IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 21, 2019, accepted February 10, 2019, date of publication February 28, 2019, date of current version March 26, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2900520*

# A Hybrid Approach Toward Research Paper Recommendation Using Centrality Measures and Author Ranking

**WALEED WAHEED, MUHAMMAD IMRAN, BASIT RAZA, AHMAD KAMRAN MALIK, AND HASAN ALI KHATTAK, (Senior Member, IEEE)**
Department of Computer Science, COMSATS University Islamabad, Islamabad 45550, Pakistan

Corresponding author: Muhammad Imran (mimran@comsats.edu.pk)

This work was supported by the COMSATS University Islamabad (CUI), Islamabad, Pakistan, under Grant CUI/ORIC-PD/19.

**ABSTRACT** The volume of research articles in digital repositories is increasing. This spectacular growth of repositories makes it rather difficult for researchers to obtain related research papers in response to their queries. The problem becomes worse when a researcher with insufficient knowledge of searching research articles uses these repositories. In the traditional recommendation approaches, the results of the query miss many high-quality papers, in the related work section, which are either published recently or have low citation count. To overcome this problem, there needs to be a solution which considers not only structural relationships between the papers but also inspects the quality of authors publishing those articles. Many research paper recommendation approaches have been implemented which includes collaborative filtering-based, content-based, and citation analysis-based techniques. The collaborative filtering-based approaches primarily use paper-citation matrix for recommendations, whereas the content-based approaches only consider the content of the paper. The citation analysis considers the structure of the network and focuses on papers citing or cited by the paper of interest. It is therefore very difficult for a recommender system to recommend high-quality papers without a hybrid approach that incorporates multiple features, such as citation information and author information. The proposed method creates a multilevel citation and relationship network of authors in which the citation network uses the structural relationship between the papers to extract significant papers, and authors' collaboration network finds key authors from those papers. The papers selected by this hybrid approach are then recommended to the user. The results have shown that our proposed method performs exceedingly well as compared with the state-of-the-art existing systems, such as Google scholar and multilevel simultaneous citation network.

**INDEX TERMS** Citation networks, collaboration networks, recommender systems, research paper recommendation systems.

## I. INTRODUCTION

The process of literature review starts with finding relevant research articles using search engines. The number of freely available academic articles on the web have risen up-to 25 million [1]. The task of recommending related articles from such huge volume is non-trivial, as the search system has to deliver best results by handling big data. The problem becomes worse when beginners cannot find their relevant articles due to lack of experience in using these search engines [2]. The process of filtering relevant papers manually is also a time-consuming and tedious task due to such a large scale of research data available. Therefore, an efficient research paper recommender system is needed which produces high quality recommendations from these digital repositories [3], [4].

There are many recommender systems implemented but few of them focus on recommendation of academic papers [5]. These methods consist of collaborative filtering, content-based filtering and citation analysis based techniques. Collaborative filtering is mostly used recommendation technique in academic recommender systems. It recommends articles based on the paper-citation matrix which shows past preferences of the users. However, this

The associate editor coordinating the review of this manuscript and approving it for publication was Congduan Li.

technique can cause cold start problem due to not having sufficient number of paper citations which are needed for recommendations. It also generates data sparsity problem due to having huge size of the paper-citation matrix [6]. The drawbacks of collaborative filtering are overcome using content-based filtering techniques where recommendations are based on the comparison of textual information between research articles [7]. However, this method does not capture the semantics of the user interests and cannot handle the ambiguity due to natural language [8].

Citation analysis comprises of co-citation analysis [9] and bibliographic coupling [10]. Research papers cite papers that are closely related with them. Therefore, relations between papers are more meaningful and purposeful. The disadvantage of using citation-based method is that it only considers the citations and does not consider the content of paper which may lead to inappropriate results. For example, when the cited paper is only added in the reference section without being used in the content of research paper then these citations become useless. Google's PageRank is another approach for recommending research papers [11]. It is used by Google Scholar to recommend articles in the "related work" section of the web page. PageRank measures the authority of paper and ranks it based on the number of citations it receives from other academic articles. Its major drawback is that it uses citation count as a metric to recommend articles which fails to recommend quality articles when recently published paper is selected as the paper of interest [12].

This paper proposes a hybrid technique for research paper recommendation that combines multi-level citation network and an author's relationship network. It considers the structural relationship of papers with the paper of interest and creates a ten-level citation network by placing paper of Interest $I$ as an ego-node and using references at the end of each paper to expand the network in both directions. The resultant network is shown in Figure 3. The reason for using ten-level network lies in the state-of-the-art literature, as it has been recommended as a reasonable size and using more than ten levels may include papers not related to the paper of interest [13]. The importance of each paper with the paper of interest is examined by applying four centrality parameters named betweenness centrality, eigenvector centrality, degree centrality and closeness centrality. The traditional recommendation approaches do not focus on the importance of authors and hence recommend articles which does not match the expectation of the users. This approach applies another filter to the recommended articles by creating relationship network of authors and identifies key author using the above mentioned parameters. After the identification of key authors, quality of paper is examined, and top *'n'* high quality papers are recommended to users.

The main contribution of this research work is a novel approach towards research paper recommendations which combines multi-level citation network and collaboration network of authors to generate high quality recommendations as compared to existing techniques.

