Journal of Intelligent & Fuzzy Systems 46 (2024) 10135–10150
DOI:10.3233/JIFS-237975
IOS Press

ICRM: An intelligent citation recommendation mechanism based on BERT and weighted BoW models

Chih-Yung Chang[a,*], Yu-Ting Yang[a], Qiaoyun Zhang[a], Yi-Ti Lin[b] and Diptendu Sinha Roy[c]
[a]*Department of Computer Science and Information Engineering, Tamkang University, New Taipei, Taiwan*
[b]*Department of English, Tamkang University, New Taipei, Taiwan*
[c]*Department of Computer Science and Engineering, National Institute of Technology, Shillong, India*

Abstract. With the field of technology has witnessed rapid advancements, attracting an ever-growing community of researchers dedicated to developing theories and techniques. This paper proposes an innovative ICRM (Intelligent Citation Recommendation Mechanism), designed to automate the process of suggesting the appropriate number of citations for individual brackets within a document. The proposed ICRM comprises three phases: Coarse-grained Weighted Bag of Word (WCBW), Fine-grained SciBERT (FSB) and Citation Adjustment phases. Firstly, the WCBW phase employs TF-IDF to extract keywords from both target and candidate documents, forming vectors that capture word significance along with metadata like authorship, keywords, and titles. It aims to identify relevant papers from a database, serving as initial candidates for each bracket. Secondly, the FSB phase employs the SciBERT model to assess the similarity between candidate documents and the local context around brackets, enhancing the precision of recommendations. It refines this selection by analyzing candidate-document relationships within the proximity of the brackets. Lastly, the Citation Adjustment phase tackles overlapping citations and ensures that recommended citation numbers align with user-defined criteria, resolving issues of imbalance. The simulation results demonstrate that the proposed ICRM outperforms existing models significantly in terms of precision, recall and F1-score.

Keywords: Citation recommendation, TF-IDF, weighted bag of word, BERT

1. Introduction

Research in the field of technology has been experiencing rapid advancements in recent years. An increasing number of researchers are involved in developing theories and techniques in this field and sharing their research results with others in no time. Therefore, researchers often struggle to keep up with the pace of new publications, leading to difficulties in summarizing related work. The goal of this study is to develop a system that provides accurate citation recommendations, reducing redundant citations and ensuring the appropriateness of citations in each bracket.

The advanced development of natural language processing (NLP) [1–3] in AI technology makes extracting a few keywords and formulating a summary from a document possible. The technique of extracting keywords from a document has been now widely applied in the areas of news processing, article recommendations, sentiment analysis, and so on. Making a summary from a document can also improve work efficiency and reduce the effort of readers for processing the document. To research related

*Corresponding author. Chih-Yung Chang, Department of Computer Science and Information Engineering, Tamkang University, New Taipei 25137, Taiwan. E-mail: cychang@mail.tku.edu.tw.

ISSN 1064-1246/$35.00 © 2024 – IOS Press. All rights reserved.

work in a specific field, researchers are now able to use the AI technique to find the connections among a large number of documents. Therefore, it is an essential research topic to develop the mechanism of automatic citation recommendations to specific brackets in the research paper. The purpose of this paper is to develop the technique of citation recommendations based on the technology in natural language processing.

In the literature, previous researchers have proposed citation recommendation mechanisms falling into two main categories: word vectors and deep learning. The first category [4, 5] applied TD-IDF [6, 7] to find keywords from the document and then adopted word2vec to embed keywords into word vectors. Based on the vectors of words, the developed techniques facilitated the extraction of connections among various documents. In addition, related studies falling in the second category developed the mechanisms of citation recommendations based on deep learning. These mechanisms adopted either LSTM [8–12] or BERT-based [13–15] models to find the connections among documents and recommend suitable citations to certain brackets. However, most of them still had some room for improvement. First, the representation of each document was generated in hidden states which were inexplicable and not transparent. The connections among documents built based on this inexplicable technique led to inaccurate citation recommendations. Moreover, most of the proposed approaches recommended citations to a certain bracket, leading to overlapping citations in different brackets or an improper number of citation recommendations.

This study proposes the ICRM, aiming at automatically recommending an appropriate number of citations for individual brackets in research papers. The main drawback of current systems is their overly simplistic approach to citation feature extraction, mostly considering only the contextual embedding features, which lack diversity, leading to inaccurate citation recommendations. Furthermore, most existing methods can only recommend citations for specific brackets, resulting in overlaps or inappropriate quantities of citations between different brackets.

This paper aims to propose a citation recommendation mechanism that consists of three phases. In the first phase, the coarse-grained approach is proposed, aiming to make some candidate recommendations. To achieve this, this phase applies the technique of the Weighted Bag of Words to construct a set of features as a basis to generate a vector for each document. These document vectors are transparent and explicable. Therefore, the candidate citation recommendations generated in this phase are more accurate and trustworthy. In the second phase, the fine-grained approach is proposed, aiming to examine the relationship between each of the candidate citations and the local context of a certain bracket. Therefore, this phase can further recommend the best citations to a certain bracket from the candidate citations. Nevertheless, the best citations in different brackets are very likely to be identical. Also, the number of citations might be improper. As a result, the third phase aims to remove overlapping citations and adjust the number of citations for each bracket. From the user's perspective, the proposed ICRM can help scholars save time in searching and comparing a large amount of related work, thereby enhancing their work efficiency. Moreover, it offers more precise and personalized citation recommendations, assisting researchers in quickly locating relevant literature, thus improving the quality and accuracy of their research.

The following gives the major contributions of this paper:

(1) ***Explicable and transparent document vectors***

In the first phase of the proposed mechanism, the document vectors are generated by using the Weighted Bag of Words technology, which extracts some keywords to play the role of features for the documents. Therefore, the document vectors are explicable and transparent. Compared to studies [10] and [14], the candidate citation recommended in this paper are more accurate and reliable.

(2) ***Diverse techniques of natural language processing***

Various techniques are used in different phases. In the coarse-grained approach, the Weighted Bag of Words is used to extract keywords from documents, while in the fine-grained approach, BERT [16, 17] is used to select a certain number of the best candidate citation recommendations based on how similar the candidate recommendations are to the local context of the target bracket. On the contrary, studies [10] and [14] used only the techniques of deep learning, such as LSTM or BERT, to achieve the proposed goals. The use of various techniques can help increase the accuracy and reliability of the results of recommended citations.

(3) ***The elimination of the repeated citation recommendations***

The third phase further applies the Voronoi Diagram [18, 19] to partition the overlapped brackets such that the repeated citations falling in Voronoi Cell

will be assigned to the bracket in that cell. As a result, all repeated citations can be removed.

(4) ***The recommendation of an appropriate number of citations to the target bracket***

Based on the user-predefined lower and upper bounds on the number of citations for each bracket, the third phase also adjusts the number of citations for each bracket. This guarantees the number of citations in each bracket satisfies the user requirement.

The remaining part of this paper is organized as follows. Section II presents the existing studies related to this paper. Section III presents the assumption and problem formulation of this study. Section IV presents the proposed citation recommendation mechanism. The performance of the proposed mechanism is then evaluated in Section V. Finally, Section VI concludes this work and gives the future work of this study.

2. Related work

In the literature, many studies have been proposed for citation recommendation. According to different NLP technologies of citation recommendation, they can be divided into two categories: Context-based and Attention-based Recommendation.

2.1. Context-based recommendation

Many studies [8–10] used traditional product recommendations for uses. These approaches adopted the content-based, user-based or collaborative filter mechanisms and considered the papers as the product. Therefore, these mechanisms were applied to extract the relationship between authors (user) and papers (product). Then a set of papers can be recommended as the candidates of the citations to a certain author. The process of relationship extraction between authors and papers mainly applied the Bag-of-Words model to generate a vector for each paper or for each author. Then the cosine similarity mechanism was applied to find a set of candidates as the citation recommendation.

