

Srvycite: a hybrid scientific article recommendation system

Sivasankari R, J. Dhilipan

Department of Computer Applications (MCA), SRM Institute of Science and Technology-Ramapuram Campus, Chennai, India

Article Info

Article history:

Received May 7, 2024

Revised Aug 29, 2024

Accepted Sep 22, 2024

Keywords:

BERT

Dimensional reduction

Scientific article

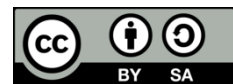
recommendation

Survey citation network citation
network

ABSTRACT

A recommendation system is becoming part of every work done today to reduce the effort of work done by the users in searching for items in need by recommending new items that may be useful. This theme has also been used in research article recommendation systems for recommending articles of interest to researchers from a bulk of digital research documents spread across different databases on the internet. To ease the task of this article recommendation process, we have proposed a novel approach, Srvycite, by utilizing the survey article citation network along with the original research article network. The purpose of utilizing the survey article citation network is to detect the most influential articles that are considered to be important by other researchers in the same field. The Srvycite approach utilizes the text and meta features of articles to recommend papers. To preprocess the text features utilized, we have employed Word2Vec and bidirectional encoder representations from transformers (BERT) for vectorization. Then citation graph and survey citation graphs are generated to find the most influential nodes. The weighted text similarity score is finally computed by combining the cited by values and the text similarity score from the citation and survey citation graph to list articles as recommendations for the user. This system is proven to increase the accuracy of the article recommendation by 3.8 and 2.1 in the case of the precision and recall measures for performance evaluation.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Sivasankari R

Department of Computer Applications (MCA)

SRM Institute of Science and Technology-Ramapuram Campus

Chennai, Tamil Nadu, India

Email: manjurmoni@gmail.com

1. INTRODUCTION

A scientific article recommendation system aims to recommend research papers for researchers. With a rapid increase in research activities, it takes time for researchers to gather the data needed for their research. The data needed also includes previous research done in the preferred area. This is required for the researchers to stay updated and to focus on the required area and part of the research. Not only that, but with the help of previously published articles, researchers can easily identify the concept and fulfil the requirement to enhance the idea further. In today's digitally connected world, it is effortless to gather needed articles through a digital support system. Yet the need for the recommendation system is to identify the most relevant articles without spending much time. Scientific recommendation systems help researchers by identifying the most relevant articles for them with the help of a search query. These recommendation systems can be categorized into content-based systems, collaborative systems, graph-based systems, and hybrid recommendation systems. The content-based filtering methods (CBF) is a classic recommendation approach that provides recommendations by discovering the association between user modelling and textual features in the paper. The advantage of CBF allows user-based personalization and requires less classification work.

CBF requires addressing problems like high computing power, low serendipity, and overspecialization, and also ignores the quality and popularity of items. It is dependent on access to the item's features, making it unsuitable for many users [1].

The idea of collaborative filtering systems (CF) is to propose a paper to users who share common interests. It implies that if two users A and B have similar interests, the articles viewed by user A can be recommended to user B. CF is content-independent, considers real-world quality ratings, and makes serendipitous recommendations. CF suffers from the "cold start" problem, which occurs when the user or the item is new and no proper rating information is supplied by the users [2], [3]. In graph-based recommendation systems (GB) the features of an articles such as textual content, bibliography, authors meta data, and article metadata can be used to construct the graph. After building a graph from the required features recommendations are computed based on the graph algorithms. GB systems can handle data sparsity but require improvements to handle complex queries. To overcome the challenges of the above-mentioned system, two or more types can be combined to produce a most effective and powerful recommendation. Such systems are called hybrid recommendation systems, and they have the capability to address two or more issues. Existing recommendation algorithms illustrate the viability of recommending multiple types of features and relationships from scientific articles, such as historical behaviour and records, research content, citation networks, and so on [4]. An efficient scientific paper recommender system is needed to generate high-quality recommendations for novice researchers who lack knowledge in the research area [2], [5]. Scholars have conducted research on recommendation systems, but there is a lack of exploration of the research status and future trends [6]. After exploring several research articles [7]-[9], it is concluded that the state-of-the-art articles or review articles in the field of study are completely ignored and not considered as important to compute the recommendation. The review articles are primary because they synthesise past works, summarise the idea of the work, understand well the problem that has been discussed, and provide a potential way to think about the solution to the problem under discussion. Literature reviews are included in all research projects, regardless of the discipline. A literature review is a research article where the author typically begins by summarising prior research to map and evaluate the research area, inspire the study's purpose, and support the research question and hypotheses. A good literature review would discuss the important works produced over time in the selected field of study as well as the current trends in the field of research. This results in a list of citations that are regarded as significant and highly pertinent to the study. These citations may be used to recommend a relevant research article. If a citation network for such a survey paper is built, it may result in a network that comprises significant articles from recent times that can be used for the recommendation systems. This work suggests a Srvycite, a hybrid scientific article recommendation system that builds the survey citation network (SCN) along with the CBF and citation networks (CN) to recommend relevant articles.

The objective of this study is summarized as follows:

- The main objective of this paper is to improve the accuracy of recommendations by using a hybrid technique that integrates content-based filtering and citation graphs. To prove this objective, we have introduced a novel citation graph called the SCN graph, which is constructed using all types of survey articles from the dataset with their citations.
- Apart from the SCN, we also employed the CN graph to improve the accuracy of our results.
- For clustering, we used bidirectional encoder representations from transformers (BERT) in text embedding along cosine similarity and the K-means algorithm.
- Here, the data we used is high-dimensional text data that suffers from memory and time complexity. To address this problem, LASSO or L1 regularization is used.
- Furthermore, the experiments with the proposed approach show that the recommendation system now works better and gives more accurate results compared to the other methods presented in this study.

To summarize the contributions of the proposed system, the first step is to cluster the entire data into two clusters: survey article cluster and original article clusters. To cluster text data, this system follows several steps, such as text preprocessing, text embedding, dimension reduction, similarity calculation, and clustering. Followed by clustering is graph generation, where two different graphs (SCN and CN) are constructed by matching the search query of the user. Finally, the results of both graphs are combined to generate the list of final articles relevant to the search query. Here, the proposed system is hybrid, where the collaboration-filtering method is used to find important nodes in the network and content-filtering is used in the later process to compute the text similarity score along with citation values computed in the graph to generate the final recommendations. Figure 1 explains the architecture of the proposed system. The rest of the paper is structured as follows: section 2 discusses several relevant works that are related to this work. Section 3 describes the proposed method in detail. Section 4 presents the outcomes of the experiments and evaluations. Section 5 concludes the article and addresses future work.