The remaining paper is structured as follows. Literature review of existing recommendation approaches is presented in Section II. Section III elaborates the proposed methodology. Detail about Experiments and evaluation are presented in Section IV. While results are discussed in Section V. Finally Section VI concludes the paper.

## II. RELATED WORK

There are several research paper recommendation methods which focus on finding similarity between research articles [14]. These methods include: (1) collaborative filtering [15] (2) meta-data based [16], [17] (3) content-based [18], [19] (4) citation-based [9], [10], [20], [21] (5) multi-level citation network [13] (6) and (7) user profile-based [22]–[24] approaches.

Collaborative filtering finds relationship between research papers and is used in most recommendation systems [25], [26]. This method takes citing paper and cited papers corresponding to users and items in e-commerce respectively and generates paper-citation matrix from the citation network. This approach takes commonly cited papers as a measure and computes similarity between the papers using citation-score metric. There are many limitations of using this approach in which the most common one is called cold start problem. Papers are recommended to users based on their citations by other articles. Therefore, if a new paper is selected as a paper of interest, it has to be cited by a number of research papers for generating recommendations.

Meta-data based methods [16], [17] find similarity between research papers by comparing the meta-data of research papers which includes title of the research paper, name of authors, keywords and date of publication. The main advantage of using meta-data methods is the free availability of research paper meta-data even if they are published in paid journals. However, these methods do not always provide correct recommendations. For instance, when the common author has published research papers in different research fields then the recommendations provided by meta based methods are not accurate.

Content-based [18], [19] approaches find relationship between two papers by matching their contents. It gives improved results and is proven to be a better option than merely relying on meta-data based techniques. The main drawback of using this technique is that the whole content of research papers are not available online in most digital libraries. Furthermore, the process of matching the whole text of research paper takes a lot of time and proves to be very costly.

Co-citation analysis technique measures the similarity of two papers cited together by one or more common papers [9], [10], [20]. In this technique, papers are recommended based on the fact that co-cited papers belong to the similar area of research and can be potential set of papers of user's interest. However, this technique does not consider content of paper or any other feature in the paper, which leads to inconsistent recommendations [28]. Bibliographic

**TABLE 1.** Comparison between different recommender systems.

| Systems | Drawbacks | Related Document Approach | Citation Relationship | Semantic Relationship | Author Analysis | Papers Quality |
|---|---|---|---|---|---|---|
| Google Scholar [11], [27] | Once a highly cited paper is selected as the paper of interest, Google Scholar will recommend other highly cited papers filtering out recent papers or papers with low citation count. Similarly, when an older paper is selected as paper of interest, most of the Google Scholar's recommended papers are out of date. | PageRank | Not considered | Not considered | Not considered | Considered only in case of prominent paper selection |
| Multilevel Simultaneous Citation Network (MSCN) [13] | This technique does not consider quality of the paper for recommendations. The quality of paper can be incorporated by analyzing authors of individual papers. | Multi-level Citation Network | Considering the relationship of "cite"and "cited by"equally | Structural and semantic relationships are inspected through indirect links to paper of interest | Not considered | Not considered |
| **Proposed** Citation Network of Papers and Relationship Network of Authors (CNRN) | – | Multi-level Citation Network + Relationship Network of Authors | Considered direct, indirect relations by co-citation and bibliographic coupling | Considered semantic relationship between papers | Identify key authors from relationship network of authors | Recommend high quality papers |

coupling is another technique for recommending related papers [21]. It measures the similarity of two papers that cite one or more common papers. Like co-citation, this technique also ignores the logical structure and content of the paper and only considers structural relationship between them. Another problem with this technique arises when there is an absence of citation in the text corresponding to the references added in the reference list. These citations are known as false citations and such citations also lead to inappropriate results.

User profile-based approaches are based on user interests and access-log history [22], [23]. These approaches recommend papers to user based on their available information in digital libraries. The main drawback of using profile-based approaches is that sufficient results are not achieved when the available information is not enough. Mendeley uses profile-based technique for research paper recommendation [29]. The recommendations are based on what the user lastly read either from the Mendeley Desktop, mobile application or its web library. Furthermore, it considers reference list of user from the library and research areas mentioned in the profile description. The recommendation set also suggests the references that are popular among Mendeley users of the same discipline.

Google Scholar web search engine enables the researchers to search academic literature and scientific publications in digital repositories. It uses text mining and citation count to list the results in response to the user's search query. Google Scholar recommendation system employees making new connections philosophy as it is backed by powerful Google search algorithm. When author adds publications to their profile, the Google Scholar searches the indexes of scholarly content for presenting papers and articles that matches the given publications. By using a statistical model based on citations and co-authorships, the most relevant research articles are recommended. New relevant articles are recommended to the users by maintaining the users profile data, their interests, and area of research [12]. Another approach for generating recommendations is by analyzing the reading behavior instead of search patterns. For instance, Science Direct platform measures the reading behavior having more than 10 million unique visitors a month and over 700 million downloaded articles each year.

