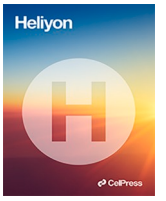
Heliyon 10 (2024) e34685

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Scientific paper recommender system using deep learning and link prediction in citation network

Weijuan Li

Dean's Office, Yellow River Conservancy Technical Institute, Kaifeng, 475004, Henan, China

ARTICLE INFO

Keywords:
Recommender system (RS)
Content-based paper recommendation
Citation network of scientific papers
Text processing

ABSTRACT

Today, the number of published scientific articles is increasing day by day, and this has made the process of searching for articles more difficult. The need to provide specific recommender systems (RSs) for suggesting scientific articles is strongly felt in this situation. Because searching for articles based only on matching the titles or content of other articles is not an efficient process. In this research, the combination of two content analysis and citation network is used to design an RS for scientific articles (RECSA). In RECSA, natural language processing and deep learning techniques are used to process the titles and extract the content attributes of the articles. For this purpose, first, the titles of the articles are pre-processed, and by using the Term Frequency Inverse Document Frequency (TF-IDF) criterion, the importance of each word in the title is estimated. Then the dimensions of the obtained attributes are reduced by using a convolutional neural network (CNN). Then, by using the cosine similarity criterion, the content similarity matrix of the articles is calculated based on the attribute vectors. Also, the link prediction approach is used to analyze the connections of scientific articles' citation network. Finally, in the third step of RECSA, the two similarity matrices calculated in the previous steps are combined using an influence coefficient parameter to obtain the final similarity matrix, and the recommendation operation is based on the highest similarity value. The efficiency of RECSA has been evaluated from different aspects and the results have been compared with previous works. According to the results, utilizing the combination of TF-IDF and CNN for analyzing content-based features, leads to at least 0.32 % improvement in terms of precision compared to previous works. Also, by integrating citation and content-based data, the precision of first suggestion in RECSA would be 99.01 % which indicates the minimum improvement of 0.9 % compared to compared methods. The results show that by using RECSA, the recommendation can be done with higher accuracy and efficiency.

1. Introduction

RSs make using information systems easier for users and can prevent the waste of a huge amount of processing power of the systems as well as spending time searching for the desired content of the users. Today, RSs are widely used in social networks and content provider networks [1]. These types of systems provide suitable recommendations or friend relationships for users based on their interests and characteristics in the network [2]. Nevertheless, not many studies have been done in the field of RSs for scientific articles, and the research works that have been done so far have deficiencies that make their application difficult [3]. On the other hand, the difference in the structure of RSs for scientific articles with other types of RSs has made it impossible to use these systems for the

E-mail address: 2006780043@yrcti.edu.cn.

https://doi.org/10.1016/j.heliyon.2024.e34685
Received 6 May 2023; Received in revised form 13 July 2024; Accepted 15 July 2024
Available online 15 July 2024
2405-8440/© 2024 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

intended purpose. Because researchers usually need to filter the search results of scientific articles to find items that are related to their research field. The filtering operation is very complicated and time-consuming because every day the number of published scientific articles increases [4]. Some researchers have tried to improve RSs for scientific articles by analyzing the citations made to an article or analyzing its content.

Most of the content analysis methods are performed using the process of matching keywords and therefore cannot consider the semantic context of the articles. On the other hand, techniques based on article citation analysis use the number of citations made to an article as well as the context of citing articles to provide recommendations [5]. Although citation-based methods can consider the semantic context of articles to provide recommendations; if the number of citations for an article is low (or there is no reference to a new article), they will not be effective [6]. Meanwhile, it seems that more precise suggestion can be provided by using the dynamic combination of two content and citation-based recommendation strategies; so that by using the combination of the strengths of these two models, more efficient recommendations will be made. This goal has motivated the current research. In this article, the combination of analyzing content and citation network of articles is used to solve the problems of each method. The proposed method, uses the combination of TF-IDF and CNN models to describe the content-based features of articles which is more efficient than basic Natural Language Processing (NLP)-based methods (such as Doc2Vec) in the compact description of content features. Also, this method, unlike the techniques of word filtering or threshold-based removal, does not eliminate the effect of low-repeated words, and as a result, it can provide a more efficient representation of the content characteristics of the text. Also, in the presented approach, the link prediction technique is utilized in order to analyze the citation network and predict the continuation of the authors' collaboration based on their activity records. This strategy can reveal some of the appropriate recommendations (which cannot be identified through content analysis). The contribution of the current article includes the following.

- In this article, a content analysis strategy based on natural language processing and deep learning techniques is presented, which can be used to estimate content similarities with higher efficiency. This strategy first extracts text features in TF-IDF vectors and then uses a Deep Neural Network (DNN) to reduce the number of these attributes.
- In the current article, combination of content features and citation network analysis is used to recommend articles. This combination makes it possible to consider articles in the recommendation process that have higher importance and feedback in addition to the content related to the searched title.

The remainder of the article is structured as follows: In section 2, the previous studies are reviewed. In the third section, the presented approach is described, and in the fourth section, the process of implementing the proposed approach and the results of its performance evaluation is explained. Section 5 expresses the conclusion of the research.

2. Literature review

In [7], a content-based method was presented to recommend practical articles in data sets. This paper, used a paradigm of information retrieval for recommending article, where abstract and title of papers are used for generating semantic features. In this method, several distribution methods such as doc2vec, word2vec, latent Dirichlet allocation, latent semantic analysis, BM25, and TF-IDF are used. The cosine similarity measure between the feature vectors extracted for the articles was used to recommend similar articles. Several re-ranking and normalization mechanisms were proposed in this method to improve the recommendations.

In [8], a recommendation approach based on collaborative filtering and citation network analysis for scientific articles was introduced, which does not depend on the profiles created for previous users and only uses public contextual information in the citation network. In this method, by using the citation context, two-level article-citation relationships are used to find latent connections between articles. In Ref. [9], a citation network-based method for recommending scientific articles was introduced. The advantage of this method is to produce a citation network with more than one cite level. While in other previous methods, including the method presented in Ref. [10], scientific articles are recommended using a citation network with two levels in maximum.

On the other hand, some methods of article recommendation, such as the method proposed in Ref. [11], perform the recommendation operation by processing the keywords in the scientific article. In this method, the keywords requested by the user are searched in the articles and the closest content is recommended to the user. Some methods recommend scientific articles using machine learning techniques. For example, the method presented in Ref. [12] makes recommendations using a DNN. Also, in Ref. [13], the combination of citation networks and hierarchical clustering is used to recommend scientific articles.