Lee et al. [8] presented a recommendation mechanism that recommended articles relevant to the research field of the users based on the traditional product recommendation mechanisms. Finally, the K-Nearest Neighbors was applied to estimate the preference of a target user and then recommended the most preferred papers. Nogueira et al. [9] proposed a two-stage approach: candidate generation followed by reranking. Initially, they assigned a vector to each paper based on the Bag of Word model. Then the abstract and title of the target paper are treated as a query. By taking the query and each vector of document as a pair of inputs, the BERT model will generate a score to represent the relation between the target paper and the input document. According to the scores, the papers with higher scores were selected as the candidates for citation recommendation. However, the Bag-of-Words techniques didn't consider the context of the words, which represented the semantic similarity between the articles. Dai et al. [10] proposed a novel neural network model called ASL (Attentive Stacked Denoising AutoEncoders (ASDAE) with Bi-LSTM) for context-aware citation recommendation. The proposed ASDAE was used to capture the global attention from citation context when encoding cited paper. They also extended Bi-LSTM with a local attention layer to learn the hidden vectors for citation context, which could capture the key information for effective embedding and benefit for extracting suitable citation context. Personalization also embedded valuable author information to produce personalized citation recommendations. To overcome this limitation, this paper uses the Feature-based Ruler to exploit the context relation between the target article and the considered articles. The Feature-based Ruler extracts the features based on the word2vec model which assigns a vector based on n-skip gram or CBOW. Since the n-skip gram and CBOW establishing a vector to each word is based on a large number of contexts, the vector embeds the semantic similarity among articles. This paper also utilizes the Voronoi Diagram to partition the overlapped citations, aiming to remove the redundant citations and balance the citations.

2.2. BERT-based recommendation

Many studies [13–15] proposed embedding-based approaches to capture the resemblance between the query and the considered article according to the cosine distance or the Euclidean distance between their embeddings. Gökçe et al. [13] derived a subset of papers from the databases according to a user-defined keyword-based filter. They used the document embedding model, called Sent2Vec, to rank the filtered papers. Jeong et al. [14] proposed a BERT-GCN model for context-aware paper citation recommendation. It computed for each paper embeddings of the citation graph and the query context. The learning presentation of context sentences, through

pre-trained BERT, achieved high performance. The GCN model was used to represent the citation relationship between papers and to extract a learning representation of them. The BERT-GCN model was evaluated on small datasets of only thousands of papers, partly due to the high cost of computing the GCN, which limited its scalability for recommending citations from a large paper database. In addition, similar to the disadvantage of the previous category, the recommendation only considered the global context of the target article. To recommend citations for the bracket of the target article, the local context near by the bracket is still needed to be further considered.

Gu et al. [15] proposed a novel two-stage local citation recommendation system, which contained prefetching stage and reranking stage. The prefetching stage used an embedding-based paper retrieval system, in which a Siamese text encoder first precomputes a vector-based embedding for each paper in the database. The query text was then mapped into the same embedding space to retrieve the K nearest neighbors of the query vector. To encode queries and papers of various lengths in a memory efficient way, it designed a two-layer Hierarchical Attention-based text encoder (HAtten) that first computed paragraph embedding and then computed from the paragraph embedding the query and document embedding using a self-attention mechanism. After that, the reranking stage fine-tuned the SciBERT to rerank the candidates retrieved by the HAtten prefetching model. However, the functions of the proposed two stages were similar, which major relied on the attention scheme to extract the relation between citations of previous referenced papers and suffered with overlapping references. This paper proposes Feature-based Ruler and attention-based mechanisms that recommend candidates in two stages, the coarse grain and fine-grain. The two mechanisms can extract the features of each article based on different aspects. In addition, the proposed ICRM adopts the Voronoi Diagram to tackle the overlapping brackets. An Improper-Size Adjustment (ISA) Mechanism is proposed to further balance the number of citations for brackets.

Table 1 presents a detailed comparison between the proposed ICRM and prior related work in various aspects, including context extraction, embedding method, scalability, overlapping citations and citation balance. It highlights the limitations of existing research methods and illustrates how ICRM addresses these limitations with its innovative solutions and advantages.

For context extraction, previous research [8–10] depended heavily on the Bag-of-Words model and lacked a comprehensive understanding of context. The proposed ICRM utilizes a combination of TF-IDF and WCBW to extract more diverse features, enhancing the context and relevance of citations. In terms of the embedding method, ICRM introduces SCI-BERT with features derived from both TF-IDF and WCBW, which contrasts with the homogeneous feature extraction via embedding [13–15] used in previous studies, offering more accurate and complementary feature extraction. In addressing scalability, models like BERT-GCN [14] demonstrated limited scalability. The proposed ICRM uses Voronoi Diagrams for overlapping citations, allowing for more efficient handling of large databases and providing unique categorization. For the challenge of overlapping citations, prior research [13–15] suffered from overlapping references. The ICRM effectively utilizes Voronoi Diagrams for better management and differentiation of such citations. Lastly, in the aspect of citation balance, the ICRM introduces an Improper-Size Adjustment Mechanism, addressing the absence of a balancing mechanism in previous studies [9, 10] and ensuring a balanced distribution of citations across various brackets.

3. Network environment and problem formulation

This section introduces the problem statement and objectives of this study.

3.1. Problem statement

Assume that there is a set $D=\{D_1,D_2,\ldots,D_m\}$ which consists of m papers D_i, for $1\le i\le m$. Let U_i,A_i,T_i,K_i,C_i denote the authors, abstract, title, keywords, and conclusion of the paper D_i, respectively. The paper D_i can be denoted by Exp. (1).

$$D_i = \{U_i, A_i, T_i, K_i, C_i\}\,(1 \le i \le m) \qquad (1)$$

Let D^t denote the target paper that needs to find citations. Let U^t,A^t,T^t,K^t,C^t, and L^t denote the authors, abstract, title, keywords, conclusion, and citations of the paper D^t, respectively. The paper D^t can be denoted by Exp. (2).

$$D^t = \left\{U^t, A^t, T^t, K^t, C^t, L^t\right\} \qquad (2)$$

Table 1
The proposed ICRM compared previous related work

Aspect	Related research limitations	The proposed ICRM	The proposed ICRM solution and advantages
Context Extraction	Relies on Bag-of-Words, lacks context understanding ([8–10])	TF-IDF+WCBW for diverse features	Enhanced context and relevance in citations
Embedding Method	Homogeneous feature extraction via embedding ([13–15])	SCI-BERT with diverse features from TF-IDF/ WCBW	Complementary feature extraction for accuracy
Scalability	Limited scalability in models like BERT-GCN ([14])	Voronoi Diagrams for overlapping citations	Efficient handling of large databases, unique categorization
Overlapping Citations	Suffers from overlapping references ([13–15])	Voronoi Diagram	Effective management and differentiation of overlapping citations
Citation Balance	No mechanism for balancing citations ([9, 10])	Improper-Size Adjustment Mechanism	Balanced distribution of citations across brackets

Assume that there is an ordered list of citations, considered as the labels, which are cited in n positions of D^t. The citations L^t can be denoted by Exp. (3).

$$L^t = \left(L_1^t, L_2^t, \ldots, L_n^t\right) \tag{3}$$

Let M be a mechanism that predicts the citations of D^t. Let $P_j^{M,t}$ denote the prediction of the j-th position of citations L_j^t using mechanism M. Assume that there are α prediction cited references $P_j^t = \left(p_{j,1}^t, p_{j,2}^t, .., p_{j,\alpha}^t\right)$. Let L_j^t denote a list of actually cited references of *j-th* position in the paper D^t. Assume that there are β actual cited references $L_j^t = \left(l_{j,1}^t, l_{j,2}^t, .., l_{j,\beta}^t\right)$.

Let Boolean variable $\lambda_{i,j}^M$ denote whether the paper D_i is in prediction $P_j^{M,t}$. It can be derived by Exp. (4).

$$\lambda_{i,j}^M = \begin{cases} 1, & D_i \in P_j^{M,t} \\ 0, & otherwise \end{cases} \tag{4}$$

Let Boolean variable $\mu_{i,j}$ denote whether the paper D_i is in label L_j^t. It can be derived by Exp. (5).

$$\mu_{i,j} = \begin{cases} 1, & D_i \in L_j^t \\ 0, & otherwise \end{cases} \tag{5}$$

According to the values of $P_{i,j}^M$ and $L_{i,j}$, the confusion matrix can be calculated. Let $TP_{i,j}^M$, $FP_{i,j}^M$, and $FN_{i,j}^M$ denote the true positive, true negative, false positive and false negative of the confusion matrix whether D_i is cited in the *j-th* position L_j^t, respectively $TP_{i,j}^M$ denotes correct prediction that D_i is cited in the *j-th* position L_j^t. $FP_{i,j}^M$ denotes incorrect prediction that D_i is not actually cited in the *j-th* position L_j^t. $FN_{i,j}^M$ denotes incorrect prediction that D_i is cited in the *j-th* position L_j^t.