The related research papers returned by traditional recommender systems are mostly based on either structural based approaches where relationship between papers is assessed for recommendation or content based approaches where meta-data or content of the research paper is analyzed for recommendation. The existing work only considers citation network for generating recommendations [13], which is insufficient to generate high quality results. The potential extensions to improve the performance are to include author and journal information. Our proposed study moves one step towards that milestone and additionally incorporates author information along with the citation network for generating recommendations. The contribution of this work is a hybrid approach which incorporates both of the citation network and author ranking to recommend better quality results. The proposed recommender approach named citation Network of papers and Relationship Network of authors (CNRN) is based on Multilevel Simultaneous Citation Network (MSCN) and overcomes problems when either old or new papers are selected as paper of interest. MSCN evaluates the importance of each essential paper through centrality measures. The three basic steps consist of initially generating directional multilevel citation networks, then selecting candidate papers thereby computing candidate score of each paper and finally average ranking of each candidate paper for final recommendation. Network is generated up to ten levels where the nodes represent papers and the links between them represent citation with forward and backward links. For papers citing paper *I* backward link is exploited whereas forward link identifies the paper cited by paper *I*. To calculate the level of multilevel citation network, the sum of forward and backward direction links is computed [13].

Table 1 shows the comparison between existing recommender systems and our proposed recommendation approach CNRN. The limitations mentioned in Table 1 are addressed in the proposed approach.

## III. PROPOSED METHODOLOGY

This study proposes a recommendation approach by evaluating the importance of each paper using centrality measures (Equations 4, 5, 6, 7). Fig. 1 shows the block diagram of our proposed CNRN approach. It passes information through a set of sequential steps before generating recommendation for users. The detailed working of the proposed CNRN recommendation system is shown in Fig. 2. In the first step (a), citation network of papers is generated with the paper of interest using cited and cited by relationship. Then (b), candidate score is calculated for each paper and relevant papers are selected based on the candidate score. In the third step (c), centrality measures are calculated for each paper and are converted into ranks. We calculated average rank of each paper and extracted authors from top papers in fourth step (d). Fifth step (e) generates author's collaboration network and calculates author's collaboration score by applying centrality measures and top authors are selected based on their collaboration score. In the final step (f), top papers published by top authors are recommended to user in *related work* section of the results. The description of each of the step is provided below:

### A. CREATION OF A CITATION NETWORK

In the first step, a citation network is created based on reference list appearing at the end of the paper of interest. Citation network has ten levels, five levels both in the forward and backward direction. Forward direction includes papers cited by the paper of interest and backward direction constitutes of those papers which cite the paper of interest. In existing studies, citation analysis creates ten-level citation networks [13]. This work also generates ten-level citation network because considering more than ten levels may include papers that are not related to the paper of interest. The algorithm for creation of the citation network is given in Algorithm 1. It takes paper of interest as an input and returns all related papers to the paper of interest up to five levels in the both directions. The set of related papers are maintained in the form of a list.

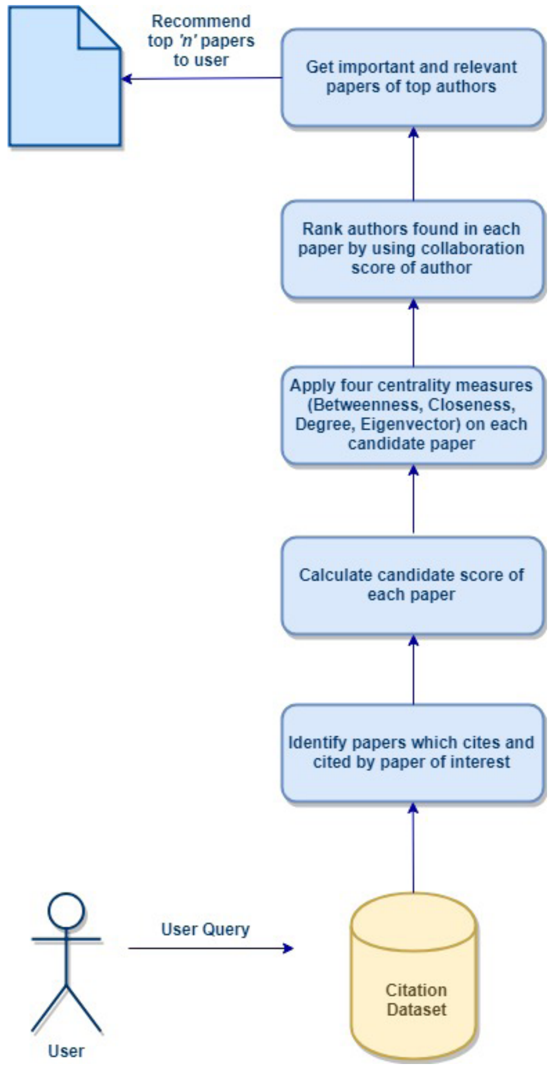



**FIGURE 1.** **The proposed CNRN block diagram.**

### B. SELECTION OF RELEVANT PAPERS FROM THE CITATION NETWORK

The candidate score of each paper is calculated to select relevant papers from the citation network. The relevancy of papers is measured by using bibliographic coupling and co-citation analysis. Figure 4 describes the bibliographic coupling (B.C) and co-citation (C.C) for two sample documents X and Y. If both the documents X and Y are citing papers A, B and C then the B.C($X_A$, $Y_A$) is three. It represents the number of documents mutually cited by any two papers. If both the documents X and Y are cited by documents A, B and C, then the C.C($X_A$, $Y_A$) of documents X and Y is three, which shows the number of articles mutually citing documents X

