approach generalizes, we experimented by taking different papers as an input, and the results obtained were similar to Figure 8. The algorithm missed 25- 30% of the documents collected by Doc2vec for the top-50 documents. We gathered the missing papers and examined the cosine values and ranks of the top 50 documents for further analysis, as shown in Figure 9. Figure (A) depicts the representation of missing papers and their corresponding ranks. The cosine value can be found on the document label. Most of the missing documents are ranked 30–50, with a cosine similarity value of less than 0.47. As an additional measure of the input document's similarity to other documents, we counted the number of common terms. Most missing documents were not identical to the input documents, so even with 25–30% of missing documents, our proposed approach provides better results compared to the number of comparisons.

Calculating the accuracy of the recommendation model is difficult because there is no optimal output in the recommendation. The ideal output is significantly reliant on the user's perspective. Forming offline expert committees to assess the results of proposed systems is one of the evaluation techniques in the recommendation.[38] The output of the trained Doc2vec model, on the other hand, is considered an ideal output in this case. Accordingly, we can calculate the recall value for the top-10, 20, 30, and 40 for the document cluster produced by our proposed method. We obtained 1, 0.95, 0.9, and 0.75 for Recall@10, Recall@20, Recall@30, and Recall@40. We needed to perform 11, 119, 1000, 1728, and 836 comparisons for each of the depths of 1 to 5. Against more than 16 thousand, a total of 3,694 comparisons were made by the proposed approach. These results demonstrate that the calculation costs were greatly lower due to the proposed approach. Furthermore, the proposed approach aids Doc2vec in identifying actual positive cases while locating related documents from various perspectives.

CONCLUSION

We identified the cost of comparisons for large amounts of data as a critical obstacle to a content-based document recommendation system. While working with literature papers, bibliographic and co-citation approaches are simple, fast, and produce good results. However, these techniques completely ignore the documents that are not directly or indirectly connected and the importance of textual data. The framework presented in the paper uses semantic relatedness to calculate the weights among links and a unique way to produce the cluster of arguably similar documents, solving the mentioned issues. Our approach may include some irrelevant documents, and it may miss some of the related documents, but as shown in the results and discussion, our approach reduced around 80% of unnecessary comparisons.

In order to reduce the harmful effects of an increasing number of irrelevant documents, it would be possible to incorporate co-citation contexts into the process of calculating similarity scores. The cluster length for dense and sparse networks should be different. The same articles are recommended to all researchers without paying attention to the researcher's previous studies. Here the recommendation is made by accepting each researcher equally. The future aims to create a user-specific article recommendation system by considering the features such as the publications in the researcher's profile and study field.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Liu H, Kong X, Bai X, Wang W, Bekele TM, Xia F. Context-Based Collaborative Filtering for Citation Recommendation. IEEE Access. 2015;3:1695–703.
2. Gipp B, Beel J. Citation Proximity Analysis (CPA)-A new approach for identifying related work based on Co-Citation Analysis [Internet]. Available from: www.scienstein.org
3. Gipp B, Beel J, Hentschel C. Scienstein: A research paper recommender system. In: Proceedings of the international conference on Emerging trends in computing (ICETiC'09) 2009;309-315.
4. Sugiyama K, Kan MY. Scholarly paper recommendation via user's recent research interests. Proc ACM Int Conf Digit Libr. 2010;29-38.
5. McNee SM, Albert I, Cosley D, *et al*. On the recommending of citations for research papers. In: Proceedings of the 2002 ACM conference on Computer supported cooperative work 2002; 116-125.
6. Sugiyama K, Kan MY. Exploiting potential citation papers in scholarly paper recommendation. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. 2013;153-62.
7. Bulut B, Gündoğan E, Kaya B, Alhajj R, Kaya M. User's Research Interests Based Paper Recommendation System: A Deep Learning Approach. 2020;117-30.
8. Le Q, Mikolov T. Distributed representations of sentences and documents. 31st Int International Conference on Machine Learning. 2014;4:2931-9.