The $TP_{i,j}^M$, $FP_{i,j}^M$, and $FN_{i,j}^M$ can be derived by Exps. (6)–(8), respectively.

$$TP_{i,j}^M = \begin{cases} 1, & \lambda_{i,j}^M * \mu_{i,j} = 1 \\ 0, & otherwise \end{cases} \tag{6}$$

$$FP_{i,j}^M = \begin{cases} 1, & \lambda_{i,j}^M - \mu_{i,j} = 1 \\ 0, & otherwise \end{cases} \tag{7}$$

$$FN_{i,j}^M = \begin{cases} 1, & \mu_{i,j} - \lambda_{i,j}^M = 1 \\ 0, & otherwise \end{cases} \tag{8}$$

Let TP_j^M, FP_j^M, and FN_j^M denote the true positive, true negative, false positive and false negative of the confusion matrix for all citations of the *j-th* position, respectively. TP_j^M denotes correct prediction that all citations are cited in the *j-th* position L_j^t. FP_j^M denotes incorrect prediction that all citations are not actually cited in the *j-th* position L_j^t. And FN_j^M denotes incorrect prediction that all citations are cited in the *j-th* position L_j^t.

The TP_j^M, FP_j^M, and FN_j^M can be derived by Exps. (9)–(11).

$$TP_j^M = \sum_{i=1}^{m} TP_{i,j}^M \tag{9}$$

$$FP_j^M = \sum_{i=1}^{m} FP_{i,j}^M \tag{10}$$

$$FN_j^M = \sum_{i=1}^{m} FN_{i,j}^M \tag{11}$$

Let TP^M, FP^M, and FN^M denote the true positive, true negative, false positive and false negative of the confusion matrix for n citations of the paper D^t, respectively. TP^M denotes correct prediction that n citations are cited in all brackets. FP^M denotes incorrect prediction that n citations are not actually cited in all brackets. FN^M denotes incorrect prediction that n citations are cited in all brackets.

The TP^M, FP^M, and FN^M can be derived by Exps. (12)–(14).

$$TP^M = \sum_{j=1}^{n} TP_j^M = \sum_{j=1}^{n}\sum_{i=1}^{m} TP_{i,j}^M \tag{12}$$

$$FP^M = \sum_{j=1}^{n} FP_j^M = \sum_{j=1}^{n}\sum_{i=1}^{m} FP_{i,j}^M \tag{13}$$

$$FN^M = \sum_{j=1}^{n} FN_j^M = \sum_{j=1}^{n}\sum_{i=1}^{m} FN_{i,j}^M \tag{14}$$

Let PC^M denote the precision of prediction. It can be calculated by Exp. (15).

$$PC^M = \frac{TP^M}{TP^M + FP^M} = \sum_{j=1}^{n}\sum_{i=1}^{m} TP_{i,j}^M \Bigg/ \left(\sum_{j=1}^{n}\sum_{i=1}^{m} TP_{i,j}^M + \sum_{j=1}^{n}\sum_{i=1}^{m} FP_{i,j}^M \right) \tag{15}$$

Let RC^M denote the recall of prediction. It can be calculated by Exp. (16).

$$RC^M = \frac{TP^M}{TP^M + FN^M} = \sum_{j=1}^{n}\sum_{i=1}^{m} TP_{i,j}^M \Bigg/ \left(\sum_{j=1}^{n}\sum_{i=1}^{m} TP_{i,j}^M + \sum_{j=1}^{n}\sum_{i=1}^{m} FN_{i,j}^M \right) \tag{16}$$

Let $F1^M$ denote the $F1$–*score* which can be calculated by Exp. (17).

$$F1^M = \frac{2 * PC^M * RC^M}{PC^M + RC^M} \tag{17}$$

3.2. Problem objectives

Let Φ denote the set of all possible citation mechanisms, $M \in \Phi$, that predict the citations for the target document D^t.

3.2.1. Objective function

The objective of this paper is to find the best mechanism M_{best} which satisfies:

$$M_{best} = \arg \max_{M \in \Phi} F1^M \tag{18}$$

where the $F1^M$ combines PC^M and RC^M. The PC^M measures the accuracy of the mechanism M when predicting the positive class, while RC^M measures the mechanism M's ability to correctly identify instances of the positive class.

4. The proposed intelligent citation recommendation mechanism

This paper introduces an Intelligent Citation Recommendation Mechanism, called ICRM. The proposed ICRM is designed to automate the process of suggesting the appropriate number of citations for individual brackets within a target document. It consists of three phases: Coarse-grained Weighted Bag of Word (WCBW), Fine-grained SciBERT (FSB), and Citation Adjustment Phases. In the WCBW phase, candidates are identified from document D for each bracket using WCBW and TF-IDF schemes. The FSB phase employs the SciBERT model to evaluate how closely the candidate documents align with the specific context surrounding the brackets in the target document. However, the initial two phases may result in some brackets sharing identical citations. Therefore, the citation adjustment phase aims to resolve this issue by eliminating overlapping citations among different brackets based on user-defined constraints. The following presents the details of each phase.

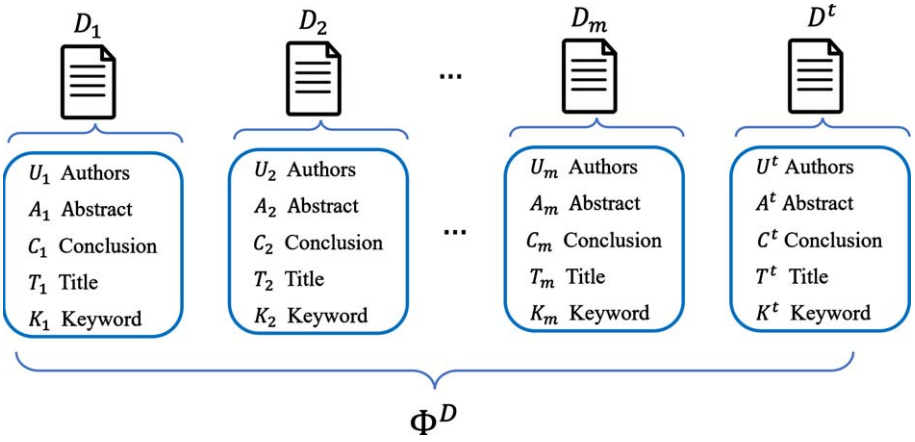


Fig. 1. The content of Φ^D.

4.1. *Coarse-grained weighted bag of word (WCBW) phase*

This phase aims to find the top η appropriate candidates from D for each bracket of the target document. The η candidates are selected based on Weighted Bag of Words and TF-IDF schemes.

Recall that there are m articles in database D. Recall that notations U_i,A_i,T_i,K_i and C_i denote authors, abstract, title, keywords, and conclusion of the article D_i, respectively. That is $D_i = \{U_i,A_i,T_i,K_i,C_i\}$. And D^t is the target paper. Let Φ^D denote the set of words in all documents D_i, for $1 \le i \le m$, as shown in Fig. 1. That is,

$$\Phi^D = \left\{ w \,\middle|\, w \in D_i \text{ or } w \in D^t, \text{ for } 1 \le i \le m \right\} \tag{19}$$

Let notations U^t,A^t,T^t,K^t and C^t denote authors, abstract, title, keywords, and conclusion of the target paper D^t, respectively. Let Φ_i denote the union of A_i and C_i. That is,

$$\Phi_i = A_i + C_i \tag{20}$$

Let Φ_t denote the union of A^t and C^t. That is,

$$\Phi_t = A^t + C^t \tag{21}$$

The next step aims to extract the most important q keywords to organize a Weighted Bag of Words from Φ_t and Φ_i for $1 \le i \le m$. Let $f^{TF-IDF}(w, D_i, \Phi^D)$ denote the TF-IDF function applied on word w in D_i. Let v_{w,Φ_i}^{TF-IDF} denote the TF-IDF value of each word $w \in \Phi_i$. That is,

$$v_{w,\Phi_i}^{TF-IDF} = f^{TF-IDF}\left(w, \Phi_i, \Phi^D\right) \tag{22}$$

The value of v_{w,Φ_i}^{TF-IDF} indicates the importance of the word w which did not frequently appear in other documents but appeared in document Φ_i frequently. Let $max^q(V)$ be the maximum function that returns the q largest values in set V. Let $F_i^{A,C}$ denote the set of the q most important words in Φ_i, where notations A and C denote the words in Abstract and Conclusions. The $F_i^{A,C}$ can be obtained by applying Exp. (23).

$$F_i^{A,C} = \arg \max_{w \in \Phi_i}^{q} \left(v_{w,\Phi_i}^{TF-IDF} \right) \tag{23}$$

Similarly, we may extract the most important p words to organize another Bag of words from Φ_t. Let v_{w,Φ_t}^{TF-IDF} denote the TF-IDF value of each word $w \in \Phi_t$. That is,

$$v_{w,\Phi_i}^{TF-IDF} = f^{TF-IDF}\left(w, \Phi_t, \Phi^D\right) \tag{24}$$

Let $F^{t,A,C}$ denote the set of the p most important words in D^t, where A and C denote the words in Abstract and Conclusions. The $F^{t,A,C}$ can be obtained by applying the following Exp. (25).

$$F^{t,A,C} = \arg \max_{w \in \Phi_t}^{p} \left(v_{w,\Phi_i}^{TF-IDF} \right) \tag{25}$$

Till now, we have obtained the p and q most important words from Φ_t and Φ_i, respectively.