In [14], a RS for scientific articles named RefSeer was introduced, which performs the recommendation process using a weighted citation network. It was also emphasized in Ref. [15] that in a citation network, not all connections have the same importance. For this reason, in this article, a criterion for evaluating the importance of citations in scientific articles is presented, which can increase the efficiency of RSs for scientific articles.

Research in Ref. [16], presented a citation recommender system named SentCite, which uses the salient similarity between segments of the documents for recommendation. SentCite, utilizes a CNN model for extracting the citation-requiring sentences from input documents and computes the salient similarity between the extracted sentence and full text of the recommendable documents. Using this strategy in large databases is very time-consuming. In addition, due to the wide range of topics and the variety of topics, it is challenging to train a CNN model to recognize target sentences in real applications.

In [17], a new neural citation network model-based citation recommender system using Bidirectional Encoder Representations from Transformers (BERT) was presented. This research utilized auto-encoding mechanism for learning contextual and citation features. The content features that can be combined in this model require the review of the entire content of the articles, which is

time-consuming and complex in real applications. In Ref. [18], a multi-cell Recurrent Neural Network (RNN) model for recommending scholar articles was proposed. This model preprocesses the documents by removing stop-words and then, uses a probabilistic matching method for document canonicalization. Finally, a RNN is used for ranking document. This research does not consider citation information and also, the RNN model needs high memory for medium to large datasets.

In [19], a deep reinforcement learning model was presented for citation recommendation. This model uses an iterative process for training based on citation networks. In each iteration, it tries to improve the network by adding a new citation link in the network and evaluating it based on the reward operator of the reinforcement learning model. This method does not pay attention to content information, which limits its generality of application. Also, research in Ref. [20] presented a model named Convolutional Citation Networks (ConvCN) for citation graph-based recommendation. Unlike previous graph-based models, ConvCN considers the global citation behavior of papers which is effective in producing more diverse recommendations, but not paying attention to the distance limit, reduces the precision of the model, especially in the first recommendations. In Ref. [21], two content-based recommendation approaches based on machine learning were presented which are suitable for biomedical scientific papers. The first model uses TF-IDF for vectorizing documents and Naïve Bayes for classifier for distinguishing relevant documents. The second model utilizes BERT model for text classification. This research does not consider citation information and need separate training instances for each topic which is hard to implement in real-world applications.

3. The proposed RS for scientific articles (RECSA)

In this section, the details of RECSA for recommending scientific articles using the combination of content analysis techniques and the citation network of articles are expressed. In RECSA, the combination of natural language processing and deep learning techniques is used to process the titles and extract the content features of the articles. Also, link prediction technique has been used for analyzing the connections of the citation network of scientific articles. Finally, by combining these two categories of features, the most suitable articles are recommended as the output of RECSA. The steps of RECSA can be summarized as follows.

1 Processing the titles of articles and extracting content features
2 Analysis of the citation network of articles using the link prediction algorithm
3 Combining content and citation features and providing recommendations

These steps are shown in Fig. 1. Based on Fig. 1, a database consisting of scientific articles is considered. The titles of each article and its citation network, and then keywords in each title are extracted in the first step. To extract the keywords, the stop words in the title are removed and the remaining set of words are stemmed. In the next step, the importance of each word in the title is identified

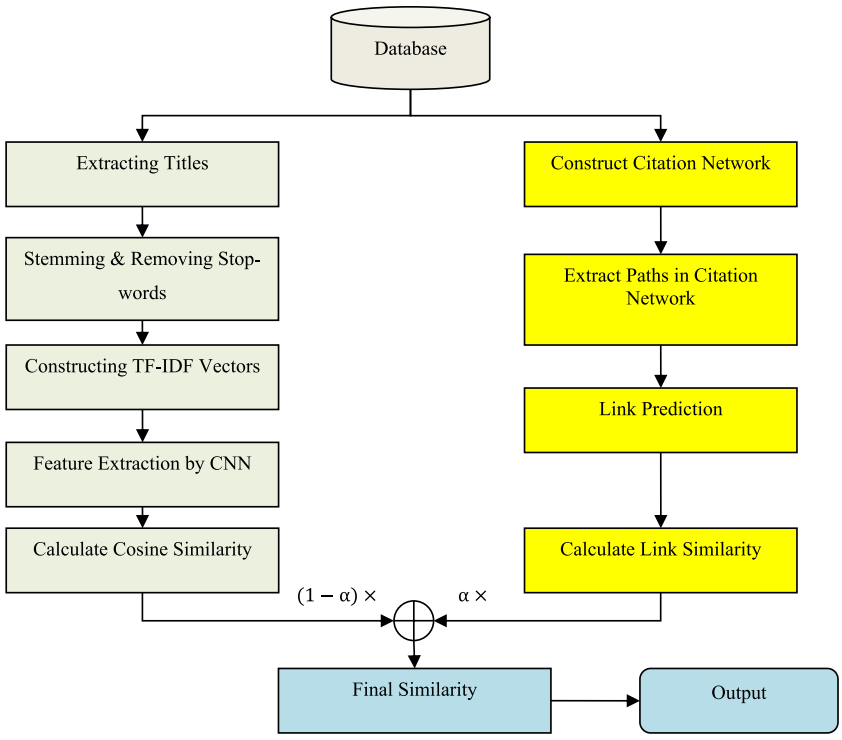


Fig. 1. The process of recommending scientific articles in RECSA.

with the TF-IDF criterion. After doing this for each article, the dimensions of the TF-IDF vectors are reduced using a CNN. At the end of this step, the content similarity matrix of the articles in the database is obtained with the cosine similarity criterion. In the second step of RECSA, the citation network of the articles is evaluated by using the link prediction algorithm and the similarity matrix of the articles' connections in the citation network is calculated. Then, in the third step of RECSA, the two similarity matrices calculated in the previous steps are combined with the influence coefficients α and β to obtain the final similarity matrix. Finally, the recommendation operation is performed based on the maximum value of similarity between the query article and other articles in the database. The details of each step are explained below.