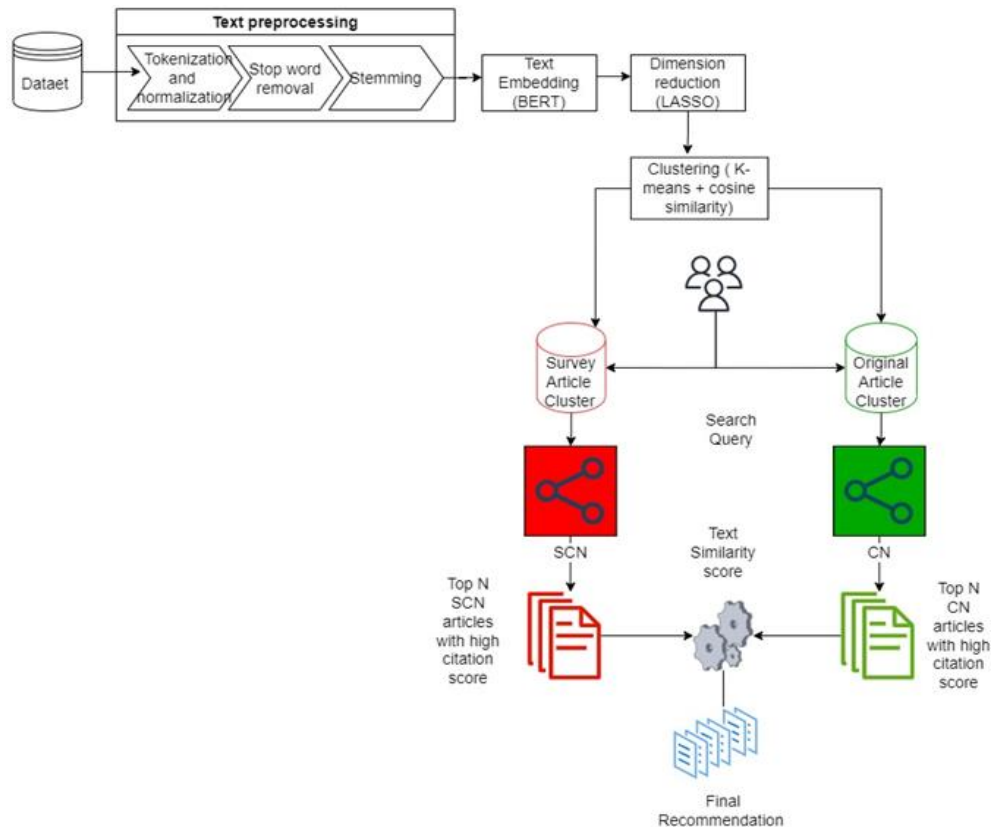


Figure 1. Srvycite architecture

2. RELATED WORK

Search engines for academic literature such as Google Scholar, Semantic Scholar, and Science Gate are used to compile the articles required for the literature review. "Research paper recommendation", "research article recommendation", "scientific article recommendation", and "scientific paper recommendation" are used as search terms. Nearly 2,429 items over the previous 10 years were identified to fit the search, and the most pertinent and important articles from this group were compiled and used for this work. Figure 2 articles publication count per year represents the yearly count of articles published from 2014 to 2023. A hybrid recommendation system was created in the work of [10]. Three separate approaches demographic-based, relationship network-based, and content-based make up this hybrid recommendation system. In this study, the authors tested the recommendation system using a data set from ResearchGate. Latent factors were integrated with content-based filtering to propose articles to researchers in the recent work by [11]. The first stage in creating an appropriate subject for the inquiry is topic modelling. Latent dirichlet allocation (LDA) and non-negative matrix factorization (NMF), two techniques, are employed in topic modelling. A list of articles that are similar is then suggested based on the semantic. Hadhiatma *et al.* [12] put forth the PPR_TC framework, which uses a modified PageRank algorithm to facilitate the selection of multi-topic communities. The initial section of the framework begins with the extraction of significant characteristics from the academic citation network, which is followed by the identification of multi-topic communities that are relevant to the query. Finally, ranking is calculated by a modified PageRank algorithm to suggest relevant articles that fit the communities. A deep learning-based hybrid article recommendation system was proposed by the academics [13]. This approach groups papers from several disciplines, including computer science, economics, and medicine, by combining document similarity, hierarchical clustering, and keyword extraction. Additionally, it has the ability to provide consumers with items that are quite similar in terms of semantics. The researchers employed the users' browsing and personal information in the study provided by [14] to make article recommendations. The density-based spatial clustering of applications with noise (DBSCAN) clustering method is used to group pertinent articles, followed by the firefly algorithm for data optimisation and the genetic algorithm for data prediction. Finally, users are recommended articles using a recommendatory system based on participatory filtering. Chaudhuri *et al.* [15] introduced systematic hidden attribute-based recommendation engine (SHARE), a multi-criteria-based personalized research paper

recommendation system that used many aspects to discover important insights to uniquely represent publications. Authors also takes into account the desire of user in order to capture the dynamic notion of users. SHARE used hybrid ranking method with the multi-criteria decision analysis (MCDA) and support vector machine (SVM) rank algorithms. In this case, MCDA is used to rank papers by analyzing multiple characteristics, whereas SVM rank is used to rate articles depending on user preferences. SHARE employed a rank aggregation method to generate customized list of rankings.