Next, we will illustrate how to present the document vector using the features in F. Let f_j denote the j-th feature in $F^{t,A,C}$. That is,

$$F^{t,A,C} = \left\{ f_i^t \,|1 \le i \le p \right\} \tag{26}$$

Till now, the set of $F^{t,A,C}$ has collected the p most important words which are extracted from the abstract and conclusion of the target document D^t. The next step is to include the authors, keywords and title

in $F^{t,A,C}$. Let F^t be the set extended by including authors, keywords and title of D^t to $F^{t,A,C}$. That is,

$$F^t = F^{t,A,C} + U^t + T^t + T^t \tag{27}$$

Let β be the number of words in F^t. The β words will be used as the features to measure the similarity between the target document D^t and each document D_i.

Similarly, each D_i should extract the important features to compare the similarity between itself and D^t. Recall that $F_i^{A,C}$ denotes the set of the most important q words extracted from the abstract and conclusion of the document D_i. Since the words in authors, title and keywords are very important, they should be reserved to represent the features of each D_i. Let F_i be the set extended by including authors, keywords and title of D_i to $F_i^{A,C}$. That is,

$$F_i = F_i^{A,C} + U_i + T_i + K_i \tag{28}$$

The words in F^t will be organized as a Bag of Words to construct a vector for each D_i. However, the Bag of Words uses the words in the Bag to represent a vector of a sentence, a paragraph or a document. One disadvantage of the Bag of Words is that the vector only presents the property of appearance of those words in the Bag. That is, the relationship of the words that appeared in a sentence has been ignored. Since the word2vec adopts the *n-skip* gram or *CBOW* to generate the vector for each word w by looking at the relationship of the word w and the neighboring words that appeared in the same sentence of a large number of documents in WiKi, the vector can represent a certain level of semantic.

To cope with this problem, the Weighted Bag of Words is proposed. Each word in F^t should be transferred to a vector by applying the Word2Vec. Let V^t denote the ordered list each element is a vector transferred from each word in F^t. The V^t will be treated as the Bag of Words. When the Word2Vec generates the vector for one word, it has been trained to extract the relationship of those words in the same sentence. Therefore, the embedded vector of each word has the advantage of exploiting the relationship among words in the same sentence. To achieve this, each feature word in F^t and F_i should be transferred to an embedded vector. Let f^{w2v} be the function of word2vec which returns the embedding vector of a given input word. Let $w_{i,j} \in F_i$ denote the j-th word in F_i. That is, $w_{i,j}$ is an important keyword of the article Φ_i. Let $v_{i,j}$ be the embedding vector generated by f^{w2v} for a given input word $w_{i,j} \in F_i$. Let V_i denote the set of all $v_{i,j}$. That is,

$$\begin{gathered} v_{i,j} = f^{w2v}\left(w_{i,j}\right) \quad \text{for each } w_{i,j} \in F_i \\ V_i = \left\{ v_{i,j} \,\middle|\, v_{i,j} = f^{w2v}\left(w_{i,j}\right) \quad \text{for each } w_{i,j} \in F_i \right\} \end{gathered} \tag{29}$$

Similarly, let $w_i^t \in F^t$ denote the i-th word in F^t. That is, w_i^t is an important keyword of the target article D^t. Let v_k^t be the embedding vector generated by f^{w2v} for a given input word $w_i^t \in F^t$. Let V^t denote the set of all v_i^t. That is,

$$\begin{gathered} v_i^t = f^{w2v}\left(w_i^t\right) \quad \text{for each } w_i^t \in F_i \\ V^t = \left\{ v_i^t \,\middle|\, v_i^t = f^{w2v}\left(w_i^t\right) \quad \text{for each } w_i^t \in F^t \right\} \end{gathered} \tag{30}$$

Figure 2 gives a conceptual explanation of the abovementioned operations. As shown in Fig. 2, F_i collects α important words from each document D_i. That is, it uses TF-IDF to initially identify the top q important words from the abstract and conclusions of D_i. It then selects α-q important words from authors, title, and keywords to create F_i. Afterward, these words are converted into word vectors using word2vec. Similarity, F^t collects β important words from D^t. The V_i and V^t and be derived through the Exps. (29) and (30), respectively.

Till now, we have transformed each important word of the article D_i and D^t to a vector by applying the word2vec scheme.

The next step is to represent each document V_i as a vector DV_i=($\rho_{i,1}$,$\rho_{i,2}$,… ,$\rho_{i,\beta}$) with length β. Then the cos similarity scheme can be applied to select the η most related documents as the candidates for the target document D^t. Let v_k^t be the k-th vector in V^t. Let $v_k^{best} \in V_i$ be the vector with the smallest distance with v_k^t. That is,

$$\begin{gathered} v_k^{best} = arg\,max\, d(v_k^t, v_i) \text{ for each} \\ v_k^t \in V^t \end{gathered} \tag{31}$$

where $d(v_k^t, v_i)$ denotes the distance of the vector $v_k^t \in V^t$ and the vector $v_i \in V_i$. Since v_k^t and v_i are two vectors, the distance function $d(v_k^t, v_i)$ can be implemented by applying cos similarity as follows.

$$d(v_k^t, v_i) = \frac{v_k^t, v_i}{\| v_k^t \| \times \| v_i \|} \tag{32}$$

Let $Pi, k = v_k^{best}$. That is, put $d(v_k^t, v_i)$ as the value of the k-th position of the vector DV_i. Then each document vector can be obtained. As shown in Fig. 3, each document D_i uses a Weighted Bag of Words DV_i to presents its features. The DV_i is

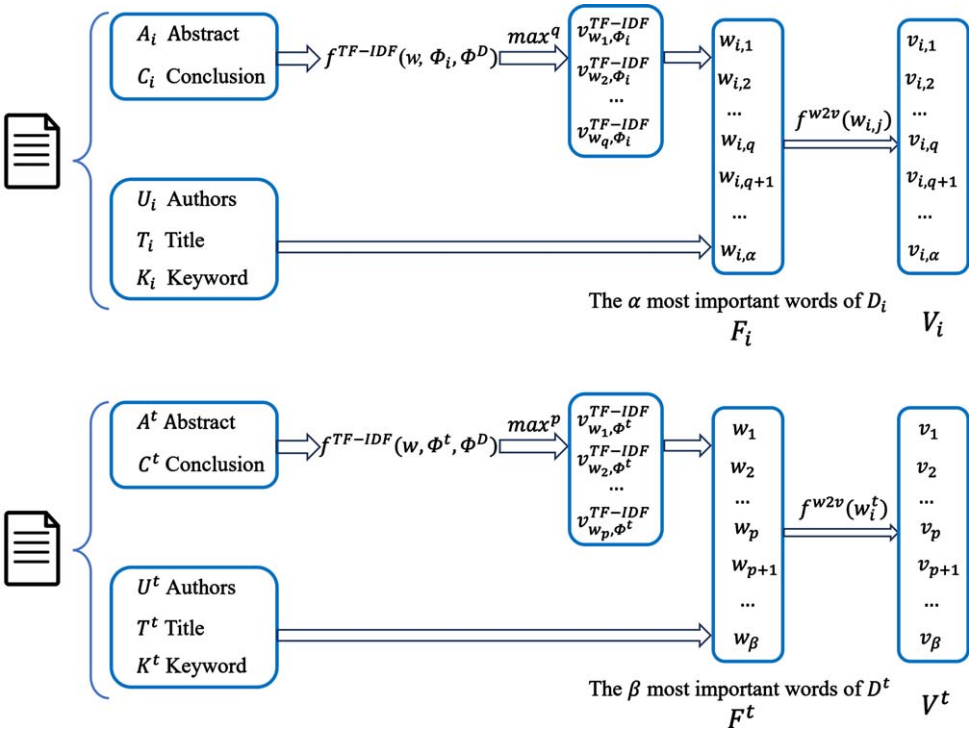


Fig. 2. The conceptual explanation of WCBW.