3.1. Processing the titles of articles and extracting content features

For a database consisting of scientific articles where each article has a title, the first phase in the presented approach is to detect the content similarity between the articles. Considering that the title of each article reflects the content and purpose of that article, therefore, by processing the titles of the articles, the similarity between the content of the articles can be examined. For this purpose, the title of all articles should be described as vectors with the same length. In RECSA, TF-IDF vectors and CNN are used to do this. The TF-IDF method has a suitable performance in processing short texts and on the other hand, unlike methods such as word2vec or doc2vec, it is higher adaptable in case of processing unknown expressions. These characteristics have caused TF-IDF strategy to be preferred over word2vec and doc2vec methods in the proposed method. However, the TF-IDF method leads to long feature vectors, which may be challenging in processing huge databases. To solve this problem, CNN model is utilized to reduce the dimensions of this vector.

For this purpose, we first identify the stop words in the titles and remove them. Stop words are words like "be", "from", "as", etc., which do not reflect any key concept in the text. Therefore, removing these words in the title of the models can improve the feature extraction process.

After removing the stop words from the title of the articles, the obtained text is decomposed into its component words, and the process of stemming is performed. Using stemmed words, limits the dimensions of the vector of words or in other words, the vector of features. Accordingly, for each word in the title, suffixes and prefixes are ignored. It should be noted that in RECSA, Porter algorithm is used for stemming [22].

Before extracting TF-IDF features from an article, first the list of unique words of the dataset samples are extracted to form a list such as F with length of n. By doing this, each stemmed word in each article title can be assigned with a weight value. This means that the TF-IDF vector extracted for each article is of length n and the weight value of stemmed word w in an article is calculated using Eq. (1) [23]:

$$TFIDF_w = \frac{T_w}{T_D} \times \log_e\left(\frac{N}{N_w}\right) \tag{1}$$

where T_w shows the frequency of *w* in the article title, and T_D refers to the length of the article title in terms of words. Also, N shows the count of samples in the dataset. Finally, N_w shows the frequency of *w* in the dataset. By weighting each word in each sample of the dataset, a feature matrix will be obtained. The number of rows in this matrix corresponds to the number of dataset samples and can describe its article title, while each column corresponds to a unique stemmed word in list F.

Considering that TF-IDF vectors contain the stem words in the text, the dimensions of these vectors are very large and uncontrollable for large databases that have a high variety of words. To solve this problem, feature reduction techniques can be used. In some previous solutions, word filter techniques or threshold-based methods were used. But it should be noted that some infrequent words

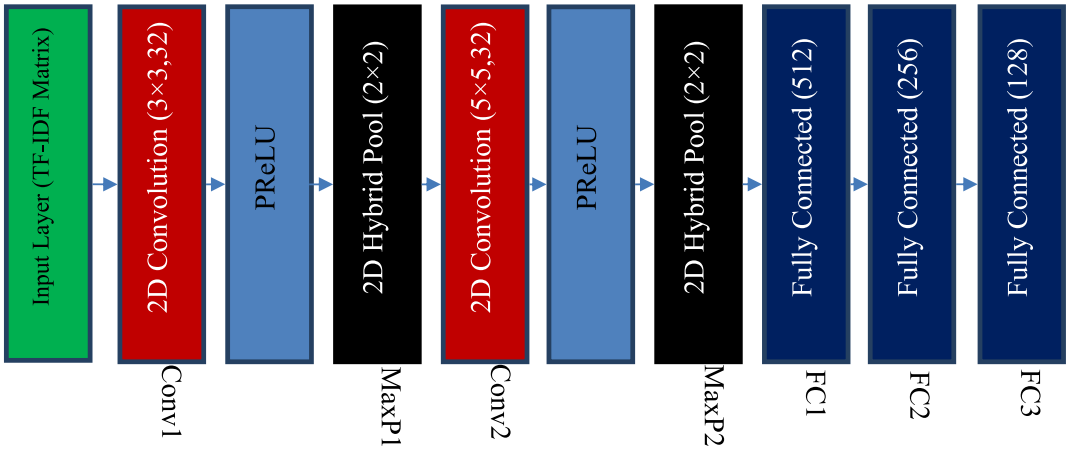


Fig. 2. The proposed CNN structure used in RECSA to reduce the dimensions of TF-IDF features.

can be of great importance; As a result, these solutions cannot be considered suitable [24]. In RECSA, a CNN is used to reduce the dimensions of TF-IDF vectors. Fig. 2 presents the structure of this CNN.

The proposed CNN model is fed with matrix form of TF-IDF vectors. In order to convert each TF-IDF vector into matrix, first all extracted vectors are normalized as follows:

$$N_{TFIDF}(D) = \frac{TFIDF_w(D) - \min_w TFIDF}{\max_w TFIDF - \min_w TFIDF} \tag{2}$$

In equation (2), $TFIDF_w(D)$ refers to weight of term w in document D. Also min and max refer to the minimum and maximum weight of this term in all documents, respectively. After normalizing all TF-IDF vectors, the interval [0,1] is divided into 256 equal intervals (e.g. [0,0.0039), [0.0039,0.0078), …, [0.9961,1]) and each weight value in each normalized vector is placed in one of these intervals based on its value. For example, a word with normalized weight of 0.00533 belongs to interval [0.0039,0.0078). Then each interval is coded as a binary string with length of 8. For example, weights within interval [0,0.0039) are coded as 00000000, and weights within interval [0.0039,0.0078) are coded as 00000001. Finally, the weight values in interval [0.9961,1] are coded as 11,111,111. By applying this process on every weight values of TF-IDF vector, it can be considered as a grayscale pixel. Finally, the resulted values are organized as a matrix which are fed to the proposed CNN.

As shown in Fig. 2, the CNN used in RECSA does not have the necessary layers to classify the samples. In RECSA, CNN is utilized for extracting features from TF-IDF vectors. As a result, the output of the last fully connected layer (FC3), which for each input is a numerical vector with length of 128, describes the characteristics of that text. Each PReLU layer after the convolution layers in this CNN acts as an activation function. Replacing conventional ReLU activations with PReLU layers is effective in solving the problem of vanishing gradients in CNNs. The PReLU layer, uses a small learnable parameter for determining the slope for negative values. Compared to ReLU, the PReLU layer has a nearly zeros added computational cost and at the same time has lower overfitting risk. This operation can be described as Eq. (3) [25]:

$$Y(H) = \begin{cases} H. & if\ X > 0 \\ \alpha H. & else \end{cases} \tag{3}$$

In Eq. (3), H represents the input of the layer and α is the learnable parameter for slope of negative values. Hybrid pooling layers are used in the proposed CNN model for extracting feature maps of convolution layers. Hybrid pooling can improve the model's generality. The max pooling layer may not be efficient for reducing overfitting in small datasets. Also, the average pooling layers may produce sparse feature maps in combination with PReLU layers. To overcome the shortcomings of the conventional pooling layers, hybrid pooling layers are used in the proposed CNN. In the hybrid pooling layer, a learnable parameter such as β is used for the heterogeneous combination of max and average pooling functions. This operation can be described as Eq. (4) [26]:

$$S_{hyb} = p \times S_{avg} + (1-p) \times S_{max} \tag{4}$$

where S_{max} and S_{avg} represent the max and average pooling functions, respectively. The last three layers employed in the proposed CNN are fully connected layers that describe features and reduce its size. The input of the defined CNN is a set of TF-IDF vectors. These vectors are created through the process described at the beginning of this section.