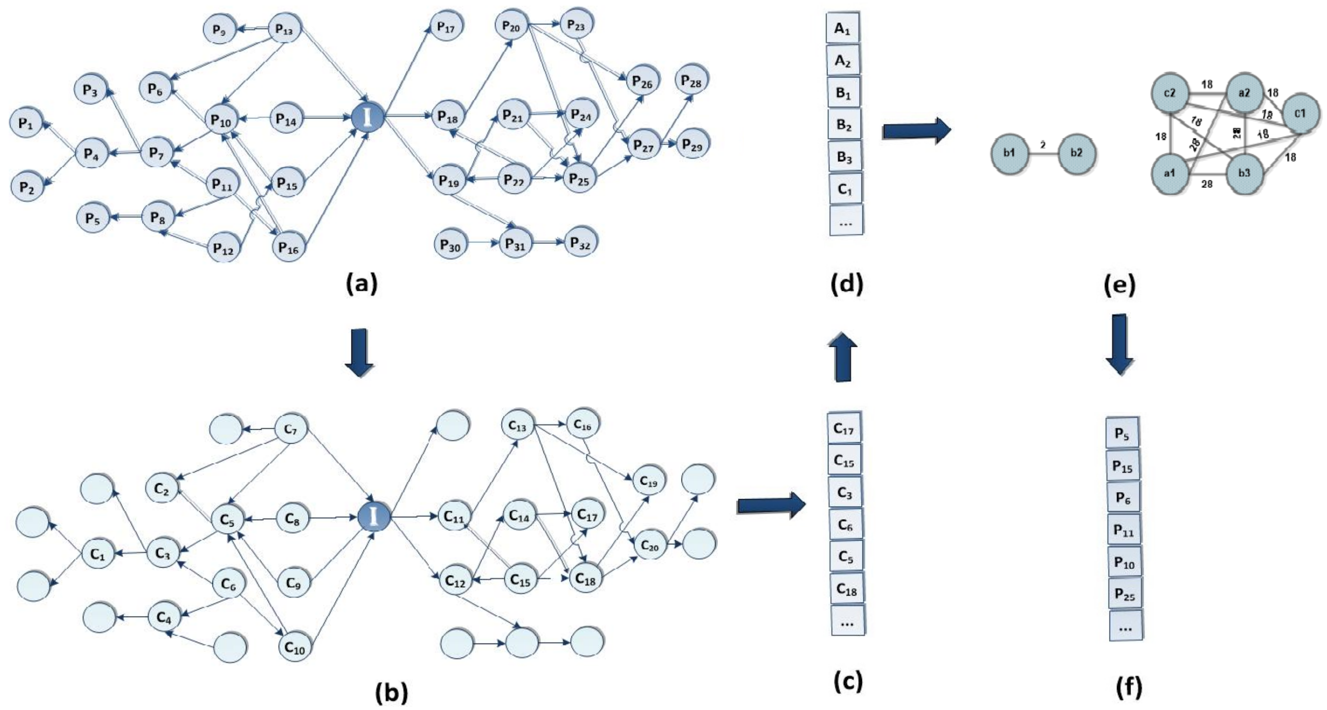



**FIGURE 2.** Overview of the Proposed CNRN recommendation approach. (a) Citation network of papers is generated with the paper of interest. (b) Candidate score is calculated for each paper and relevant papers are selected based on the candidate score. (c) Centrality measures are calculated for each paper and are converted into ranks [13]. (d) Average rank of each paper is calculated and authors are extracted from top papers. (e) Authors collaboration network is created and top authors are selected based on the collaboration score. (f) Top papers published by top authors are recommended to user.

**Algorithm 1** Creation of Citation Network

**Input** PAPER of INTEREST
**Output** RELATED PAPERS
create array of related papers
if paper of interest is selected
  get all papers which are cited by paper of interest
  get all papers which cites paper of interest
  add all papers to related papers list
return related papers

and Y. B.C($X_A$, $Y_A$) and C.C($X_A$, $Y_A$) are calculated using Equations 1 and 2 respectively.

$$B.C(X_A, Y_A) = \begin{cases} 1, & \text{if doc } X \text{ and } Y \text{ both cite doc } A \\ 0, & \text{else} \end{cases} \tag{1}$$

$$C.C(X_A, Y_A) = \begin{cases} 1, & \text{if doc } X \text{ and } Y \text{ are cited by doc } A \\ 0, & \text{else} \end{cases} \tag{2}$$

The C.C and B.C metrics are used in calculating C-Score, which gives the relevancy of papers to the paper of interest. The C-Score for the paper of interest *P* can be calculated using Equation 3. The *J* represents set of all papers excluding paper *P* and the denominator gives the distance between paper *J* and the paper of interest *I*. A high value of numerator specifies that the paper *P* is closely related to the paper *J* while a low value tells that *P* is not relevant to paper *J*. Similarly, the denominator *d (I, P)* specifies the number of hops between papers *I* and *P*. The higher number of hops indicate that the two papers are not closely related to each other.