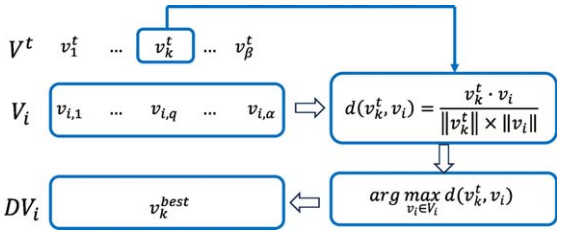


Fig. 3. Weighted Bag of Words (WBOW).

derived based on V^i and V^t. for the document vector DV_i=($\rho_{i,1}$,$\rho_{i,2}$,... ,$\rho_{i,\beta}$) with β tuples, the value of each tuple $\rho_{i,j}$ is a weight, denoting the distance of the whole document and the j-th keyword $v_j^t \in V^t$ of document D^t. The value of $\rho_{i,j}$ is a weight value in a range (0, 1), rather than 0 or 1 as defined in the traditional Bag of Word. Therefore, this mechanism is called Weighted Bag of Word, denoted as WBOW in short.

Next, we will select the η candidates that are mostly related to target paper D^t using the document vectors. As shown in Fig. 4, let D^I denote the candidates found by the first phase of the proposed mechanism. Let $\delta_{i,t} = f^{cos}(DV_i, V^t)$ denote the cosine similarity function which compares the documents vectors V_i and V_t and outputs a similarity value $\delta_{i,t}$. That is,

$$\delta_{i,t} = \frac{v_i \cdot V^t}{\|v_i\| \times \|v^t\|} \tag{33}$$

The set D^I of k candidates can be derived by applying Exp. (34) as shown in the following.

$$D^I = arg\ max^n_{\tilde{D}_i \in D}(\delta_{i,t}) \tag{34}$$

4.2. Fine-grained SciBERT (FSB) phase

During this phase, the SciBERT model is utilized to assess the degree of alignment between the candidate documents and the contextual information surrounding the brackets within the target document.

Let $D^i = \{\tilde{D}_1, \tilde{D}_2, \ldots, \tilde{D}_\eta\}$ denote the set of η candidates selected by Phase A. Let $\tilde{A}_i$,$\tilde{T}_i$, $\tilde{K}_i$ and $\tilde{C}_i$ denote the abstract, title, keywords and conclusions of the i-th candidate, respectively. Let L_j denote the local context of the bracket b_j. The L_j could be the closest r words before and after the bracket b_j. Let $\Phi^{Local}_{t,j}$ denote the union of Φ_t and L_j. That is,

$$\Phi^{Local}_j = A^t + T^t + K^t + C^t + L_j \tag{35}$$

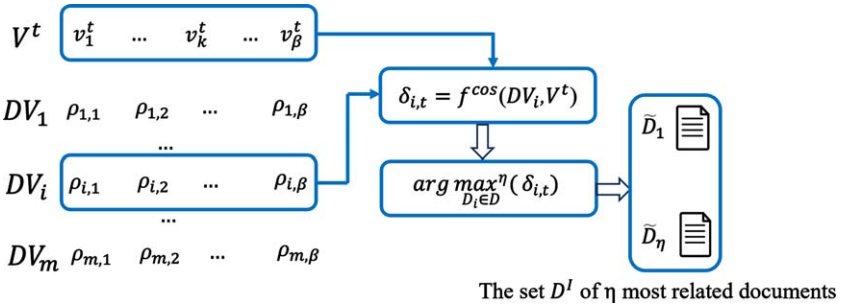


Fig. 4. The most related candidate documents are selected according to their WBOWs.

Let function $f^{sciBERT}$ take $\Phi^{Local}_{t,j}$ and Φ_i as inputs and returns value $v^{sci-BERT}_{D^t,\,D_i}$ as its output. That is,

$$dv^{sci-BERT}_{D^t,\,\tilde{D}_i} = f^{sciBERT}(\Phi^{Local}_{t,\,j},\,\Phi_i) \tag{36}$$

A large value of $v^{sci-BERT}_{D^t,\,D_i}$ indicates that document D_i is more related to the local context of the bracket b_j of D^t.

Let n_j denote the number of required citations for the bracket b_j. The next step is to select the n_j documents that are most related to the local context of the bracket b_j from k candidates. Let C_j be the set of recommended citations for bracket b_j. The value of C_j can be derived by Exp. (37).

$$C_j = arg\,max^{n_j}_{\tilde{D}_i \in D^t}(v^{sci-BERT}_{D^t,\,\tilde{D}_i}) \tag{37}$$

4.3. Citation adjustment phase

This phase aims to adjust the recommendation results obtained from the previous two phases. Each bracket b_j of D^t has been recommended with a set C_j consisting of n_j recommended citations for b_j. However, this result still has two problems: overlapping and improper size problems. Firstly, the overlapping problem refers to the overlapping citations existing in two or more brackets. This occurs because our model recommends the same citations in different brackets, especially for the neighboring brackets. The other problem is the improper size that some certain brackets might have too many or not enough citation recommendations.

4.3.1. Overlapping adjustment mechanism

This paper introduces the "distance-based mechanism," which employs Voronoi diagrams to partition citation points, effectively addressing the issue of overlapping citations by considering the distances between all possible pairs of brackets and citations.

Let $R^{citation}$ denote the set of all recommended citations. That is,

$$R^{citation} = \cup^{n}_{j=1} C_j \tag{38}$$

Let $O^{citation}$ denote the set of citations appearing in more than one bracket. Let $O^{bracket}$ denote the set of brackets with overlapping citation recommendations. That is,

$$O^{citation} = \{c_i \,|c_i \in C_j \cap C_k \;\; for\; j \neq k\}$$

$$O^{bracket} = \{b_j \,|c_i \in C_j \cap C_k \, for\; j \neq k\} \tag{39}$$

The next step is to assign a vector to each element in $O^{citation}$ and $O^{bracket}$. The vector therefore can be treated as a point in pace, aiming to define the distances among these points. The distances among these points are useful to decide the bracket to which each overlapped citation should be finally belonging.

Herein, we will apply the function f^{BERT} which takes Φ_i as its input. Let $v^{BERT}_{c_i}$ denote the output of f^{BERT}. That is,

$$v^{BERT}_{c_i} = f^{BERT}(\Phi_i) \tag{40}$$

Therefore, each $c_i \in O^{citation}$ already has a vector which is the vector corresponding to the vector V_i of document $D_i \in D \cup D^t$. However, each bracket still needs a vector. The next step is to assign a vector to each bracket. Recall that notation L_j denotes the closest r words before and after the bracket b_j. The r words in L_j represent the local features of the bracket b_j. Let v^{BERT}_{bj} denote the vector of the bracket b_j. The value of v^{BERT}_{bj} can be derived by applying Exp. (41).

$$v^{BERT}_{bj} = f^{BERT}(L_j) \tag{41}$$

Till now, each element in $O^{citation}$ and $O^{bracket}$ has its own vector and can be mapped as a point in the

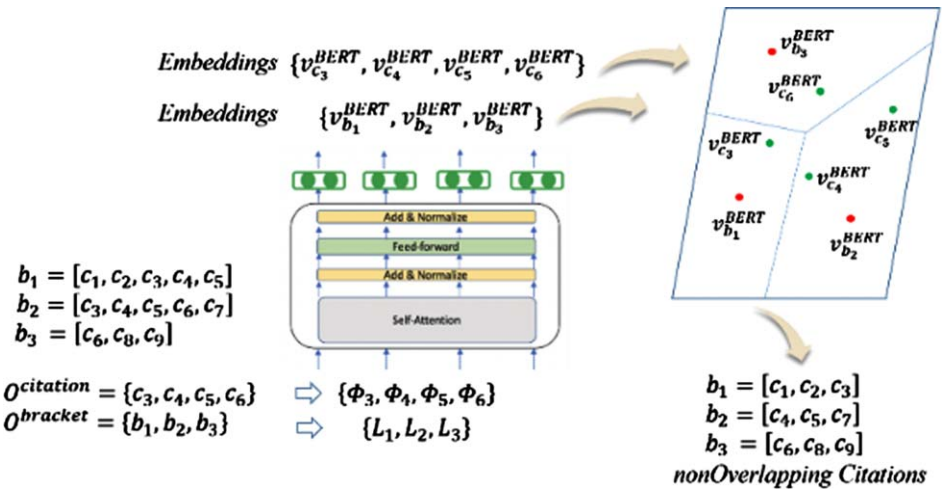


Fig. 5. The set of overlapped citations Ocitation = $\{c_3, c_3, c_3, c_3\}$ of brackets b_1, b_2 and b_3 can be partitioned into different brackets.

vector space. The next step is to define the distance $d(a, b)$ between any pair of two points (a, b) as

$$d(a,b) = 1 - f^{cos}(\vec{\to} a, \vec{\to} b) = 1 - \frac{\vec{\to} a \ \vec{\to} b}{\left\| \vec{\to} a \right\| \times \| \vec{\to} b \|}, \tag{42}$$

where $\vec{\to} a$ and $\vec{\to} b$ could be a citation vector or bracket vector.