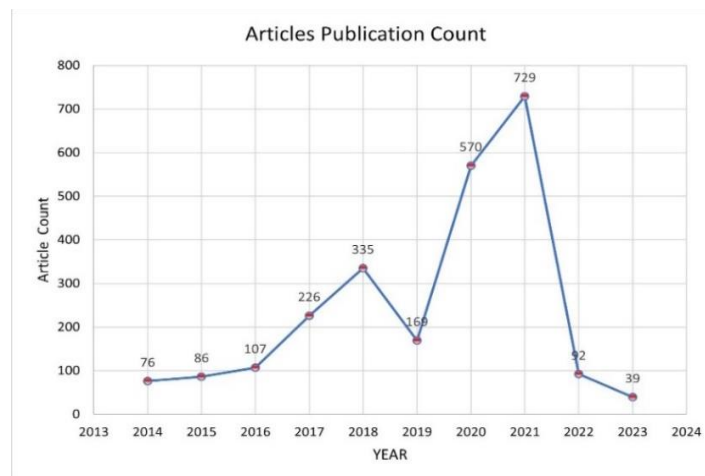


Figure 2. Articles publication count per year

The HNTA approach for academic article recommendation was put out by [16]. This technique suggests a strategy by fusing temporal features from the SCHOLAT dataset with heterogeneous networks. Researchers in this case combined user research interests with content information from the papers. In order to determine commonalities, the users' research interests are further divided into those that are immediate and those that are ongoing. In order to propose papers to scholars, all the estimated variables are weighted. A graphical framework for citation relationships was created by [17] as a hybrid recommendation model that incorporates citation and content-based techniques. Weights are allocated for the citation of the paper based on the section in which it is applied. The content-based similarity algorithm is paired with this citation weight to get the final list of recommended papers. Liu *et al.* [18] build an undirected paper citation graph from users' queries. They modelled the problem as a Steiner tree problem, constructed a tree, and recommended the papers. The authors constructed the graph by seeking all the keywords from the users' query, and then articles that match each word are retrieved. After retrieving an individual list of related papers for each keyword, a relation between all the keyword sets is formed as a subgraph based on the similarity between the papers. Based on the final Steiner tree generated after merging selected trees, recommendations are made. Tao *et al.* [19] presented an article to offer a list of references for further study. Word2vec is also used for topic vectorization, while LDA is used to extract keywords from subjects and papers. Paper vectorization is represented by Doc2Vec. The degree of similarity between topics and papers is determined using these two vectors. A list of N publications is suggested to the user as a resource based on the final similarity values. By comparing the similarity of two articles, Du *et al.* [20] created a heterogeneous information network. For network construction, factors such as article content, citations, paper area, and co-authorship are taken into account. The authors create word sequences using natural language models, and they employ random walks to move over the network.

Ma *et al.* [21], built a heterogeneous bibliographic network. The suggested approach builds the network by employing the weights of topological characteristics, meta-paths, and meta-graphs together with them. The W-rank method is proposed by the researchers in [22]. The W-rank algorithm ranks articles by weighting the linkages in the citation and authorship networks. To calculate the citation, network- and semantic-based algorithms are coupled to propose the most appropriate citation. Kanakia *et al.* [23] created a hybrid recommender system that combines both content-based (CB) and co-citation-based (CCB) suggestions to generate a final static list of recommendations for each article in the MAG. This method uses a weighted mixed hybridization with a customizable weighting function to directly contrast each recommendation pair from either list and merge both lists to get a final paper recommendation list for practically every paper in the graph. Son and Kim [24] presented an academic paper recommendation system

that combines citation analysis and network analysis. It is based on multilevel citation networks that compare all indirectly linked papers to the paper of interest to inspect the structural and semantic relationships among them. The authors considered the mutual relationships among the papers in a broad network beyond a single level and evaluated the significance of each paper through certain centrality measures. It was concluded to outperform other existing models based on user satisfaction data. Haruna *et al.* [25] explained how a scientific paper dataset is converted into an academic citation network by converting it into an academic citation network, identifying communities within that network, and identifying multi-topic communities and papers by using the LDA topic model, which is a topic model that is based on the LDA concept. They employed this technique to select and combine the most pertinent n -topic communities from a set of multi-topic communities. Modified PageRank, PPR_TC, is employed to create a bias probability vector and transition probability matrix for topic queries and multi-topic communities. In the work [26], a research paper recommendation system that combines a cross-crawling-based document gathering technique with a paper connection analyzer. Traditional keyword analysis and citation networking analysis are combined in the suggested paper relationship analyzer. A cross-crawling engine was created to collect relevant articles from various digital libraries, while a paper citation network analyzer was built to identify the relevance of papers based on textual and paper citation network analysis. Sharma *et al.* [27] developed a method for encoding each document in terms of its concept or semantics. They accomplished this by building a distributed representation in high-dimensional space. This individual representation is used to discover items that meet the needs of the user. The entire recommendation procedure is divided into two stages: vector creation for candidate papers and a recommendation algorithm.

Dhanda and Verma [28] suggested a personalized research article recommendation system that uses an incremental research paper repository and the incremental high-utility item set mining approach efficient incremental high-utility itemset mining algorithm (EIHI). The technique performs the suggestion process in two steps. First, articles are selected for the researcher's area of interest using the probabilistic latent semantic analysis (PLSA) algorithm, which categorizes research papers into ' k ' clusters, and then valuable publications are recommended to the researcher using EIHI to mine dynamic datasets. Xia *et al.* [29] are the authors of the common author relation-based recommendation technique (CARE). Researchers frequently seek out works by the same authors, but not all researchers have author-based search habits. These two facts form the basis of the CARE technique. It has a researcher selection module and a graph-based article ranking module. The first module gathers data from prior preferences and suggests that academics with author-based search habits are its target audience. To create article rating lists, the second module uses a graph-based random walk technique. Amami *et al.* [30] suggested an LDA-based approach for recommending scholarly articles. According to this strategy, the content that the researcher self-generated should serve as the foundation for the creation of the researcher profile because it reveals the topics in which the researcher is most interested as well as the specialized language that researchers employ when they write articles. The themes the LDA algorithm extracted from the researcher's earlier publications are then combined to create the researcher profile. An official language model representation of the unread article is provided, and it is compared with each topic defining the researcher profile in order to determine whether the item would be of interest to the researcher. Chakraborty *et al.* [31] developed the faceted recommendation system for scientific articles (FeRoSA). FeRoSA is designed to properly cater to the end user's appropriate knowledge context by not only recommending relevant scientific papers but also organizing the recommendations into facets given a query paper. This approach aims to simulate the traversal mechanism of a user starting from a known article using a principled framework of random walks with restarts. To systematically generate the most relevant results, the model considers both the citation links and the content data. FeRoSA divides the suggestions into four naturally occurring categories, including background, alternative approaches, methods, and comparison.

3. METHOD

3.1. Problem formulation

The task of the recommendation system is to predict the list of articles that match the user query and to recommend them to the researchers. Let us consider an article $a \in A$, where A is the set of all articles and the search query of a researcher is represented as SQ . $SQ_r = \{a_1, a_2, \dots, a_n\}$ represents the set of all articles this system recommends as similar articles by relating the SQ of a researcher r . To recommend the articles in SQ_r , this system requires two different graph networks: the SCN and the CN. An article $sc \in SCN$, where SCN is the set of survey articles, and $cn \in CN$, where CN is the set of normal articles other than surveys. An article sc is said to be an element of SCN if and only if $sc \in A$; similarly, an article cn is said to be an element of CN if and only if $cn \in A$. SQ_r is computed by combining a list of the top recommended articles from SCN and CN networks.

3.2. Clustering

In order to group text documents into meaningful clusters based on their content and similarity, text document clustering is used. Due to its assistance in organizing, summarizing, and gathering insights from huge amounts of textual data, it plays a vital role in many natural language processing (NLP) and information retrieval tasks. The main objective of clustering in a recommendation system is to put similar documents together so that they can share common themes, topics, or subject matter for further processing of content. The entire process of clustering can be broken down into five steps: preprocessing, text representation or embeddings, dimensionality reduction, similarity measure and finally clustering.