$$C - Score = \frac{\sum_{j=1}^{n}((B.C(P,J)) + (C.C(P,J)))}{d(I,P)} \tag{3}$$

**Algorithm 2** Selection of Candidate Papers

**Input** RELATED PAPERS
**Output** CANDIDATE PAPERS
create array candidate papers
iterate over list of nodes and edges
  calculate Bibliographic-coupling of each paper by using Equation 1
  calculate Co-citation of each paper by using Equation 2
  calculate Total Similarity of each paper
  calculate Distance of each paper
  calculate Candidate-Score of each paper by using Equation 3
  add papers to candidate papers list
return candidate papers list

Candidate papers are selected based on C-Score. The algorithm for selection of candidate papers is shown in Algorithm 2. It takes list of related papers generated by 1 as input and returns a list of candidate papers using C-Score.

**TABLE 2.** Candidate score's calculation.

| Paper | Co citation | Bibliographic Coupling | Total Similarity | Total Distance | Candidate Score |
|---|---|---|---|---|---|
| $P_{10}$ | 6 | 2 | 8 | 2 | 4 |
| $P_{25}$ | 6 | 2 | 8 | 3 | 1.6 |
| $P_{31}$ | 1 | 0 | 1 | 2 | 0.5 |

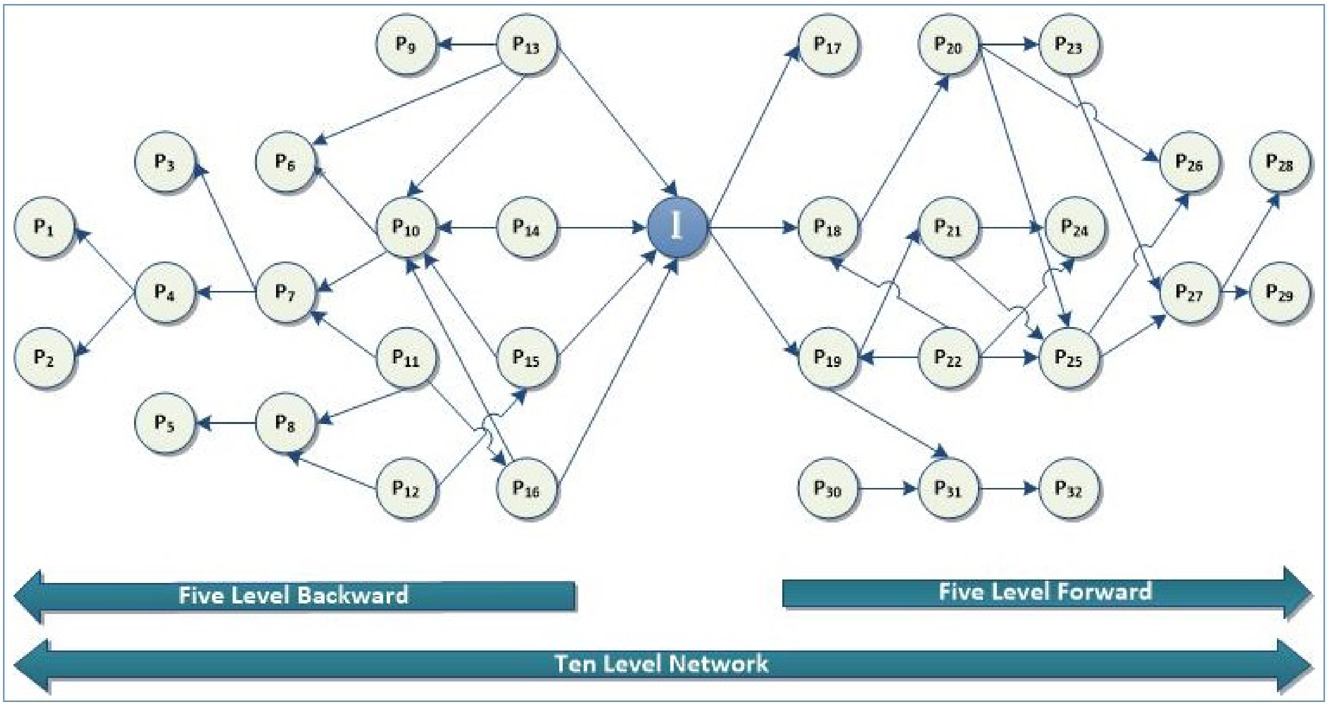



**FIGURE 3.** An example of a ten level citation network adapted from [13].

Existing studies used network size between 500-800 papers for experimentation [13]. Therefore, in this study 500-800 candidate papers are selected based on large C-Score values. Table 2 shows candidate score calculation of papers $P_{10}$, $P_{25}$ and $P_{31}$ which are shown in Figure. 3. $P_{25}$ has co-citation value 6 [($P_{23}$, $P_{25} \leftarrow P_{20}$), ($P_{26}$, $P_{25} \leftarrow P_{20}$), ($P_{24}$, $P_{25} \leftarrow P_{21}$), ($P_{24}$, $P_{25} \leftarrow P_{22}$), ($P_{18}$, $P_{25} \leftarrow P_{22}$), ($P_{19}$, $P_{25} \leftarrow P_{22}$)] and bibliographic coupling value 2 [($P_{20}$, $P_{25} \rightarrow P_{26}$), ($P_{23}$, $P_{25} \rightarrow P_{27}$)] and its distance is 3 which is the number of links between paper I and $P_{25}$. Total similarity of both papers $P_{25}$ and $P_{10}$ is same but $P_{25}$ has low value of C-Score because it is farther from paper I which means $P_{10}$ is more similar to paper I as compared to $P_{25}$. Although $P_{10}$ and $P_{31}$ have same distance from the paper of interest I, but the C-Score value of $P_{31}$ is lower than that of $P_{10}$ due to the total similarity of paper $P_{10}$ greater than the value of $P_{31}$. The papers having low value of C-Score are removed from the network. C-Score value of each paper is calculated in this way to inspect the relation of each paper with the paper of interest.

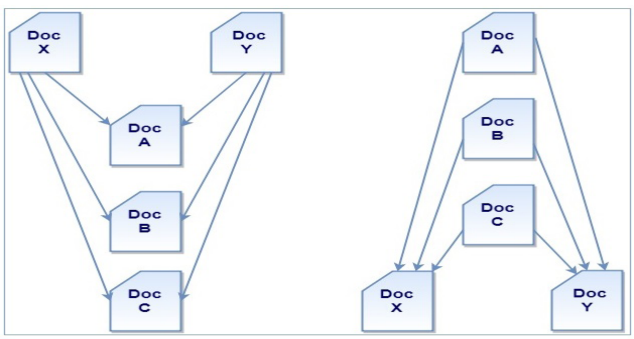



**FIGURE 4.** Example of bibliographic coupling and co-citation analysis.

### C. IDENTIFICATION OF SIGNIFICANT PAPER FROM RELEVANT PAPERS

Centrality measures are used for evaluating importance of each paper. The importance of each paper is evaluated based on its relationship with other papers [30] in the network. A range of centrality measures are applied on candidate papers. These include degree centrality, closeness centrality, betweenness centrality and eigenvector centrality. Degree centrality is the simplest way of finding important nodes. It calculates number of neighbors of each node with a node having more neighbors is considered to have greater influence. The formula for calculating degree centrality is shown in Equation 4.