The following introduces the distance-based mechanism. First, we should construct a Voronoi diagram for all points $b_j \in O^{bracket}$. Let Ω_j be the voronoi cell of bracket point b_j. One principle of Voronoi diagram is that all points in a Voronoi cell Ω_j are closest to the bracket point b_j, as compared with other bracket point b_k for all $j \neq k$. This implies all citations points falling in voronoi cell Ω_j are closest to bracket point b_j. As a result, each overlapped citation $c_i \in O^{citation}$ can be assigned to one bracket $b_j \in O^{bracket}$ if the following condition is satisfied.

$$b_j = arg \min_{b_k \in O^{bracket}} d(b_k, c_i) \tag{43}$$

Assume that the citation point c_i falling in Ω_j. This indicates that c_i and b_j satisfy Exp. (43). Therefore, the overlapped citation point c_i will be included in bracket b_j.

Figure 5 gives an example where all elements in $O^{citation}$ and $O^{bracket}$ have been mapped to the two-dimensional vector space. There are three bracket vectors $b_1 = [c_1,c_2,c_3,c_4,c_5]$, $b_2 = [c_3,c_4,c_5,c_6,c_7]$ and $b_3 = [c_6,c_8,c_9]$ where b_1 and b_2 are overlapped and b_2 and b_3 are overlapped. The following presents how to decide that each overlapped citation finally belong to a certain bracket. The set $O^{citation}$ of overlapped citations can be derived as c_3,c_4,c_5,c_6. Then the mechanism prepares Φ_3, Φ_4, Φ_5, Φ_6 according to c_3,c_4,c_5,c_6. Then the L_1,L_2,L_3 can be derived from the closest r words before and after the bracket b_1, b_2, and b_3, respectively. Each element in sets Φ_3, Φ_4, Φ_5, Φ_6 and L_1,L_2,L_3 serves as an input of BERT model. The $v_{c_3}^{BERT}$, $v_{c_4}^{BERT}$, $v_{c_5}^{BERT}$, v_{c6}^{BERT} and $v_{b_1}^{BERT}$, $v_{b_2}^{BERT}$, $v_{b_3}^{BERT}$ can be derived by applying Exps. (40) and (41). These vectors can be mapped onto the space where the voronoi diagram is applied to partition the space into three disjoint areas. According to the falling areas of $v_{c_3}^{BERT}$, $v_{c_4}^{BERT}$, $v_{c_5}^{BERT}$, v_{c6}^{BERT}, the overlapped citations can be partitioned into different brackets based on distance-based mechanism, resulting in the following brackets.

$$b_1 = [c_1, c_2, c_3]$$
$$b_2 = [c_4, c_5, c_7]$$
$$b_3 = [c_6, c_8, c_9]$$

4.3.2. *Improper-size adjustment mechanism*

This mechanism aims to cope with the problem of imbalanced number of recommended citations among brackets. Let $C_j^{distance}$ and $s_j^{distance}$ denote the set of citations obtained from distance-based mechanism and the size of $C_j^{distance}$ for bracket b_j, respectively. Let notations α and β denote the user-defined lower and upper bounds for the number of citations in each bracket, where $\alpha \leq \beta$. That is, the number of citations $s_j^{distance}$ in bracket b_j should be within the bounds (α,β). Let B^{less} and B^{more} be the sets of brackets whose number of citations is smaller

than α and larger than β, respectively. That is,

$$B^{less} = \{b_j \mid s_j^{distance} < \alpha\} \text{ and } B^{more} = \{b_j \mid s_j^{distance} > \beta\} \quad (44)$$

Let $B^{sort\_less} = (\overrightarrow{\frown} b_1, \ldots, \overrightarrow{\frown} b_{|B^{less}|})$ denote the sorted list according to the size $s_j^{distance}$ of each $b_j \in B^{less}$ in an ascending order. Similarly, let $B^{sort\_more} = (\overrightarrow{\frown} b_1, \ldots, \overrightarrow{\frown} b_{|B^{more}|})$ denote the sorted list according to the size $s_j^{distance}$ of each $b_j \in B^{more}$ in a descending order. The Improper-Size Adjustment (ISA) Mechanism is presented as shown in Table 2. It iterates through each bracket $\overrightarrow{\frown} b_i$ in the set B^{sort_less}. For each $\overrightarrow{\frown} b_i$, it checks if there is any overlapped between the citations in C_i and C_j. If there is an overlap between C_i and C_j, the algorithm calculates $k_{overlap}$, which represents the number of overlapping citations between the two brackets. It also calculates k_{need} which is the number of citations needed to be adjusted. Next, the algorithm checks if $k_{overlap}$ is greater than k_{need} and if there are enough citations ($\left|C_j\right| - k_{need} \geq \alpha$) in the bracket associated with $\overrightarrow{\frown} b_j$ to transfer the overlapping citations. If it is the case, it moves k_{need} overlapped citations from C_j to C_i and then breaks the loop. On the contrary, if $k_{overlap}$ is less than or equal to k_{need}, it transfers ($\left|C_j\right| - \alpha + 1$) overlapped citations from C_j to C_i and updates the value of k_{need} accordingly. The algorithm repeats these steps for each bracket in B^{sort_less}, aiming to balance the number of citations between the two sets while handling overlapping citations.

Table 2
The process of Improper-size adjustment mechanism

	Algorithm: Improper-Size Adjustment (ISA) Mechanism
	Input: $B^{sort\_less} = (\overrightarrow{\frown} b_1, \ldots, \overrightarrow{\frown} b_{\left\|B^{less}\right\|})$,
	$B^{sort\_more} = (\overrightarrow{\frown} b_1, \ldots, \overrightarrow{\frown} b_{\left\|B^{more}\right\|})$
	Output: k_{need}
1	for each in B^{sort_less}
2	if C_iÇC$_j \neq$ Æ
3	for each in B^{sort_less}
4	let $k_{overlap} = \|C_i \cap$;
5	let k_{need}=$\alpha - s_i^{distance} + 1$;
6	if $k_{overlap} > k_{need}$and $\left\|C_j\right\| - k_{need} \geq \alpha$:
7	move k_{need} overlapped citations from C_j to C_i;
8	break;
9	elif $k_{overlap} \leq k_{need}$:
10	move $\left\|C_j\right\| - \alpha + 1$ overlapped citations from C_j to C_i;
11	$k_{need} = k_{need} - (\left\|C_j\right\| - \alpha + 1)$;
12	end if
13	end for
14	end if
15	end for

5. Performance evaluation

This section evaluates the performance improvement of the proposed ICRM against Coarse-grained WCBW and the existing HAtten-SciBERT (Hierarchical-Attention Text Encoder and SciBERT-based Reranking) [15]. The WCBW recommends the most related top k papers from D in a coarse-grain manner. The existing HAtten-SciBERT proposed a two-stage local citation recommendation, including prefetching and reranking stages. The prefetching stages scored and ranked all papers in the database to fetch a rough initial subset of candidates by a two-layer Hierarchical Attention-based text encoder (HAtten). On the other hand, the reranking stage was used to fine-tune the SciBERT to rerank the candidates retrieved by the HAtten prefetching model. Finally, the proposed ICRM applies coarse-grain and fine-grain policies to recommend the citations for each bracket. The code of the proposed ICRM can be found at http://aiit.csie.tku.edu.tw/. The following gives the simulation environment and results.

5.1. Data sets

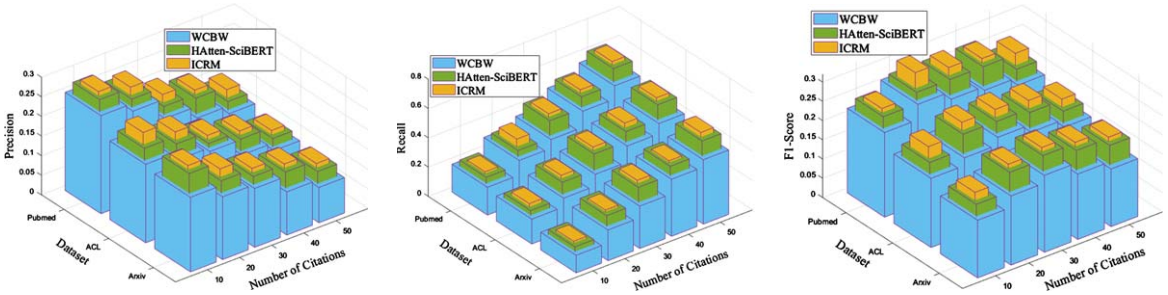
The datasets including PubMed, ACL (Association for Computational Linguistics) and arXiv are used for measuring the performances of the compared studies. The PubMed is a biomedical literature database maintained by the U.S. The National Library of Medicine contains over 37 million abstracts and full-text articles. The ACL Anthology currently hosts 83313 articles on the study of computational linguistics and natural language processing. The arXiv is an online preprint platform, covering papers in fields such as physics, computer science, mathematics, statistics, and biology. This dataset contains a large number of academic papers that can be used for research in areas such as machine learning and natural language processing. It contains over 1.7 million articles.