3.2.1. Text preprocessing

Text documents distributed via digital media are unprocessed and cannot be processed immediately by intelligent systems. The text document containing raw text data can take various forms, such as plain text, HTML documents, JSON files, and so on. This text data also includes some special characters, words, letters, and numbers that are irrelevant to the context or have no meaning. Text processing is required to convert it into a form that this intelligent system can understand. Text processing is a process where unwanted data from raw data is removed to proceed with further processing.

Text processing can be subdivided into tokenization, normalization, stop word removal and stemming. Tokenization and normalization: tokenization is a fundamental text processing technique that entails converting raw text from text files or digital media into meaningful units known as tokens. This process divides a document's text into discrete units by inserting white spaces or distinct markers, allowing each unit to correspond to a different word in the text. During normalization, text elements such as code tags, numbers, symbols, special character tags, URLs, and characters that delineate sections within entire documents, such as underscore (_), at (@), percentage (%), ampersand (&) and so on, are removed. Normalization also eliminates symbols with ambiguous meanings, non-English words, and punctuations to preprocess text. Stop word removal: stop words are primarily used to provide grammatical structure in text documents. Stop words are tokens that are characterized as having less importance, and they tend to appear frequently in the document. The removal of these stop words has a significant impact on the system's storage and processing. The, this, is, are, he, these, and that are some examples of stop words in English. Stemming: the process of removing all prefixes and suffixes in order to obtain root words is known as stemming. It is not necessary to apply stemming to all generated tokens. Only a few tokens, such as playing, singing, and processing, require stemming to obtain the corresponding root words play, sing, and process.

3.2.2. Text embeddings

Due to the complex nature of text tokens, it is still impossible for machine learning (ML) algorithms to understand them. To simplify this problem, text embeddings are used, which is the process of representing complex tokens into a form that is simple to understand by the ML algorithms. The converted form of these complex text tokens into a simplified representation is called a vector. Each word or text token is assigned individual numerical vectors. Traditional methods, neural networks with word embeddings, and deep learning methods are among the similarity measures used to compute the vectors. Traditional methods are classified as either simple or complex. One-hot encoding, bag of words (BoW), term frequency-inverse document frequency (TF-IDF), and word count vectors are examples of simple traditional embedding methods. Vectors can be measured using complex methods such as latent semantic analysis (LSA), LDA, and vector space model (VSM). Word2Vec, Doc2Vec, and Glove are some similarity measures that use neural networks with word embeddings. Deep learning methods such BERT, long-short term memory (LSTM), bidirectional long short-term memory (Bi-LSTM), and convolutional neural network (CNN) can also be used. We have selected BERT for our proposed model to generate text embedding. This method involves using computers to train machine learning models with text data. BERT, which stands for Bidirectional Encoder Representations from Transformers, gains an understanding of words in context during training. Unlike some modern language representation theories, BERT's main objective is to create highly detailed bidirectional representations from unlabeled text. It accomplishes this by considering both left and right contexts throughout its layers.

3.2.3. Dimensionality reduction

Dimensionality reduction is a technique for reducing the number of features or variables in a dataset while retaining as much useful information as possible. This method is used in this case to deal with high-dimensional data in a recommendation system. It improves computational efficiency, alleviates the curse of dimensionality, and improves data interpretability. The most commonly used dimensionality reduction algorithms are principal component analysis (PCA), LDA1, T-distributed stochastic neighbor embedding (t-SNE), least absolute shrinkage and selection operator (LASSO) and uniform manifold approximation and projection (UMAP). In the proposed method, LASSO is used for dimension reduction. The reason for choosing LASSO or L1 regularization is that the features of the text data we use are high-dimensional. When

high-dimensional data is reduced, the resultant dataset will become sparse. LASSO is proven to handle sparse data by eliminating features that are not important using a zero coefficient.

3.2.4. Similarity and distance measures

The similarity measure assesses the similarity of several terms, such as words, phrases, documents, or concepts. The goal of assessing the similarity between two phrases is to determine the degree to which two concepts are conceptually related. It will not compare words that are lexicographically similar. The shortest distance between two items will be determined in the case of the distance measure. The primary goal of these metrics is to group papers that are similar together. Cosine similarity, Jaccard coefficient, and inner products are the most often used similarity metrics, whereas distance measurements include Euclidean distance, Manhattan distance (L1 distance), Jensen-Shannon divergence, and hamming distance.

We employ cosine similarity to measure the similarity between the vectors because we have used the K-means algorithm for clustering. The cosine similarity measure is very versatile when combined with K-means. It also concludes that two vectors are close if the distance calculated between them is small.

3.2.5. Clustering algorithms

Documents are divided into groups using clustering techniques. The goal of clustering algorithms is to produce separate cluster based on the document similarity. In other words, papers within one cluster should be as similar as possible, and documents inside separate clusters should be as dissimilar as possible. Clustering algorithms are basically classified into three types partitioned-based clustering, hierarchical-based clustering and density-based clustering. For our proposed approach, the K-means algorithm is chosen for its simplicity and efficiency in handling high-dimensional data. Also, this is the root of all other clustering algorithms. The primary reason for selecting this algorithm is that the dataset we used is high-dimensional and the number of clusters is clearly known. So, to reduce the complexity of our approach and to provide an accurate result, this clustering was chosen for the study.

3.3. Citation networks

The primary objective of this study is to propose article recommendations to researchers by using a citation network formed with all survey articles for the search topic. To have a better understanding of this work, it is mandatory to know about the types of research. A research article published would be: original research, retrospective study, case study, methodological study, opinion study, short communication, and review articles. In our system, we categorize all types of research articles except review articles as one class and review articles as a separate class. The review article itself may be a narrative, systematic, or meta-analysis review that summarizes existing trends and gaps in the field of research. By accounting for a single survey article, it is possible to find the imperative research discoveries in the field. With these articles, it is possible to obtain citations for these significant research articles in a field over a period of study that can be used to recommend relevant articles. For this purpose, this proposed work constructs two different networks: SCN and CN.