$$C_{(D)}(P) = \frac{d(P)}{n-1} \tag{4}$$

where *d(P)* represents the number of papers referring to paper $P$ and $n$ are the total number of papers. To calculate significance, only the in-degree centrality is considered in this work. In closeness centrality, a paper is considered central and important if it is linked with many other nodes. The formula for calculation of closeness centrality is shown in Equation 5.

$$C_{(c)}(P) = \frac{n-1}{\sum_{j \neq T} d(P, J)} \tag{5}$$

where $n$ is the total number of papers and $d\ (P,\ J)$ defines the distance between paper $P$ and $J$. Similarly, *n-1* defines the minimum distance of paper to all other adjacent papers. Betweenness centrality is the number of shortest paths that pass through any particular node in the network. The formula for calculating betweenness centrality is given in Equation 6.

$$C_{(B)}(P) = \sum_{j \neq V \neq P} \frac{g_{jv}(P)}{g_{jv}} \tag{6}$$

where the metric $g_{jv}$ provides the number of links that pass through shortest route and $g_{jv}(P)$ shows the number of links that pass through paper $P$. The last centrality measure, Eigenvector centrality is used for measuring influence of a node in the network. It is a variant of PageRank algorithm and measures importance of node based on referral of other important nodes in the network. The formula for calculating eigenvector centrality is provided in Equation 7.

$$C_{(E)}(P) = \frac{1}{\Lambda} \sum_{j=B_P} A_{P,J} X_J \tag{7}$$

where $A_{P,J}$ is the adjacency matrix and $X_j$ is the score and is the eigenvalue. Calculate the average rank of each paper by using Equation 8.

$$AR(P) = \frac{\sum_{k=1}^{M} rank^k(P)}{M} \tag{8}$$

where M is the total centrality measure and *rank* is the ranking result on paper $P$. To set the rank in the same range, centrality measures are scaled in the range (1:50). The algorithm for extraction of top papers is shown in Algorithm 3. It takes a list of candidate papers as input and returns the ranked list of top papers with respect to the centrality measures.

**Algorithm 3** Extraction of Top Papers

**Input** CANDIDATE PAPERS
**Output** TOP PAPERS
create an array of top papers
iterate over the list of candidate papers
calculate Degree Centrality of each paper by using Equation 4
calculate Closeness Centrality of each paper by using Equation 5
calculate Betweenness Centrality of each paper by using Equation 6
calculate Eigenvector Centrality of each paper by using Equation 7
convert all Centrality measures to Rank
calculate Average Rank of each paper among all four centrality measures by using Equation 8
add papers to top papers list
return top papers

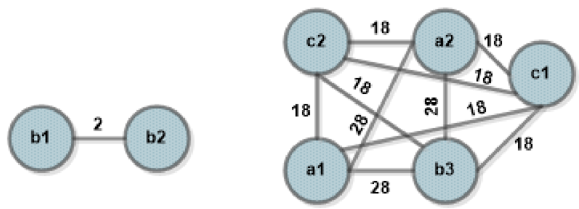



**FIGURE 5.** Example of Relationship Network of Authors in which Nodes a1, a2, b1, b2, b3, c1, c2 represents Authors and Edges represents relationship Score between them.