The feature matrix obtained through the convolutional neural network is used to form the similarity matrix between the articles. If the feature vector of each database article is considered as $x_i = \{t_1, \ldots, t_n\}$, using the cosine similarity criterion, the similarity between the titles of two articles x_s and x_t can be calculated as follows [27]:

$$ContentSim(s,t) = \frac{x_s\, x_t'}{\sqrt{(x_s\, x_s')(x_t\, x_t')}} \tag{5}$$

In (5), the similarity value between two vectors is described as a criterion in the interval [0,1]. If the similarity is equal to zero, the two articles x_s and x_t have no similarity, and if the two articles x_s and x_t are completely identical, their similarity is one. After forming the content similarity matrix of the database articles, the analysis of the citation network of the articles and the formation of the communication similarity matrix is done. This process is described below.

3.2. Citation network analysis of articles using link prediction algorithm

In the second step of RECSA, link prediction is used for analyzing the connections in the citation network of articles. This algorithm considers the citation paths between articles with variable lengths in the citation network and makes recommendations based on it. In the following, the function of the link prediction algorithm using FriendLink method is described [28].

If we consider the articles of the citation network in the database as nodes and consider an edge for each citation between them, then paths can be generated between multiple articles. To calculate the similarity of the articles, the similarity matrix is updated by predicting links and considering different lengths for paths between the articles. If n_x and n_y are considered as vertices of the graph, the matrix of paths of lengths two and three can be generated with the condition that there are no duplicate vertices. The higher the number of these paths, the higher the probability of citation. Therefore, assuming a matrix containing paths of length one and two for all pairs of network vertices, the similarity between two articles n_x and n_y can be calculated as follows [28]:

$$CitationSim(n_x, n_y) = \sum_{i=2}^{l} \frac{1}{i-1} \cdot \frac{\left|path^{i}_{n_x,n_y}\right|}{\prod_{j=2}^{i}(N-j)} \quad (6)$$

In (6), l shows the maximum path length considered between two n_x and n_y. Also, N refers to the database size (number of vertices in the graph). There can be no cycle in these paths. $\frac{1}{i-1}$ is a damping coefficient that weights paths with different lengths.

For instance, paths with a length of two are considered with coefficient $\frac{1}{2-1} = 1$; while paths with a length of 3 are considered with coefficient $\frac{1}{3-1} = 0.5$ in the relationship of similarity calculation. In (6), $\left|path^{i}_{n_x,n_y}\right|$ is the number of all acyclic paths with length i between n_x and n_y.

It should be noted that if the query article (for which the recommendation is made) does not have any previous citation connection; a virtual citation link between it and the article that has the most content similarity (in the previous step) will be established. After applying this technique on the articles in the citation network, a similarity matrix is obtained that predicts the probability of citation between the articles. In this way, articles that have the highest values in the q-th row of the similarity matrix are recommended as query q. In the third step of RECSA, the operation of generating the similarity matrix and providing recommendations is performed.

3.3. The combination of content and citation features and recommending

The third step of RECSA is combining content features (first step) and citation features (second step) to generate a total similarity matrix. The total similarity matrix is calculated using the following equation:

$$Similarity(i,j) = \alpha \times CitationSim(i,j) + (1-\alpha) \times ContentSim(i,j) \quad (7)$$

In (7), $ContentSim(i,j)$ is the content similarity between two articles i and j, and is calculated through Eq. (5). Also, $CitationSim(i,j)$ specifies the communication similarity of i and j, which is calculated with Eq. (6). In (7), α is the influence coefficient of content and citation similarities in providing recommendations. This value is in the range of $0 \leq \alpha \leq 1$. If $\alpha < 0.5$, the priority of providing recommendations is based on the content similarity of the articles. and in the case that $\alpha > 0.5$, priority is given to recommending articles that have more citation similarity. It should be noted that if $\alpha = 1$; Articles are recommended only based on the citation network and if $\alpha = 0$; Articles are recommended only based on content similarity.

The article with the highest degree of similarity will be recommended as the output of RECSA. For example, to provide a recommendation as query i, an article will be recommended that its corresponding column has the highest similarity value in the i-th row of the total similarity matrix. According to the solution used in RECSA, the recommended articles have the most similarity with the title of the query both in terms of content and citation connections.

4. Implementation and evaluation of RECSA

In this section, the details of RECSA implementation and the results of its testing are discussed. RECSA for recommending scientific articles has been implemented using MATLAB software. All the tests have been performed using a computer with intelli7, a 1.8Ghz processor and main 8 GB RAM, and Windows 7 64-bit operating system. To test RECSA, a collection of articles collected from the DBLP website has been used. To check the performance of RECSA, the recommendations' correctness is checked. To test RECSA, a collection of practical articles collected from the DBLP website has been used. This database contains 700 documents in XML format, each document contains content information of an article. The citation network of articles is defined as an adjacency matrix. In this matrix, the element of row i and column j stores the value 1 if article i has used article j in its reference list. Otherwise, this value is equal to 0.

Performance evaluation of RECSA is done using Double Cross-Validation. In this evaluation method, the citation network connections in the database are divided into 10 overlapped subsets and the modeling and recommendation process is repeated 10 times. In each experiment, 9 subsets of database citation connections are used to model and form the total similarity matrix, and the remaining subset is used to evaluate the performance of RECSA in recommending articles. During the experiments, a modified approach for creating test datasets was utilized that incorporates temporal information. Specifically, removing links was prioritized based on their temporal information. In other words, connections of each paper in the citation network was prioritized by date and then 10 % of most recent connections belonging to 10 % of citation network nodes were removed. The removed connections were considered as ground truth data. Each time the test is performed, the total similarity matrix is formed by using 90 % of connections (9 parts) and then the recommendations provided by RECSA are generated using this matrix. Then the items recommended by RECSA are compared with the remaining 10 % of actual connections. By comparing the recommendation of RECSA with ground truth connections in citation network, one of the following cases will happen for each recommended item.