5.2. Simulation results

Table 3 compares the WCBW, HAtten-SciBERT and ICRM in terms of precision, recall and F1-score. The number of articles in dataset varies ranging from 100 to 500. All the compared mechanisms have the common trend that the precision,

Table 3
Comparison of WCBW, HAtten-SciBERT and ICRM in terms of precision, recall and F1-score

Number of aticles in dataset	Mechanisms	500	400	300	200	100
Precision	WCBW	0.2	0.18	0.17	0.14	0.11
	HAtten-SciBERT	0.25	0.23	0.21	0.19	0.17
	ICRM	0.3	0.27	0.25	0.22	0.19
Recall	WCBW	0.64	0.57	0.52	0.49	0.46
	HAtten-SciBERT	0.73	0.67	0.63	0.59	0.54
	ICRM	0.7	0.69	0.65	0.61	0.56
F1-score	WCBW	0.22	0.2	0.15	0.12	0.09
	HAtten-SciBERT	0.29	0.26	0.23	02	0.17
	ICRM	0.31	0.29	0.26	0.23	0.2



(a) Precision (b) Recall (c) F1-Score

Fig. 6. Comparison of WCBW, HAtten-SciBERT and ICRM with different datasets.

Table 4
Comparison of WCBW, HAtten-SciBERT and ICRM with different datasets in terms of MRR

Dataset	Method	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000
ACL	WCBW	0.29	0.29	0.28	0.28	0.26	0.26	0.25	0.18	0.18	0.17
	HAtten-SciBERT	0.48	0.47	0.43	0.38	0.38	0.36	0.34	0.33	0.31	0.31
	ICRM	0.49	0.48	0.45	0.40	0.38	0.36	0.35	0.34	0.32	0.32
arXiv	WCBW	0.29	0.29	0.29	0.28	0.27	0.26	0.23	0.20	0.19	0.16
	HAtten-SciBERT	0.47	0.46	0.42	0.39	0.37	0.36	0.33	0.32	0.32	0.31
	ICRM	0.49	0.47	0.42	0.40	0.38	0.37	0.36	0.34	0.34	0.31
PubMed	WCBW	0.30	0.30	0.29	0.28	0.26	0.25	0.25	0.22	0.21	0.18
	HAtten-SciBERT	0.50	0.48	0.44	0.38	0.37	0.35	0.34	0.32	0.30	0.31
	ICRM	0.52	0.49	0.45	0.40	0.38	0.36	0.35	0.32	0.32	0.32

recall, and F1 are increased with the number of articles in dataset. More articles in dataset help the model distinguish between relevant and irrelevant citations with greater precision. Consequently, the values of precision, recall and F1-score are increased. The proposed ICRM outperforms the WCBW and HAtten-SciBERT mechanisms in all cases. The main reason is that the proposed ICRM incorporates the WCBW and TF-IDF schemes to recommend the most related papers from D as the potential candidates. Additionally, it further examines the detailed relationship between each candidate paper and the given sentences to enhance the Precision, Recall and F1-score of citation recommendation.

Figure 6 (a), (b) and (c) compare the PubMed, arXiv and ACL in terms of Precision, Recall and F1-Score, respectively. The number of citations in dataset varies ranging from 10 to 50. It is obvious that the performance of ACL dataset is generally better than the other datasets.

Table 4 further compares WCBW and ICRM with the PubMed, arXiv and ACL datasets in terms of MRR (Mean Reciprocal Rank). The value of MRR can be calculated by the Exp. (45).

$$MRR = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{r_j} \tag{45}$$

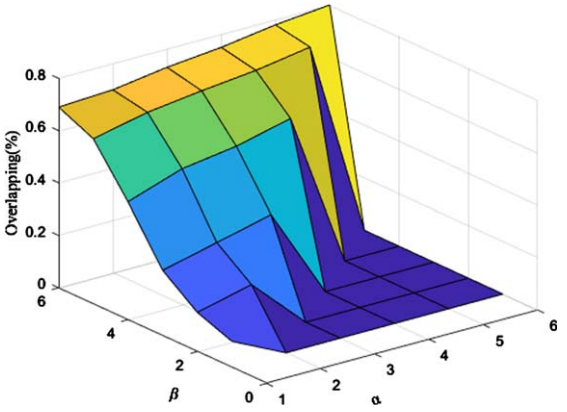


Fig. 7. The relationship between Overlapping and bounds (α, β).

where r_j denotes the rank of document D_j among the predicted papers and document D_i is the target document of this citation bracket. As shown in Table 4, the number of articles in dataset varies ranging from 1000 to 10000. It is obvious that the MRR is decreased with the number of articles in datasets. The reason is that it is difficult for the lower-ranked papers to be accurately recommended in the top positions when the number of articles increases. The proposed ICRM outperforms the WCBW and HAtten-SciBERT in all cases. The reason is that the proposed ICRM adopts FSB phase to select the documents that are most related to the local context and citation adjustment phase to adjust the remediation results obtained from the WCBW phase.

Recall that notations α and β denote the user-defined lower and upper bounds for the number of citations in each bracket, respectively. The values of α and β will highly impact the number of citations for each bracket and also affect the number of overlapping citations.

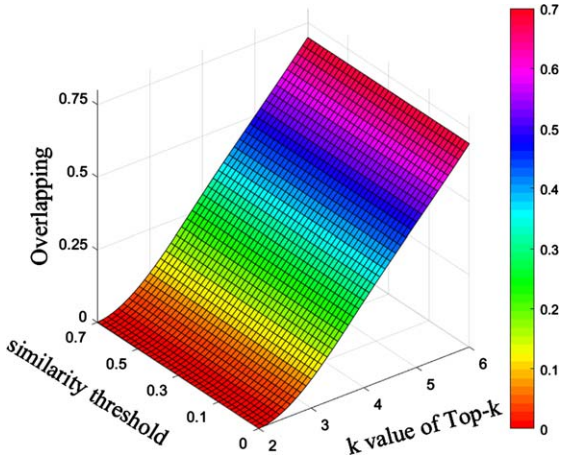
Figure 7 compares the overlapping of different α and β ($\beta \geq \alpha$). The values of α and β vary ranging from 1 to 6. The overlapping refers to the commonality or shared citations among different brackets under various α and β settings. It is calculated by Exp. (46).

$$overlapping = \max_{\forall b_j, b_k \in O^{bracket}} \frac{b_j \cap b_k}{b_j \cup b_k}, \qquad (46)$$

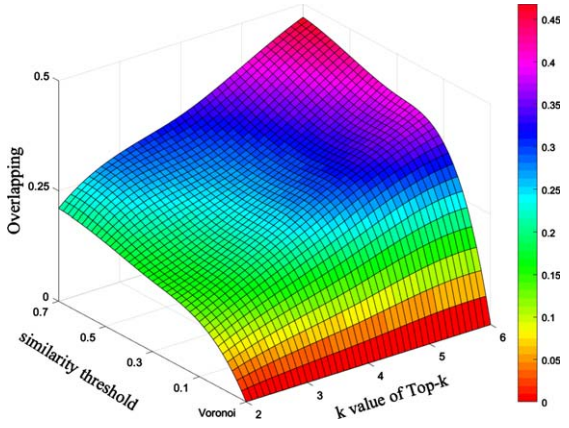
where b_j denotes the set of citations at j^{th} bracket. As shown in Fig. 7, with the increasing of α and β, the overlapping seems to follow an increasing trend, especially in the higher range of α and β. This implies that higher α and β settings may result in more overlapping between citations.

Figure 8 (a) and (b) compare the HAtten-SciBERT and ICRM in terms of overlapping in each bracket, respectively. The top-k documents will be recommended, where $2 \leq k < 6$. The similarity can be calculated by applying Exp. (42). The similarity threshold varies from 0 to 0.7, which is used to allow the maximal similarity for any pair of brackets. A small similarity threshold will restrict the overlapping citations in different brackets. This constraint can reduce the overlapped citations. However, the HAtten-SciBERT did not cope with the overlapping problem. As a result, as shown in Fig. 8(a), the similarity threshold did not impact the overlapping ratio. On the contrary, as shown in Fig. 8(b), the proposed ICRM significantly reduces the overlapping ratio for a smaller similarity threshold value. When the similarity threshold is zero, the ICRM mechanism utilizes Voronoi diagram to assign each citation to a distinct branch, ensuring that there is no duplication as shown in Fig. 8(b).