### D. GENERATING AUTHORŚ RELATIONSHIP NETWORK FROM SIGNIFICANT PAPERS

In this step, the relationship network of authors is extracted from significant papers from top papers. It is generated by placing a link between those authors who co-author one or more articles [31]. Figure 5 shows a sample of such network with frequency of the co-authorship is placed on the links connecting them. The relationship network is created using a network matrix generated by the following approach: Let's say, there are three papers *P1*, *P2* and *P3* in which two papers *P1* and *P2* are journal papers and *P3* is a conference paper. *P1* has three authors *a1*, *a2* and *b3* with ten citations and *P2* has two authors *b1* and *b2* with two citations. Similarly, *P3* has five authors *a1*, *a2*, *b3*, *c1*, *c2* with eighteen citations. Figure. 5 shows the output network where the authors are represented by nodes and are linked using the weighted links a1-a2, a1-b3, a2-b3, b1-b2 and a1-a2, a1b3, a1-c1, a1-c2, a2-b3, a2-c1, a2-c2, b3-c1, b3-c2, c1-c2.

The algorithm for authors selection from top papers is shown in Algorithm 4. The input to this algorithm are the top papers and it returns list of authors along with their co-authorship frequency.

### E. KEY AUTHOR ANALYSIS OF RELATIONSHIP NETWORK

The objective of this step is to find a set of key authors from a relationship network of authors. This network returns

**Algorithm 4** Authors Selection From Top Papers

**Input** TOP PAPERS
**Output** LIST of AUTHORS
create array of authors
iterate over a list of top papers
    select all authors from each top paper
    add authors to authors list
return authors list

**Algorithm 5** Ranking of Authors

**Input** LIST of AUTHORS
**Output** TOP AUTHORS
create an array of top authors
iterate over the list of authors
    calculate citations of authors
    calculate collaboration score of authors by applying four centrality measures
    convert author scores to rank
    add top authors to top authors list
return top authors

authors based on citation count and four centrality measures including closeness centrality, betweenness centrality and eigenvector centrality. The algorithm for ranking of authors is shown in Algorithm 5. List of authors from Section III-D are sent as input and it returns list of top authors based on citation count of an author and centrality measures.

In citation count, the frequency measure of the citations is evaluated. It is the number of times other people have referred research articles of an author. For Example, when an article has three authors and eight citations then each author has a citation count of eight. Closeness centrality helps in determining the closeness of any author with other authors in the network. An author with more number of coauthor relationships in the network is considered as a key author because of having high value of closeness centrality. Betweenness centrality is also an important factor for finding key authors. An author having high value of this metric connects researchers from two different sub-networks and hence becomes an important candidate to be considered as a key author. Eigenvector centrality also plays a vital role in finding key authors. An author linked with other key authors in the network having high eigenvector centrality is also a key author. The papers authored by ranked authors are then considered as high quality papers for recommendation.

### F. RECOMMEND TOP N PAPERS TO USERS

Finally, papers published by key authors are selected which are identified from relationship network of authors. Papers are sorted according to high eigenvector values of author and top 10 papers are recommended to users. According to the literature recommending more than items will confuse users [32]. The algorithm for determination of recommend papers is shown in Algorithm 6. It performs relatively straightforward task of sorting papers with respect to the eigenvalue and returns top ten articles in the list as a recommendation to the user.

**Algorithm 6** Determination of Recommend Papers

**Input** TOP AUTHORS
**Output** RECOMMENDED PAPERS
create an array of recommended papers
iterate over the list of top authors
    select papers of top authors and append them to recommended papers list
return recommended papers[n]

## IV. EXPERIMENTAL SETTING

There are several approaches used in the literature for measuring accuracy and user satisfaction of recommender systems [33], [34]. The 69% of these approaches use offline methods while the rest use online methods of evaluation. Offline methods make use of existing datasets which are already being used by others and are considered standard datasets among the research community. While online methods generate new dataset on the fly and are considered to be a time consuming as evaluator has to wait for days or week for the results. Offline methods are thus considered as a more reliable approach as they can reproduce the same setting of experiment for different evaluations. This also increases the consistency level of offline methods when compared with online methods as its result for variety of tasks can be setup for comparison. This work is also evaluated using offline evaluation techniques.

The AMiner dataset[1] is used for evaluation of the proposed approach [35]. It contains paper information, author information and citation information. The dataset is organized into three files as shown in Table 4. AMiner-paper file contains information about 2,092,356 research papers which are published over the years and 8,024,869 citations exist between them. AMiner-author contains information about 1,712,433 authors, whereas 4,258,615 collaborations are stored in AMiner-coauthor file. Figure 6 and Figure 7 provides the snapshots of the sample subset of AMiner dataset. Figure. 6 shows the citation network using Gephi tool where nodes of the graph represent paper ids and edges between them represent citations of the paper. Figure. 7 provides snapshot of the collaboration network where nodes represent author ids and edges between them represent co-authorship relation between them.