- True Positive (TP): RECSA has recommended a paper which actually exists in the remaining 10 % of ground truth connections.
- False Positive (FP): RECSA has recommended a paper which does not exist in the remaining 10 % of ground truth connections.
- False Negative (FN): each paper in the ground truth connections of an article which RECSA could not retrieve them is considered as a FN.

Then, the recall and precision criteria are utilized for evaluating the quality of recommendations and are calculated using Eq. (8) and Eq. (9), respectively:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

To determine the best parameter values in RECSA; First, the effect of changes in parameter α on the performance of RECSA is examined. The results of precision and recall values for changes in this parameter are shown in Fig. 3(a) and (b), respectively. In these two graphs, the horizontal axes represent the values used for the parameter α in each experiment. As shown in Fig. 3, the best performance of RECSA is achieved in the condition that the impact factor is $\alpha = 0.7$.

In the used database, each article "title" is included with its "abstract". For evaluating the effectiveness of the content feature extraction strategy of articles based on the combination of TF-IDF and CNN in the presented approach, the performance of this model in terms of precision and recall in two modes of analyzing title and analyzing the abstract has been compared. Fig. 4(a) and (b) illustrate the precision and recall obtained through this experiment, respectively.

In this experiment, the efficiency of the introduced feature extraction technique has been compared with the doc2vec [7] strategy. Several points can be obtained by examining Fig. 4. First, using a combination of TF-IDF and CNN to extract content features has a higher performance than the doc2vec model. Because, as mentioned, the requirement for accurate description of the content features by the doc2vec strategy is to have longer texts, which cannot be fully fulfilled only by processing the title or abstract of the articles. On the other hand, considering the abstract to extract the content features cannot lead to an increase in the performance of the proposed model. Because some additional information in the abstract of the articles may lead to the addition of non-important information to the content characteristics of the text. While abstracts typically contain more detailed information about a paper, the results show that incorporating them into feature extraction process did not significantly improve performance. This is likely due to the fact that titles are more concise and effective at capturing the key concepts of a paper. Additionally, the inclusion of abstracts introduced additional complexity to the model, which may have affected its overall performance. For these reasons, in these experiments, the recommendation process was performed based on processing the titles of the articles and the abstract part was ignored. These results demonstrate that utilizing TF-IDF, combined with the CNN for analyzing content-based features, leads to at least 0.32 % higher precision compared to other works. Also, In the continuation of this section, the output of RECSA for the optimal parameter $\alpha = 0.7$ has been compared with the content-based article recommender [7], citation network-based scientific article recommender [8], and RNN-based recommender [17]. In the following, the results of these experiments are discussed. In the evaluation of RECSA, the following scenarios have been considered in conducting tests.

a) Changing the number of recommendations
b) Changing the maximum length of the prediction path (parameter L)
c) Changing the dimensions of the database (number of articles)

The results of each of these tests are presented below.

4.1. *Changes in the number of recommendations and performance of RECSA*

In this test, the number of articles recommended by RECSA is changed and the recall and precision criteria are calculated for every state. The number of recommendations means the number of articles that are recommended for citation by the query article and based on the similarity matrix obtained by RECSA, to each article under test. The recommended number of changes ranges from one to five.

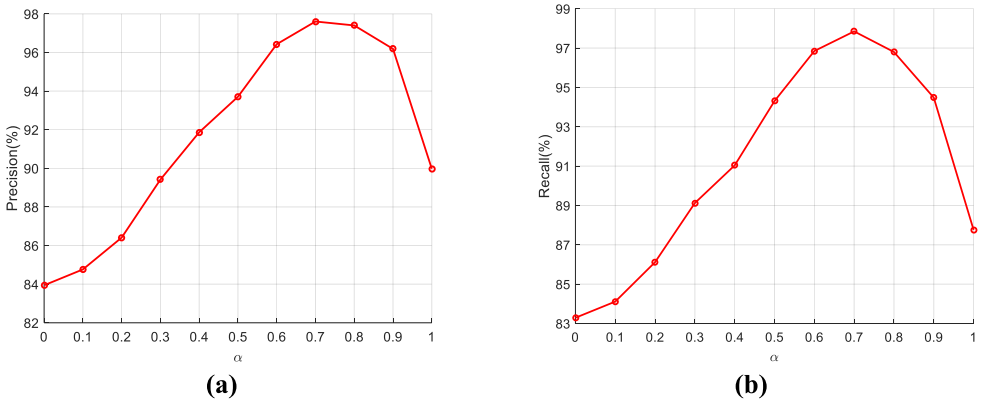


Fig. 3. Graphs of changes (a) precision, (b) recall for changes in parameter α in the proposed method.

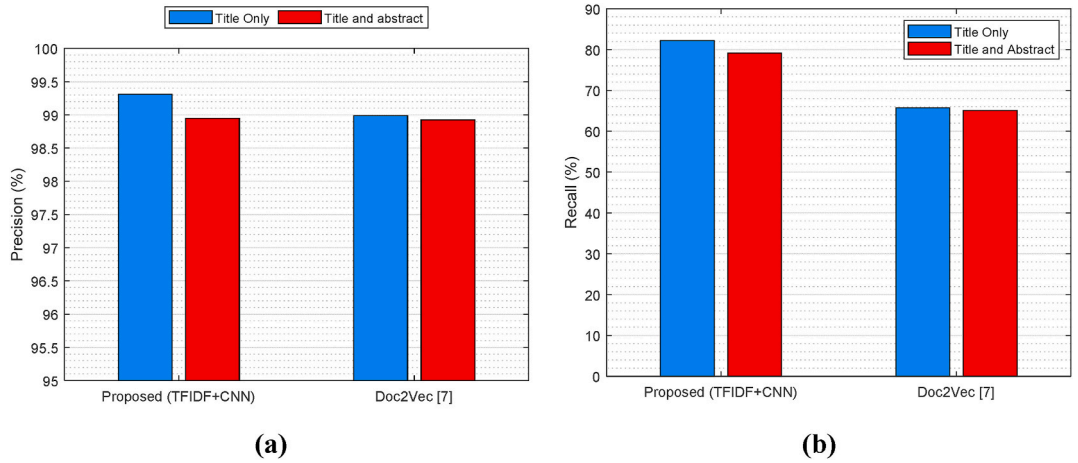


Fig. 4. Values of (a) precision, (b) recall for different content feature extraction methods in two cases of analyzing title and analyzing abstract of the articles.