9. Cristo M, Calado P, Silva De Moura E, Ziviani N, Ribeiro-Neto B. Link Information as a Similarity Measure in Web Classification. String Processing and Information Retrieval: 10th International Symposium. 2003;43-55.
10. Rodriguez-Prieto O, Araujo L, Martinez-Romo J. Discovering related scientific literature beyond semantic similarity: a new co-citation approach. Scientometrics. 2019;120(1):105-27.
11. Beel J, Gipp B, Langer S, Breitinger C. Paper Recommender Systems: A Literature Survey. International Journal on Digital Libraries. 2016;17:305-38.
12. Ahmad S, Afzal MT. Combining metadata and co-citations for recommending related papers. Turkish Journal of Electrical Engineering and Computer Sciences. 2020;28(3):1519-34.
13. Eto M. Evaluations of context-based co-citation searching. Scientometrics. 2013;94(2):651-73.
14. Eto M. Extended co-citation search: Graph-based document retrieval on a co-citation network containing citation context information. Information Processing & Management. 2019;56(6):102046. DOI:
15. Liu RL, Hsu CK. Improving bibliographic coupling with category-based cocitation. Appl Sci. 2019;9(23).
16. Khan AY, Shahid A, Afzal MT. Extending co-citation using sections of research articles. Turkish Journal of Electrical Engineering and Computer Sciences. 2018;26(6):3345-55.
17. Shahid A, Afzal MT, Qadir MA. Discovering Semantic Relatedness between Scientific Articles through Citation Frequency. Australian Journal of Basic and Applied Sciences. 2011;5(6):1599-604.
18. Habib R, Afzal MT. Sections-based bibliographic coupling for research paper recommendation. Scientometrics. 2019;119(2):643-56.
19. Achakulvisut T, Acuna DE, Ruangrong T, Kording K. Science Concierge: A fast content-based recommendation system for scientific publications. PloS One. 2016;11(7):e0158423.
20. Ganguly S, Pudi V. Paper2vec: Combining Graph and Text Information for Scientific Paper Representation. In: Advances in Information Retrieval: 39th European Conference on IR Research. 2017;383-395.
21. Kazemi B, Abhari A. Content-based Node2Vec for Representation of Papers in the Scientific Literature. Data & Knowledge Engineering. 2020;127:101794. DOI: 10.1016/j.datak.2020.101794
22. Ravi KM, Mori J, Sakata I. Cross-domain academic paper recommendation by semantic linkage approach using text analysis and recurrent neural networks. In: Portland International Conference on Management of Engineering and Technology. 2017;1-10.
23. Hassan HA. Personalized research paper recommendation using deep learning. In: Proceedings of the 25th conference on user modeling, adaptation and personalization. 2017;327-330.
24. Chen J, Ban Z. Literature recommendation by researchers' publication analysis. In: IEEE International Conference on Information and Automation. 2016;1964-69.
25. Haruna K, Ismail MA, Damiasih D, Sutopo J, Herawan T. A collaborative approach for research paper recommender system. PLoS One. 2017;12(10):e0184516.
26. Bu Y, Liu T yi, Huang W bin. MACA: a modified author co-citation analysis method combined with general descriptive metadata of citations. Scientometrics. 2016;108(1):143-66.
27. Wang B, Bu Y, Huang W. Document- and Keyword-based Author Co-citation Analysis. Data Inf Manag. 2018;2(2):70-82.
28. Weinberg BH. Bibliographic coupling: A review, Inf Storage Retr. 1974;10(5- 6):189-96.
29. Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science. 1973;24(4):265-9.
30. Tanner W, Akbas E, Hasan M. Paper Recommendation Based on Citation Relation. In: IEEE international Conference on Big Data. 2019;3053-9.
31. Boyack KW, Small H, Klavans R. Improving the accuracy of co-citation clustering using full text. Journal of the American Society for Information Science and Technology.2013;64(9):1759-67.
32. Boyack KW, van Eck NJ, Colavizza G, Waltman L. Characterizing in-text citations in scientific articles: A large-scale analysis. J Informetr. 2018;12(1):59-73.
33. Jeong YK, Song M, Ding Y. Content-based author co-citation analysis. J Informetr. 2014;8(1):197-211.
34. Ye Z, He B, Huang X, Lin H. Revisiting Rocchio's relevance feedback algorithm for probabilistic models. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2010; 151-61.
35. MacRoberts MH, MacRoberts BR. Problems of citation analysis: A critical review. Journal of the American Society for information Science. 1989;40(5):342-9.
36. Tran HN, Huynh T, Hoang K. A Potential Approach to Overcome Data Limitation in Scientific Publication Recommendation. In: Seventh International Conference on Knowledge and Systems Engineering. 2015;310-313.
37. Kenter T, De Rijke M. Short text similarity with word embeddings. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. 2015;1411-1420.
38. Beel J, Langer S, Genzmehr M, Gipp B, Breitinger C, Nürnberger A. Research paper recommender system evaluation: A quantitative literature survey. In: Proceedings of the international workshop on reproducibility and replication in recommender systems evaluation. 2013;15-22.
39. Dolatshah M, Hadian A, Minaei-Bidgoli B. Ball*-tree: Efficient spatial indexing for constrained nearest-neighbor search in metric spaces. arXiv preprint arXiv:1511.00628. 2015.

Cite this article: Makwana M, Mehta RG. Discovering Search Space Using M-distance Clustering of Semantic Relatedness Based Weighted Network for the Content-based Recommender System. J Scientometric Res. 2023;12(2):243-53.