3.3.1. Survey citation network

SCN is a directed graph constructed with a list of survey articles identified as surveys in the database. The set $UA = \{a_1, a_2, \dots, a_n\}$ consist of all articles formed by combining both graph SCN and CN. Let us consider the $UA = \{a_1, a_2, \dots, a_n\}$ and $SA = \{sa_1, sa_2, \dots, sa_n\}$ as the set of nodes formed with all articles from survey clusters along with their citations, and a relation between the nodes from sa_1 to sa_2 can be referred to as cited by. Consider any article from SA is an element of set A, such that $sa_i \in A$ is always true. Each node in this graph SCN is assigned three values: article ID (Aid), cited by value (CBv), and citing value (Cv). Article ID is the primary unique identifier that is used to identify each article separately in this network. CBv is the total count for the number of times the article has been “cited by” other articles. Similarly, Cv is the total number of articles that the current node is citing. Figure 3 illustrates a clear explanation of the SCN graph with seven nodes: A, B, C, X, Y, W, and Z. By considering the below example, the entire recommendation system can be explained as follows: $UA = \{A, B, C, Y, W, X, Y, Z\}$, and $SA = \{A, B, C\}$. Here A, B, and C are survey articles identified as relevant to the user’s search query. and then, with these nodes, the citations of the articles in A, B, and C are represented in Figure 3B). Figure 3B) represents the citations of each node; for example, article A has cited articles X and Y in the figure. The corresponding graph of the nodes in Figure 3B) is constructed into the graph as shown in Figure 3A) SCN graph. Figure 3C) explains the structure of node A and its values, which are used in the graph to find the most influential node in the graph SCN.

3.3.2. Citation network

CN is constructed with a list of articles other than the ones that are identified as survey articles. This network is formed with a set of nodes $OA = \{oa1, oa2, \dots, oan\}$, where OA represents the set of nodes from the original cluster along with their citations. The relationship of this graph, CN, can be referred to as “cited by” between two nodes from $oa1$ to $oa2$. Similar to SCN, the nodes in CN hold three values: CBv , Aid , and Cv . Consider any article from OA is an element of set A , such that $oai \in UA$ is always true and $UA = \{a1, a2, \dots, an\}$. The graph representation of the graph CN is similar that of graph SCN. The values in the nodes of these graphs (SCN and CN) are used to find a list of prominent articles, which are used by the recommendation system to find relevant articles based on the text similarity index of each recommended article.

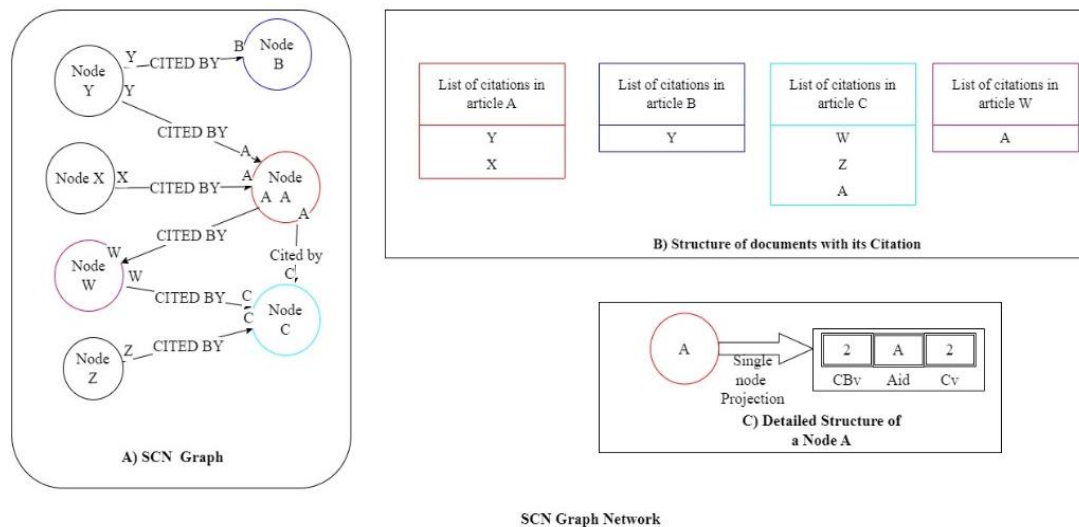


Figure 3. SCN graph and node structure

4. EXPERIMENTAL SET UP

The entire dataset is divided into two groups: groups of survey articles and groups of original research articles. Review, meta-analysis review and systematic review articles make up the articles in survey articles cluster. All other varieties of research articles, including case studies, methodological research articles, short communication articles, and others, are included in the original research cluster. The procedure used to divide the dataset into two distinct groups begins with clustering. Clustering is the initial phase of the recommendation system. Using this clustering, the data set is divided into survey cluster and original article clusters. Clustering text data is a time-consuming procedure that comprises preprocessing, text embeddings, dimensionality reduction, similarity measurements, and clustering. As a result of the clustering, two distinct clusters are generated. The next phase of the recommendation system is the construction of SCN and CN using these clusters. Further processing is computed by utilizing these two graphs, SCN and CN. The similarity score of each connection is obtained following the building of the SCN. The same procedure is done for the articles in the network CN, and they are once again grouped based on text similarity score along with computed citation value. The findings of these two clusters are then joint to calculate the weighted average of the similarity score. The recommendation algorithm makes recommendations based on the assessed similarity score. For our study, we have collected articles from sources like Google Scholar, DBLP, and IEEE Xplore, which contain over 607,331 articles from computer science majors. From the collected data, we have removed over 24,421 articles, as they have a few missing attributes that are considered important for our study. The articles in this dataset are identified as articles from machine learning, artificial intelligence, image processing, natural language processing, compiler design, internet of things (IoT), semantic web, data mining, computer vision, real-time computing, cyber security, computer networks, and algorithms. For the study, the major attributes or features we have concentrated on are text attributes: the title of the article and the abstract of the article; and meta-attributes: cited by articles and citing articles. The NLTK text data is used to preprocess this dataset. For embedding this pre-processed text data, we have utilized the BERT model. The variables used for this model during the training are listed in Table 1. Then, for dimension reduction, L1 regularization is used with an alpha value of 0.2. Finally, the clusters are formed

using K-means clustering. After clustering, two clusters formed, where the first cluster consisted of 30,454 number articles of type survey or review articles, and the second cluster consisted of 52,456 number articles clustered as original research articles. Then the articles that match the user query from both clusters are selected separately to generate the SCN and CN graphs. Then the CBv and Cv nodes for each node are calculated. With computed CBv and Cv, important nodes from the networks SCN and CN are identified separately. Then the Weighted Text Similarity (WTS) scores of these selected nodes are computed using (1). Then a final recommendation list is generated and presented as a list of recommended articles.

$$WTS(ai) = \frac{TS(ai)*SCN(CBv(ai)) + TS(ai)*CN(CBv(ai))}{2} \quad (1)$$

Where:

$WTS(ai)$: weighted text similarity score for article ai .

$TS(ai)$: text similarity score of article ai computed by comparing the researcher's query.

$SCN(CBv(ai))$: total cited by count of article ai in the SCN.

$CN(CBv(ai))$: total cited by count of article ai in the CN.

In the (1), the Cv values (representing the number of citing papers) have been effectively incorporated alongside the cited by counts (CBv) and text similarity (TS) values. The division by 2 in the formula ensures that the weighted average of the text similarity scores, considering both the SCN and CN, is computed accurately.