The default value for the prediction path length parameter in this experiment is considered equal to 3. In Figs. 5 and 6, the graphs of the changes in the precision and recall criteria for the number of recommendations in RECSA are displayed, respectively.

In both of these graphs, the results of RECSA are compared with content-based and citation network-based methods. These results show that better results can be achieved by using the method of combining content and citation information in RECSA. Based on these results, the best level of precision and recall are obtained if $\alpha = 0.7$. On the other hand, in Fig. 5, the precision has a downward trend, and in Fig. 6, the recall has an upward trend. This situation specifies that several recommendations would be appropriate to achieve better precision. The decrease in the precision value along with the increase in the number of recommendations shows that RECSA often prioritizes the appropriate recommendations. This is the characteristic that is expected from a suitable recommender algorithm. Thus, Fig. 6 shows that in RECSA, the most suitable articles are recommended first, and other recommendations are less important (in other words, with the increase in the recommendations, the probability of being included in the set FP increases and a lower proportion of the recommendations will be correct). Also, with the increase in the number of recommendations of RECSA in Fig. 6, the recall criterion increases. For 5 recommendations, the recall criterion is 100 % and all correct citation connections are recommended by the RECSA. This feature shows that with increasing the recommended items, a higher proportion of existing citation connections are correctly extracted. Accordingly, the performance of RECSA is superior to other algorithms in all cases. This improvement can be attributed to the combined solution used in RECSA.

4.2. Changes in the maximum length of the prediction path and RECSA performance evaluation

In this experiment, the prediction path length parameter (L) is changed and the recall and precision criteria are calculated for each state. In this experiment, the number of recommendations for RECSA is 2. Also, considering that the content-based recommendation method does not support the path length parameter, in this experiment, its comparison with RECSA is omitted. The range of changes in

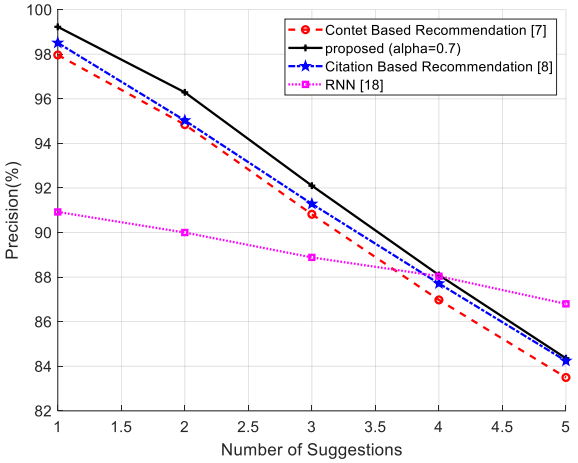


Fig. 5. Changes in precision criteria for different numbers of recommendations.

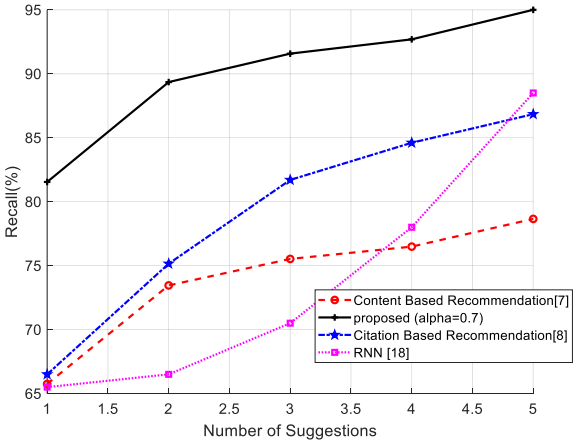


Fig. 6. Changes in the recall criteria for different number of recommendations.

the length of the prediction path is considered to be 2 to 5. In Figs. 7 and 8, respectively, the graphs of changing the length of the prediction path and its effect on the precision and recall are displayed.

Comparing the performance of RCSA with the RS based on the citation network in Figs. 7 and 8 shows that a better performance can be achieved by using RECSA. Also, these results indicate that the value of 3 for the path length achieves more accuracy. Because in the graph used for evaluation, the average shortest distance (ASD) in the graph is 3.98 and the path length parameter with value 3 can check the appropriate paths in the citation network of the articles. A value less than 3 for the path length causes some of the citation links of the articles to be ignored due to the shortness of the length parameter, and as a result, the FN rate increases. On the other hand, a value greater than 3 for the length of the path causes incorrect and unnecessary connections to be considered in the citation network, and as a result, the FP rate increases by increasing the probability of inappropriate recommendations.

4.3. Changes in database dimensions and performance evaluation of RECSA

In this section, the influence of the dimensions of the reference network on the performance of RECSA is investigated. The goal of this experiment is to investigate the efficiency of RECSA due to changes in the dimensions of the problem. In this scenario, the number of recommendations provided by RECSA is considered equal to two. Also (based on the results obtained from the previous experiment), the parameter of the predicted path length is set to three. Then, keeping these parameters constant, the number of dataset articles is changed in interval 300 to 700, and the effect of dimension changes on the performance of RECSA is investigated. Figs. 9 and 10 respectively show the graph of changes in precision and recall criteria for changes in the dimensions of the citation network in RECSA.

As the results of this test show in Figs. 9 and 10, the precision criterion increases, and the recall criterion decreases. By increasing the dimensions of the citation network, more information about the connection pattern between articles based on citations is available

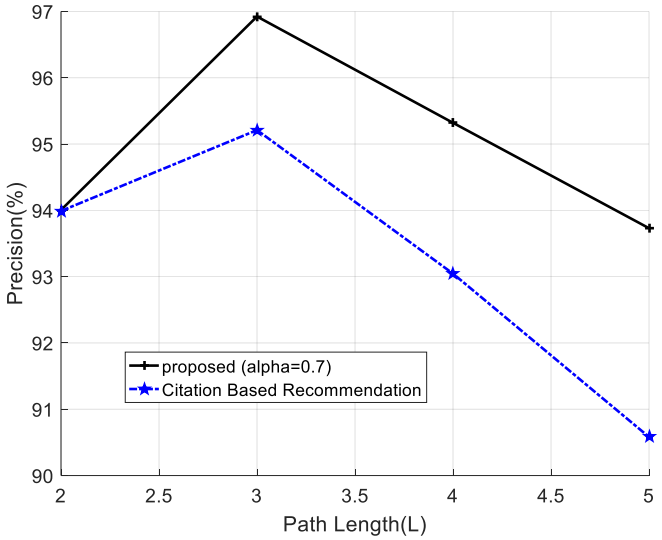


Fig. 7. Precision criterion for different lengths of the prediction path.