Figure 9 further investigates the impact of (α, β) on the Voronoi average distance. The range of (α, β) varies from (1,1) to (1,6). We intend to observe two brackets with overlapping citation, and the variation in distance between the overlapping citation and the Voronoi edge formed by the two brackets. If the distance is greater, it indicates that the overlapping citation is more likely to be assigned to a specific bracket based on semantics. In this experiment, we attempt to perform the above-mentioned analysis by introducing overlapping citations into two brackets at different distances within the article. To control the variation in distance, we select these two brackets from different sentences within the same paragraph, or from different paragraphs within the same chapter, or from different chapters within the same article for the analysis and investigation. In the experiment, the distributions of citations in different locations, including inter-sentence, inter-paragraph, and inter-chapter, are represented by the proportions (0.5,0.25,0.25), (0.333,0.333,0.333), (0.25,0.5,0.25), and (0.25,0.25,0.5), respectively. As shown in Fig. 10, we use different colors to denote different distances of the two brackets. The number of citations varies from 10 to 25. As shown in Fig. 10, the Voronoi average distance is shortest for inter-sentence citations and farthest for citations placed between chapters. This occurs because that the semantic context of different brackets will exhibit more noticeable differences as the distance between the brackets within the article increases. This makes it easier to distinguish, based on semantics, which bracket an overlapping



(a) HAtten−SciBERT



(b) ICRM

Fig. 8. Overlapping.

citation belongs to. Consequently, the distance from the Voronoi edge will also be greater. In addition, an increasing number of citations leads to a rising trend in the Voronoi average distance. This indicates that with a greater number of citations, the similarity between citations decreases, resulting in an increased average distance in the Voronoi diagram.

5.3. Limitations

The experimental results indicate that the proposed ICRM can effectively recommend the appropriate number of citations for individual brackets within a document. However, this paper has several limitations, which are presented in the following.

5.3.1. Data accessibility limitation

This study relies on having access to information such as titles, abstracts, authors, and conclusions. In this paper, we assumed the inability to access full-text journal and conference publications. However, if access to, or partial access to, many published conference and journal papers were possible, a more comprehensive and accurate analysis can be conducted. This will help to better understand the relevance of the papers, rather than solely relying on authors, titles, abstracts, and conclusions. We believe that exploring the semantics of the whole document can recommend more suitable citations.

5.3.2. Limitation of contextual sentences nearby the brackets

This paper assumed that when authors give the citation brackets, they should provide complete preceding and subsequent contextual sentences. Then the developed ICRM recommends citations based on the relevance of contextual sentences to related research. In the future, it may be possible to relax such restrictions, allowing authors to provide only the preceding context within parentheses, or even omit the parentheses altogether. In such cases, citation recommendation technology can automatically add citations.

6. Conclusion

This paper introduces the ICRM mechanism for automated citation recommendation within brackets of a target document. The proposed ICRM consists of WCBW, FSB, and Citation Adjustment phases. The WCBW phase efficiently identifies a pool of candidate papers based on authors, abstract, title, keywords and conclusions. The FSB phase adds depth to the recommendations by considering the local context around brackets. It employs advanced techniques like SciBERT to assess document similarity, resulting in more precise citation selections. Finally, the Citation Adjustment phase tackles overlapping citations and ensures that each bracket receives the desired number of citations. It utilizes a distance-based mechanism to assign overlapping citations to their appropriate brackets and maintains balance in citation numbers. The simulation demonstrates that the proposed ICRM holds promise for improving the quality of academic writing and information retrieval systems.

In future work, two aspects can be explored. The first one is the analysis of citation interconnections. This future work might analyze the interconnections and placement of citations in the literature to gain a better understanding of the relationships between

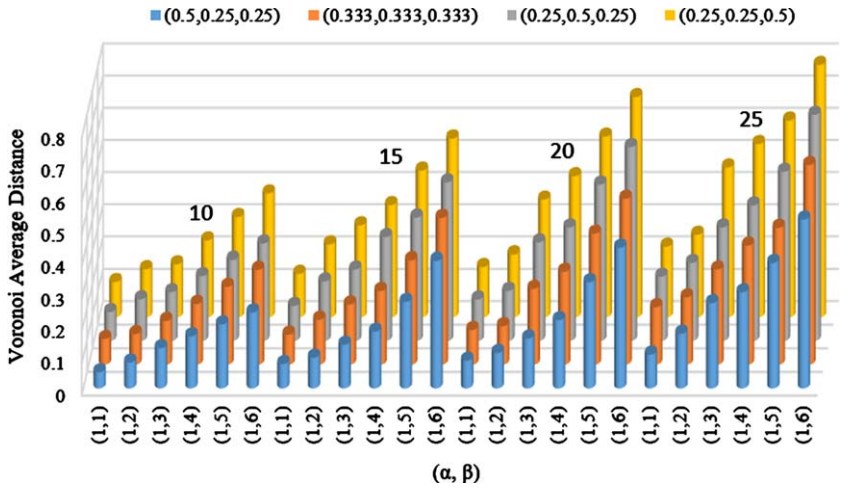


Fig. 9. The relationship between Voronoi average distance and bounds (α, β).

citations. The other future work is the development of a dynamic citation adjustment mechanism. The main task of the dynamic citation adjustment mechanism is to dynamically adjust the number of citations based on the complexity and significance of the target document's content, aiming to enhance the quality and comprehensibility of the document.

References

[1] S. Iqbal, et al., A decade of in-text citation analysis based on natural language processing and machine learning techniques: An overview of empirical studies, *Scientometrics* **126**(8) (2021), 6551–6599.

[2] K. Yue, et al., Natural language processing (NLP) in management research: A literature review, *Journal of Management Analytics* **7**(2) (2020), 139–172.

[3] X. Qin, et al., Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews, *Journal of Clinical Epidemiology* **133** (2021), 121–129.

[4] F. Enríquez, A.T. José and L.S. Tomás, An approach to the use of word embeddings in an opinion classification task, *Expert Systems with Applications* **66** (2016), 1–6.

[5] G. Costa and O. Riccardo, Jointly modeling and simultaneously discovering topics and clusters in text corpora using word vectors, *Information Sciences* **563** (2021), 226–240.

[6] D. Kim, et al., Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec, *Information Sciences* **477** (2019), 15–29.

[7] F. Lan, Research on text similarity measurement hybrid algorithm with term semantic information and TF-IDF method, *Advances in Multimedia* **2022** (2022), 1–11.

[8] J. Lee, et al., Personalized academic research paper recommendation system, *arXiv preprint arXiv* (2013), 1–8.

[9] R. Nogueira, et al., Navigation-based candidate expansion and pretrained language models for citation recommendation, *Scientometrics* **125** (2020), 3001–3016.

[10] T. Dai, et al., Attentive stacked denoising autoencoder with bi-lstm for personalized context-aware citation recommendation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **28** (2019), 553–568.

[11] L. Yang, et al., A LSTM based model for personalized context-aware citation recommendation, *IEEE Access* **6** (2018), 59618–59627.

[12] J. Wang, et al., Deep memory network with bi-lstm for personalized context-aware citation recommendation, *Neurocomputing* **410** (2020), 103–113.

[13] O. Gökçe, et al., Embedding-based scientific literature discovery in a text editor application, *The 58th Annual Meeting of the Association for Computational Linguistics*, Online (2020), 320–326.

[14] C. Jeong, et al., A context-aware citation recommendation model with BERT and graph convolutional networks, *Scientometrics* **124** (2020), 1907–1922.

[15] N. Gu, et al., Local Citation Recommendation with Hierarchical-Attention Text Encoder and SciBERT-Based Reranking, *44th European Conference on IR Research (ECIR)*, Stavanger, Norway (2022), 274–288.

[16] N. Yang, Z. Zhiqiang and F. Huang, A study of BERT-based methods for formal citation identification of scientific data, *Scientometrics* **128** (2023), 1–17.

[17] Y. Lu, et al., Research on semantic representation and citation recommendation of scientific papers with multiple semantics fusion, *Scientometrics* **128**(2) (2023), 1367–1393.

[18] Z. Hao, et al., A node localization algorithm based on Voronoi diagram and support vector machine for wireless sensor networks, *International Journal of Distributed Sensor Networks* **17**(2) (2021), 1550147721993410.

[19] Z. Zarei and B.M. Mozafar, Coverage improvement using Voronoi diagrams in directional sensor networks, *IET Wireless Sensor Systems* **11**(3) (2021), 111–119.