Table 1. Parameter settings for the proposed model

SS.No	Process	Algorithms	Value
1.	Text preprocessing	NLTK	-
2.	Text embedding	BERT	bert-large-uncased
3.	Dimension reduction	LASSO Alpha value, α	0.2
4.	K-means clustering	Number of clusters, K	2
5.	Graph construction	Threshold for text similarity, TS	0.5

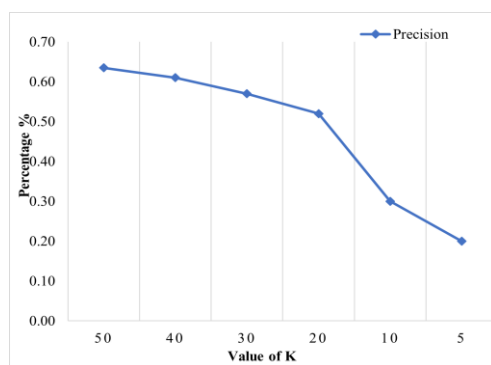
5. RESULTS AND DISCUSSIONS

The proposed Srvycite method is compared with the below-baseline methods. The accuracy of our proposed method is evaluated using the metrics precision and recall. Precision is a measure that is used to identify the proposition of the correct prediction predicted by the methods. Recall is a measure that identifies the portion of an accurate result from the predicate results. Zhu *et al.* [3] recommend using a search query based on the behaviour of the user along with the current draft of the article. This method assumes users' previous publications and cited articles will remain the same and recommends articles. Not only that, but this method does not even consider the content of the published articles. For text features, it only considers the topics of previously published articles and the draft of the current article. It is a collaborative method; hence it suffers from the cold start problem if a new user enters the system. Chaudhuri *et al.* [15] proposed a work by developing a recommendation system by using a direct text feature along with four new indirect features to calculate the diversity in keywords, complexity in sentences, citation analysis over time, and quality of the articles. Mabude *et al.* [8] proposed an integrated approach to recommend research articles and expertise recommendation systems, which are basically two different recommendation systems. But both of these systems use the same set of data sets to recommend two different outputs; therefore, these two recommendation systems are integrated into a single system. In the work proposed by Zhang *et al.* [22], created a multidimensional model based on the user's academic portrait. Here, the articles are recommended based on basic information about the user, such as age, gender, education, major, identity, and direction of the research. Then, based-on attributes like academic motivation, cognitive style, and domain knowledge of academic personality are analysed, and finally, behaviour characteristics are computed using the basic information already obtained. Li *et al.* [16] proposed a graph-based recommendation system, MARec, where the authors used an attributed heterogeneous information network (aHIN) network for information extraction, GAT for generation of network embedding to identify the user meta path, and Bi-LSTM infused with attention-aware mechanisms to capture recent articles and recommend them to users. Smail *et al.* [11] compared two top topic modelling techniques, LDA and NMF. The LDA is implemented using Scikit, and the NMF is implemented using Gensim and Mallet packages. The authors concluded that using the NMF with the Mallet package for topic modelling produced better precision for the recommendation of research articles in the dataset used. Table 2 summarizes the comparison of the proposed srvycite model with selected baseline models. In Table 2, the comparison study compares the methodology, features used by the methods

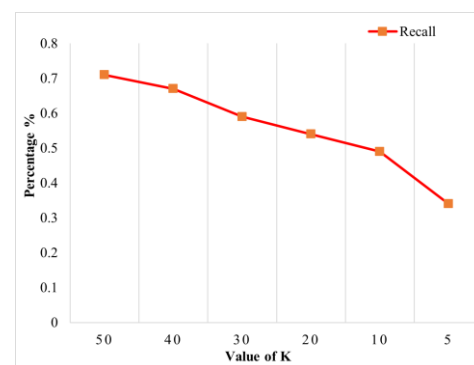
to perform recommendations, the dataset used, and evaluation metrics used to calculate the accuracy of the discussed models. Figure 4 presents the precision and recall results from the Srvycite model. We have calculated the results for precision (Figure 4(a)) and recall (Figure 4(b)) @K values for 5, 10, 20, 30, 40, and 50. From the result, it is clear that the recall measure of the proposed model shows better improvement for @K values above 10. Similarly, for precision, there is a considerable increase in the result with @K values greater than 20. With our dataset, it can be concluded that there is a constant increase in the accuracy of the results with higher values of K using the proposed Srvycite method. Based on our dataset, it can be inferred that there exists a consistent augmentation in the accuracy of the results as the @K values increase when employing the proposed Srvycite method. This observation suggests that the Srvycite model effectively enhances the precision and recall of recommended articles, particularly as the size of the recommendation list expands.

Table 2. Comparison of proposed models with the baseline models

S. No	Work	Methodology	Features of papers used	Dataset	Evaluation metrics
1.	Proposed work	K-means + survey citation + citation network	Abstract +title + citation network + survey paper network	DBLP, Google Scholar, IEEE	Precision recall accuracy
2.	Zhu <i>et al.</i> [3]	Bidirectional recurrent neural network (RNN) with attention mechanism (text feature embedding: PV-DM)	Title + author publication and citation information	DBLP	Precision@20 Recall@20 MRR NDCG@20
3.	Chaudhuri <i>et al.</i> [15]	ensemble models-AdaBoost, bagging, voting, and stacking	15 article features + paper metadata + keyword diversity measurement + sentence complexity analysis + citation analysis over time + scientific quality measurement	Scopus	Precision Recall F1 score Accuracy RMSE
4.	Smail <i>et al.</i> [11]	HIN+GAT + RNN	Topic + venue + paper + user	DBLP, Spotify, and Yelp	F1-score@5, Accuracy and NDCG@5.
5.	Mabude <i>et al.</i> [8]	Matrix factorization + KMeans	Publication count, citations, user ratings, and article key phrase	Scopus	Precision recall F1 score accuracy RMSE



(a)



(b)

Figure 4. Srvycite @K value (a) precision and (b) recall

The precision and recall results obtained from the Srvycite model were calculated for different values of K (5, 10, 20, 30, 40, and 50). Precision measures the proportion of relevant articles among the total recommended articles, while recall assesses the proportion of relevant articles retrieved by the recommendation system out of all relevant articles available in the dataset. We observed that the recall measure of the proposed model exhibited better improvement for K values above 10. Similarly, precision showed a considerable increase for K values greater than 20. These findings indicate a consistent increase in the accuracy of the results with higher values of K using the proposed Srvycite method. Furthermore, we compared the performance of our proposed Srvycite method with baseline methods by evaluating precision and recall. The percentage improvement metric was employed to quantify the extent to which our method outperformed the baseline methods. We observed a significant improvement of 15–20 percent in both

precision and recall, demonstrating the effectiveness of the Srvycite method in recommending relevant research articles.