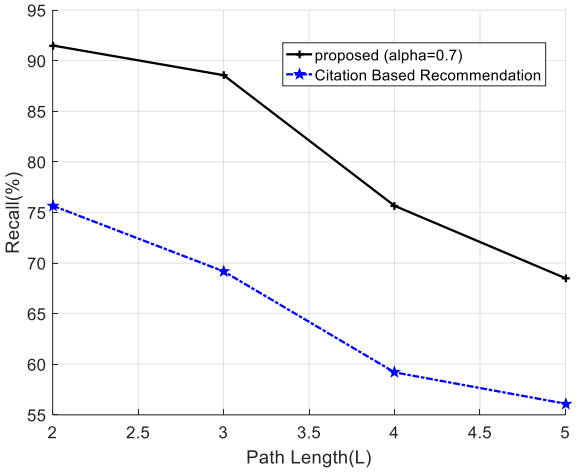


Fig. 8. Recovery criterion for different lengths of the prediction path.

to the RS. Based on this information, the RS obtains more comprehensive data of the problem and can provide recommendations with a higher probability of accuracy. For this reason, as the dimensions of the problem increase, the precision criterion increases. Also, with the increase in the number of articles and the size of the citation network, a greater proportion of citation connections is removed in the test phase, and the RS should reveal more hidden citations through recommendations. In other words, the number of test samples increases with the increase in the dimensions of the citation network, while the number of recommendations is fixed and equal to 2. As a result, the recall criterion has a downward trend in this situation.

RECSA's innovative approach of combining content analysis and citation network analysis offers several notable advantages over existing recommender systems for scientific articles. Firstly, RECSA's utilization of TF-IDF and CNN for content-based analysis enables a more comprehensive and nuanced understanding of the semantic content of scientific articles, going beyond mere title matching. This enhanced understanding leads to more accurate and relevant recommendations. Secondly, RECSA's integration of citation network analysis provides a contextual dimension to the recommendations. By analyzing the connections between articles based on citations, RECSA captures the relationships between research ideas and trends, suggesting articles that are not only thematically similar but also intellectually connected. This contextual understanding enhances the relevance and usefulness of the recommendations. Finally, RECSA's weighting scheme, which combines content-based and citation network similarities, allows for a flexible and adaptive recommendation process. The influence coefficient parameter enables the system to prioritize one type of similarity over the other, depending on the user's preferences and search context. This adaptability makes RECSA a versatile and user-friendly recommender system.

4.4. Complexity of RECSA

In order to investigate RECSA more precisely, it is necessary to study the computational complexity of this model. This model uses two parallel components based on TFIDF and link prediction to provide recommendations. The computational complexity of TF-IDF is $O(n \times L \times log\,(n \times L))$, where n is the total number of sequences in a dataset, and L is the average length of sequences in a dataset. On the other hand, the computational complexity of the link prediction model is equal to $O(d \times n^l)$ where d is the number of documents and l is the path length. This complexity is higher compared to content-based models such as [7,8], but it should be noted that this complexity is only related to the process of constructing the recommendation model, and in the recommendation phase (test phase), this complexity will be equal to O(n). Therefore, due to the superior efficiency of the presented hybrid model, this increase in computational complexity can be omitted.

According to the experiments performed in this section, RECSA outperforms existing recommender systems. This superiority can be attributed to the following features: Firstly, RECSA's combination of content analysis and citation network analysis provides a more holistic understanding of scientific articles, enabling more accurate and relevant recommendations. Previous methods that rely solely on title matching or content-based analysis often fail to capture the nuanced connections between articles, leading to irrelevant or inaccurate suggestions. Secondly, RECSA's weighting scheme, which considers both content and citation similarities, allows for a more personalized recommendation experience. This adaptive approach caters to individual preferences and search needs, ensuring that recommendations are tailored to the specific interests of each user. RECSA's evaluation results, which demonstrate significant improvements in precision compared to compared methods, provide strong evidence of its effectiveness. The precision of RECSA's first suggestion reaches 99.01 %, significantly higher than existing methods. These results highlight the superior performance of RECSA in recommending relevant and accurate scientific articles.

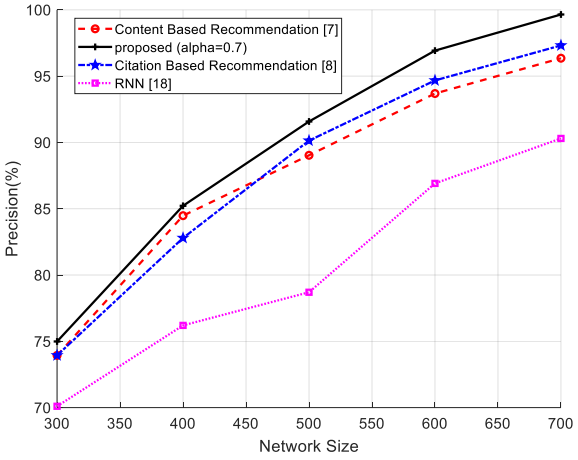


Fig. 9. The precision criterion for different dimensions of the citation network.

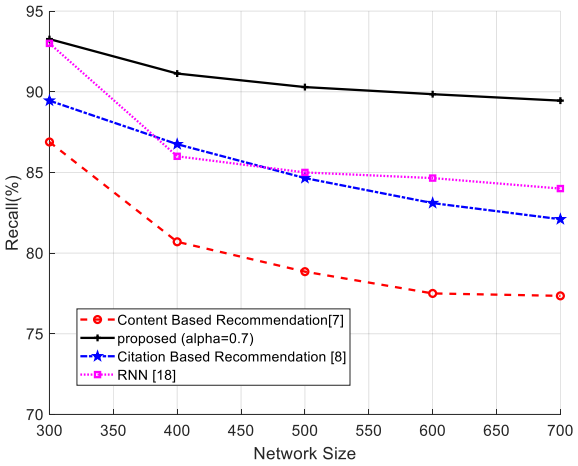


Fig. 10. The recall criterion for different dimensions of the citation network.

5. Conclusion

In this article, an RS for scientific articles based on the combination of content and citation features of the articles was presented. To test the proposed method (RECSA), a collection of practical articles collected from the DBLP website has been used. To evaluate the quality of the recommendations, the output of RECSA was evaluated using recall and precision criteria, and the results were compared with previous recommender algorithms. The results showed that RECSA has better performance than content-based and citation network-based methods, and by using RECSA in this article, the recommendation action can be performed with higher precision and recall.