Figure 5 compares the results obtained by our proposed approach with the base-line approaches selected for this study. After comparing the proposed Srvycite method with the selected baseline methods using the collected dataset, the result is presented in Figure 5 at precision and recall values of 50. From the results, it is proven that the proposed model performs far better than other baseline models. Among the base line models selected for the study, the model proposed by Smail *et al.* [11] performed better in terms of both precision and recall at a K value of 50. But the performance of the Srvycite compared with the above model produces a better result for precision and recall. The performance of precision and recall of the Srvycite has increased by 3.8% and 2.1% compared with other models for the K value 50 in our dataset. So, from the experimental results conducted from the collected dataset, it is clearly proven that this proposed Srvycite model is considered to provide better accuracy in the recommendation of research articles by considering the citation networks of the survey articles as the main feature.

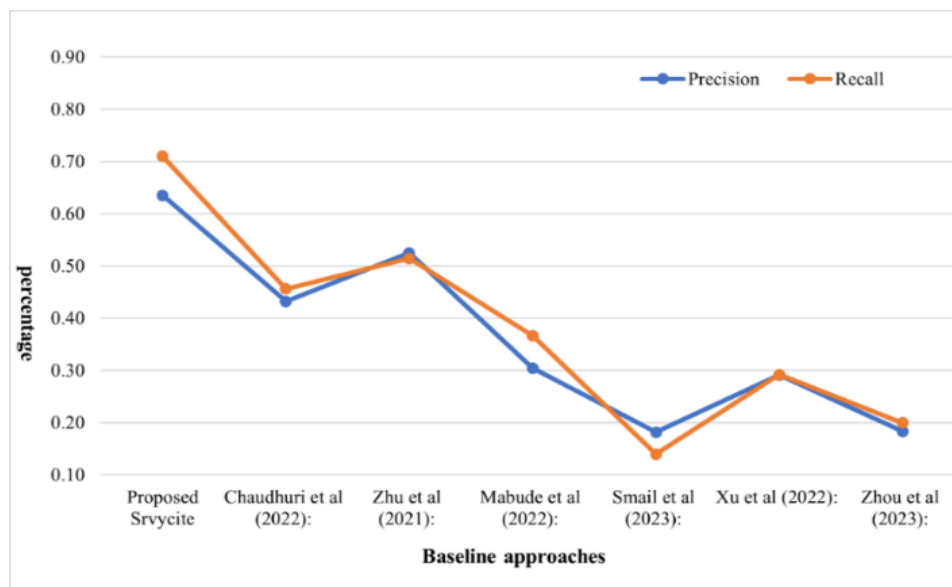


Figure 5. Comparison of baseline approaches with Srvycite

6. CONCLUSION




In the aim of recommending research articles relevant for the researcher depending on the search query, this article proposes a novel methodology called Srvycite. Our proposed method uses the citations of normal articles along with the survey articles to recommend relevant articles. In this study, we consider primary features for recommendation, such as citation networks and text features like the abstract and title of the article. Since we used text features, it requires text preprocessing that is done using NLTK and K-means for clustering articles based on the abstract, and finally, citation and survey citation graphs are utilized for the article recommendation. This study primarily focuses on implementing the survey citation network to identify the most prominent articles in the field of research. We believe that review/survey articles, if considered for the recommendation task, can be a game changer in article recommendation with less utilization of resources. This study is performed to prove the fact about the role of survey articles in finding good-quality relevant articles for the study. The experiments done in this context also prove the importance of the survey citation network for citation recommendations. By comparing the proposed model with the existing model in the dataset, we proved the increase in the performance of the system. For determining the performance of the system, we have employed precision and recall metrics. To further increase the accuracy of the article recommendation system, this proposed model could be enhanced by incorporating other meta-data of an article, like venue, author, and year of publication. The Srvycite approach has already achieved a remarkable increase in the accuracy of the result only by utilizing some of the article's features. This approach could be further improved to find the quality of the survey articles along with the time line tag to stay with current improvements.