One of the limitations of RECSA is its higher computational complexity than other methods. Because in RECSA, content information and citation network are used simultaneously, and this requires more calculations in the process of modeling. Although this difference is noticeable only in the recommender modeling phase; this time difference can be reduced by using parallel processing techniques. Deploying RECSA as an online recommender system in real-world scenarios is a plan to continue the research in future works. While the dataset used in this research provided a valuable foundation for initial testing, we acknowledge the potential benefits of investigating RECSA's performance on even larger and more diverse datasets. Future work will involve exploring the scalability of our approach by applying RECSA to significantly larger datasets encompassing a wider range of scientific disciplines.

Data availability

All data generated or analysed during this study are included in this published article.

CRediT authorship contribution statement

Weijuan Li: Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] B. Cao, J. Zhao, Z. Lv, P. Yang, Diversified personalized recommendation optimization based on mobile data, IEEE Trans. Intell. Transport. Syst. 22 (4) (2021) 2133–2139.
[2] Y. Xu, E. Wang, Y. Yang, Y. Chang, A unified collaborative representation learning for neural-network based recommender systems, IEEE Trans. Knowl. Data Eng. 34 (11) (2022) 5126–5139.
[3] X. Bai, M. Wang, I. Lee, Z. Yang, X. Kong, F. Xia, Scientific paper recommendation: a survey, IEEE Access 7 (2019) 9324–9339.
[4] A.T.M. Aymen, S. Imène, Scientific paper recommender systems: a review, in: International Conference on Artificial Intelligence in Renewable Energetic Systems, Springer, Cham, 2021, November, pp. 896–906.
[5] U. Javed, K. Shaukat, I.A. Hameed, F. Iqbal, T.M. Alam, S. Luo, A review of content-based and context-based recommendation systems, International Journal of Emerging Technologies in Learning (iJET) 16 (3) (2021) 274–306.
[6] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D.S. Weld, Specter: Document-Level Representation Learning Using Citation-Informed Transformers, 2020 *arXiv preprint arXiv:2004.07180*.
[7] B.G. Patra, V. Maroufy, B. Soltanalizadeh, N. Deng, W.J. Zheng, K. Roberts, H. Wu, A content-based literature recommendation system for datasets to improve data reusability–a case study on gene expression omnibus (geo) datasets, J. Biomed. Inf. 104 (2020) 103399.
[8] N. Sakib, R.B. Ahmad, K. Haruna, A collaborative approach toward scientific paper recommendation using citation context, IEEE Access 8 (2020) 51246–51255.
[9] J. Son, S.B. Kim, Academic paper recommender system using multilevel simultaneous citation networks, Decis. Support Syst. 105 (2018) 24–33.
[10] W. Huang, Z. Wu, L. Chen, P. Mitra, C.L. Giles, A neural probabilistic model for context based citation recommendation, in: AAAI, 2015, January, pp. 2404–2410.
[11] A.S. Raamkumar, S. Foo, N. Pang, Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems, Inf. Process. Manag. 53 (3) (2017) 577–594.
[12] H. Wang, N. Wang, D.Y. Yeung, Collaborative deep learning for recommender systems, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, August, pp. 1235–1244.
[13] J.D. West, I. Wesley-Smith, C.T. Bergstrom, A recommendation system based on hierarchical clustering of an article-level citation network, IEEE Transactions on Big Data 2 (2) (2016) 113–123.
[14] W. Huang, Z. Wu, P. Mitra, C.L. Giles, Refseer: a citation recommendation system, in: Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on, IEEE, 2014, September, pp. 371–374.
[15] X. Zhu, P. Turney, D. Lemire, A. Vellino, Measuring academic influence: not all citations are equal, Journal of the Association for Information Science and Technology 66 (2) (2015) 408–427.
[16] H.C. Wang, J.W. Cheng, C.T. Yang, SentCite: a sentence-level citation recommender based on the salient similarity among multiple segments, Scientometrics 127 (5) (2022) 2521–2546.
[17] T.N. Dinh, P. Pham, G.L. Nguyen, B. Vo, Enhanced context-aware citation recommendation with auxiliary textual information based on an auto-encoding mechanism, Appl. Intell. 53 (14) (2023) 17381–17390.
[18] K. Velkumar, Deep learning-assisted citation recommendation system using multi-cell RNN approach, Tuijin Jishu/Journal of Propulsion Technology 44 (2) (2023) 170–181.
[19] A.M. Nair, N.K. Paul, J.P. George, A citation recommendation system using deep reinforcement learning, in: Mobile Computing and Sustainable Informatics: Proceedings of ICMCSI 2021, Springer, Singapore, 2022, pp. 423–433.
[20] C. Pornprasit, X. Liu, P. Kiattipadungkul, N. Kertkeidkachorn, K.S. Kim, T. Noraset, S. Tuarob, Enhancing citation recommendation using citation network embedding, Scientometrics 127 (1) (2022) 1–32.
[21] Ö. Kart, A. Mestiashvili, K. Lachmann, R. Kwasnicki, M. Schroeder, Emati: a recommender system for biomedical literature based on supervised learning, Database 2022 (2022) 1–10.
[22] A.G. Jivani, A comparative study of stemming algorithms, Int. J. Comp. Tech. Appl 2 (6) (2011) 1930–1938.
[23] G. Sidorov, Vector space model for texts and the tf-idf measure, in: Syntactic N-Grams in Computational Linguistics, Springer, Cham, 2019, pp. 11–15.
[24] H. Liang, X. Sun, Y. Sun, Y. Gao, Text feature extraction based on deep learning: a review, EURASIP J. Wirel. Commun. Netw. 2017 (1) (2017) 1–12.
[25] J. Crnjanski, M. Krstić, A. Totović, N. Pleros, D. Gvozdić, Adaptive sigmoid-like and PReLU activation functions for all-optical perceptron, Opt Lett. 46 (9) (2021) 2003–2006.
[26] Z. Tong, G. Tanaka, Hybrid pooling for enhancement of generalization ability in deep convolutional neural networks, Neurocomputing 333 (2019) 76–85.
[27] K. Orkphol, W. Yang, Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet, Future Internet 11 (5) (2019) 1–16.
[28] A. Papadimitriou, P. Symeonidis, Y. Manolopoulos, Fast and accurate link prediction in social networking systems, J. Syst. Software 85 (9) (2012) 2119–2132.