REFERENCES




- [1] J. Beel, B. Gipp, S. Langer, and C. Breiter, "Research-paper recommender systems: a literature survey," *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, 2016, doi: 10.1007/s00799-015-0156-0.
- [2] M. H. Ahmed, S. Tiun, N. Omar, and N. S. Sani, "Short text clustering algorithms, application and challenges: a survey," *Applied Sciences (Switzerland)*, vol. 13, no. 1, p. 342, 2023, doi: 10.3390/app13010342.
- [3] Y. Zhu, Q. Lin, H. Lu, K. Shi, P. Qiu, and Z. Niu, "Recommending scientific paper via heterogeneous knowledge embedding based attentive recurrent neural networks," *Knowledge-Based Systems*, vol. 215, p. 106744, 2021, doi: 10.1016/j.knsys.2021.106744.
- [4] N. Sakib, R. B. Ahmad, and K. Haruna, "A collaborative approach toward scientific paper recommendation using citation context," *IEEE Access*, vol. 8, pp. 51246–51255, 2020, doi: 10.1109/ACCESS.2020.2980589.
- [5] B. Shao, X. Li, and G. Bian, "A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph," *Expert Systems with Applications*, vol. 165, p. 113764, Mar. 2021, doi: 10.1016/j.eswa.2020.113764.
- [6] A. Subakti, H. Murfi, and N. Hariadi, "The performance of BERT as data representation of text clustering," *Journal of Big Data*, vol. 9, no. 1, p. 15, Dec. 2022, doi: 10.1186/s40537-022-00564-9.
- [7] C. Little, D. McLean, K. Crockett, and B. Edmonds, "A semantic and syntactic similarity measure for political tweets," *IEEE Access*, vol. 8, pp. 154095–154113, 2020, doi: 10.1109/ACCESS.2020.3017797.
- [8] C. N. Mabude, I. O. Awoyelu, B. O. Akinyemi, and G. A. Aderounmu, "An Integrated approach to research paper and expertise recommendation in academic research," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, pp. 485–495, 2022, doi: 10.14569/IJACSA.2022.0130456.
- [9] M. Q. Khan *et al.*, "Impact analysis of keyword extraction using contextual word embedding," *PeerJ Computer Science*, vol. 8, p. e967, May 2022, doi: 10.7717/peerj-cs.967.
- [10] M. Mataoui, F. Sebbak, A. H. Sidhoum, T. E. Harbi, M. R. Senouci, and K. Belmessous, "A hybrid recommendation system for researchgate academic social network," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 53, Mar. 2023, doi: 10.1007/s13278-023-01056-1.
- [11] B. Smail, H. Aliane, and O. Abdeldjalil, "Using an explicit query and a topic model for scientific article recommendation," *Education and Information Technologies*, vol. 28, no. 12, pp. 15657–15670, 2023, doi: 10.1007/s10639-023-11817-2.
- [12] A. Hadhiatma, A. Azhari, and Y. Suyanto, "A scientific paper recommendation framework based on multi-topic communities and modified PageRank," *IEEE Access*, vol. 11, pp. 25303–25317, 2023, doi: 10.1109/ACCESS.2023.3251189.
- [13] E. Gündoğan and M. Kaya, "A novel hybrid paper recommendation system using deep learning," *Scientometrics*, vol. 127, no. 7, pp. 3837–3855, 2022, doi: 10.1007/s11192-022-04420-8.
- [14] R. Molaei Fard and P. Yarahmadi, "Providing a recommendation system for recommending articles to users using data mining methods," *Journal of Computer & Robotics*, vol. 15, no. 2, pp. 49–58, 2022.
- [15] A. Chaudhuri, M. Sarma, and D. Samanta, "SHARE: designing multiple criteria-based personalized research paper recommendation system," *Information Sciences*, vol. 617, pp. 41–64, 2022, doi: 10.1016/j.ins.2022.09.064.
- [16] W. Li, C. Chang, C. He, Z. Wu, J. Guo, and B. Peng, "Academic paper recommendation method combining heterogeneous network and temporal attributes," in *Computer Supported Cooperative Work and Social Computing: 15th CCF Conference, ChineseCSCW 2020*, 2021, pp. 456–468, doi: 10.1007/978-981-16-2540-4_33.
- [17] Y. Kang, A. Hou, Z. Zhao, and D. Gan, "A hybrid approach for paper recommendation," *IEICE Transactions on Information and Systems*, vol. E104D, no. 8, pp. 1222–1231, 2021, doi: 10.1587/transinf.2020BDP0008.
- [18] H. Liu, H. Kou, C. Yan, and L. Qi, "Keywords-driven and popularity-aware paper recommendation based on undirected paper citation graph," *Complexity*, vol. 2020, pp. 1–15, Apr. 2020, doi: 10.1155/2020/2085638.
- [19] M. Tao, X. Yang, G. Gu, and B. Li, "Paper recommend based on LDA and PageRank," in *Artificial Intelligence and Security: 6th International Conference, ICAIS 2020*, Hohhot, China: Springer, 2020, pp. 571–584.
- [20] N. Du, J. Guo, C. Q. Wu, A. Hou, Z. Zhao, and D. Gan, "Recommendation of academic papers based on heterogeneous information networks," in *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, Nov. 2020, vol. 2020-Novem, pp. 1–6, doi: 10.1109/AICCSA50499.2020.9316516.
- [21] X. Ma, Y. Zhang, and J. Zeng, "Newly published scientific papers recommendation in heterogeneous information networks," *Mobile Networks and Applications*, vol. 24, no. 1, pp. 69–79, 2019, doi: 10.1007/s11036-018-1133-9.
- [22] Y. Zhang, M. Wang, F. Gottwalt, M. Saberi, and E. Chang, "Ranking scientific articles based on bibliometric networks with a weighting scheme," *Journal of Informetrics*, vol. 13, no. 2, pp. 616–634, 2019, doi: 10.1016/j.joi.2019.03.013.
- [23] A. Kanakia, Z. Shen, D. Eide, and K. Wang, "A scalable hybrid research paper recommender system for Microsoft Academic," in *The World Wide Web Conference*, May 2019, pp. 2893–2899, doi: 10.1145/3308558.3313700.
- [24] J. Son and S. B. Kim, "Academic paper recommender system using multilevel simultaneous citation networks," *Decision Support Systems*, vol. 105, pp. 24–33, Jan. 2018, doi: 10.1016/j.dss.2017.10.011.
- [25] K. Haruna, M. Akmar Ismail, D. Damiasih, J. Sutopo, and T. Herawan, "A collaborative approach for research paper recommender system," *PLOS ONE*, vol. 12, no. 10, p. e0184516, Oct. 2017, doi: 10.1371/journal.pone.0184516.
- [26] X.-Y. Liu and B.-C. Chien, "Applying citation network analysis on recommendation of research paper collection," in *Proceedings of the 4th Multidisciplinary International Social Networks Conference*, Jul. 2017, vol. Part F1296, pp. 1–6, doi: 10.1145/3092090.3092138.
- [27] R. Sharma, D. Gopalani, and Y. Meena, "Concept-based approach for research paper recommendation," in *Pattern Recognition and Machine Intelligence: 7th International Conference, PREMI 2017*, 2017, pp. 687–692, doi: 10.1007/978-3-319-69900-4_87.
- [28] M. Dhanda and V. Verma, "Recommender system for academic literature with incremental dataset," *Procedia Computer Science*, vol. 89, pp. 483–491, 2016, doi: 10.1016/j.procs.2016.06.109.
- [29] F. Xia, H. Liu, I. Lee, and L. Cao, "Scientific article recommendation: exploiting common author relations and historical preferences," *IEEE Transactions on Big Data*, vol. 2, no. 2, pp. 101–112, 2016, doi: 10.1109/tbdata.2016.2555318.
- [30] M. Amami, G. Pasi, F. Stella, and R. Faiz, "A LDA-based approach to scientific paper recommendation," in *Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016*, 2016, pp. 200–210, doi: 10.1007/978-3-319-41754-7_17.
- [31] T. Chakraborty, A. Krishna, M. Singh, N. Ganguly, P. Goyal, and A. Mukherjee, "FeRoSA: a faceted recommendation system for scientific articles," in *Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference, PAKDD 2016*, 2016, pp. 528–541, doi: 10.1007/978-3-319-31750-2_42.

BIOGRAPHIES OF AUTHORS



Sivasankari R    is an Assistant Professor in the Department of Computer Science and Applications. She earned her Master of Computer Applications (MCA) degree from Anna University in 2014. Currently, she is pursuing her research as a scholar at SRM Institute of Science and Technology (SRMIST) in Ramapuram, Chennai. Her primary research interests lie in the field of the internet of things (IoT), where she focuses on innovative applications and security mechanisms. With a strong academic background and ongoing research, she contributes significantly to the advancement of computer science and IoT technologies. She can be contacted at email: manjurmoni@gmail.com.



Prof. Dr. J. Dhilipan    is currently serving as the Vice Principal Admin and the Head of Computer Applications (MCA) at SRM Institute of Science and Technology, Chennai with an experience of more than 25 years. He obtained his Ph.D. degree in the year 2014 from Manonmamiam Sundaranar University in the field of E-Commerce and Data Analytics. He has published several research papers in peer reviewed reputed scopus indexed journals in the areas of machine learning, cloud computing, IoT, and data analytics. He can be contacted at email: hod.mca.rmp@srmist.edu.in.