The proposed CNRN recommender system is evaluated using Information Retrieval metrics, namely Normalized discounted cumulative gain (NDCG), Mean Reciprocal Rank (MRR) and Average Precision (AP) [36]. These metrics are extensively used in the literature for measuring the performance of the ranking algorithms. NDCG is the average measure of the graded relevance of recommended docu-

[1] https://aminer.org/billboard/aminernetwork

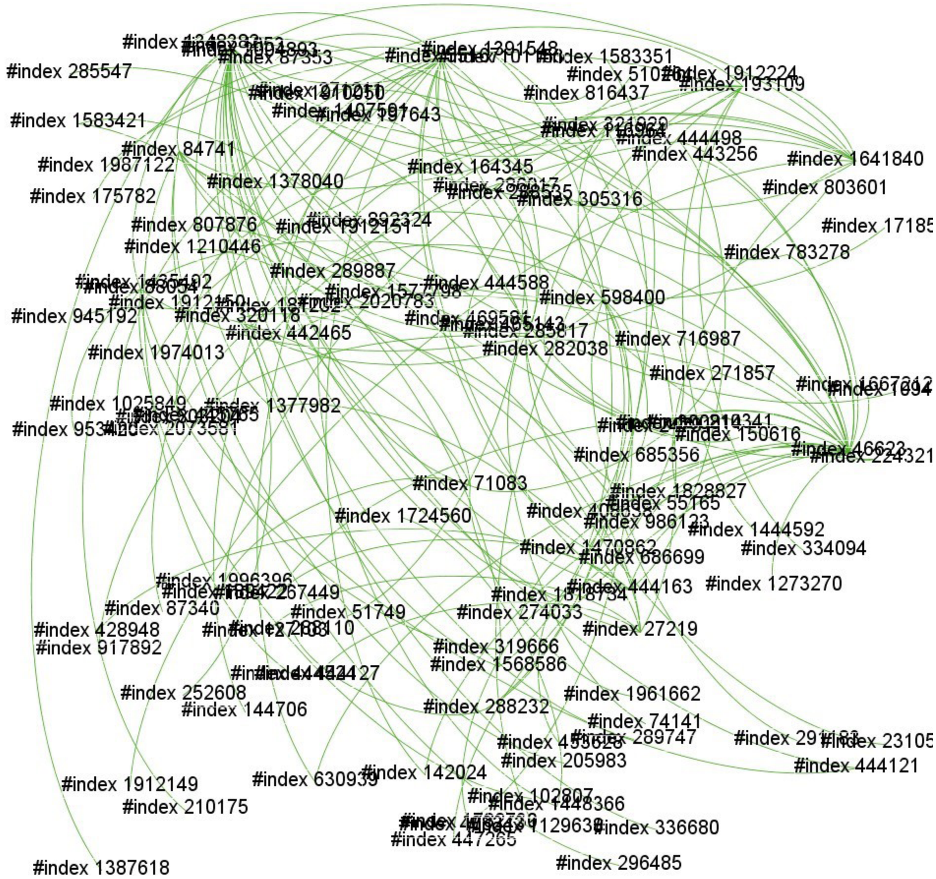



**FIGURE 6.** Citation network analysis using Gephi where nodes represents paper's id and edges represents citations of paper.

ments. It assesses the extent to which the ranked set of recommendations are near to the ideal ranking of the recommendations. The value of NDCG is calculated using Equation 9.

$$NDCG_p = \frac{DCG_p}{IDCG_p} \tag{9}$$

where $NDCG_p$ is the normalized gain accumulated at a particular rank p. DCG stands for Discounted Cumulative Gain and is the weighted sum of the degree of relevancy of the ranked items. Its value is calculated using Equation 10.

$$DCG_p = \sum_{i=1}^{P} \frac{2^{rel_i} - 1}{log_2(i+1)} \tag{10}$$

where $DCG_p$ represents total accumulated gain at a particular rank p and $rel_i$ is the graded relevance of the recommended article at a particular rank p. NDCG normalizes DCG using the Ideal Cumulative Discounted Gain (IDCG) which is the DCG measure of the best ranking result [37]. Hence, the value of NDCG always lies between 0 and 1 and its value for the perfect recommendation will be one. The value of IDCG is calculated using Equation 11.

$$IDCG_p = \sum_{i=1}^{P} \frac{1}{log(i+1)} \tag{11}$$

Traditional NDCG metric used for evaluation in [13] has two drawbacks; firstly, it ignores missing documents in the result and secondly, it does not consider irrelevant documents. To describe it for the first case, for example, two results of a query having score of 1, 1, 1, 0 and 1, 1, 0 are considered equally good while 1, 1, 0 has a missing document which is not reflected in the result (here, 1 represents valid result if the user is satisfied with the paper and 0 represents irrelevant result.).

To overcome this limitation, number of results should be fixed and a result with missing document should be replaced with 0 in the result set. So, for the scenario described above the result with missing document would be 1,1,1,0 and 1,1,0,0 instead of 1, 1, 1, 0 and 1, 1, 0. Secondly, for example, the results 1,1,1 and 1,1,1,0 are considered equally good by NDCG as they both contain three relevant document while the 1,1,1,0 contains bad document represented by 0 in the result. To address this limitation, we use numerical values 0, 1, -1 for ranking judgments excellent, fair and bad respectively. This

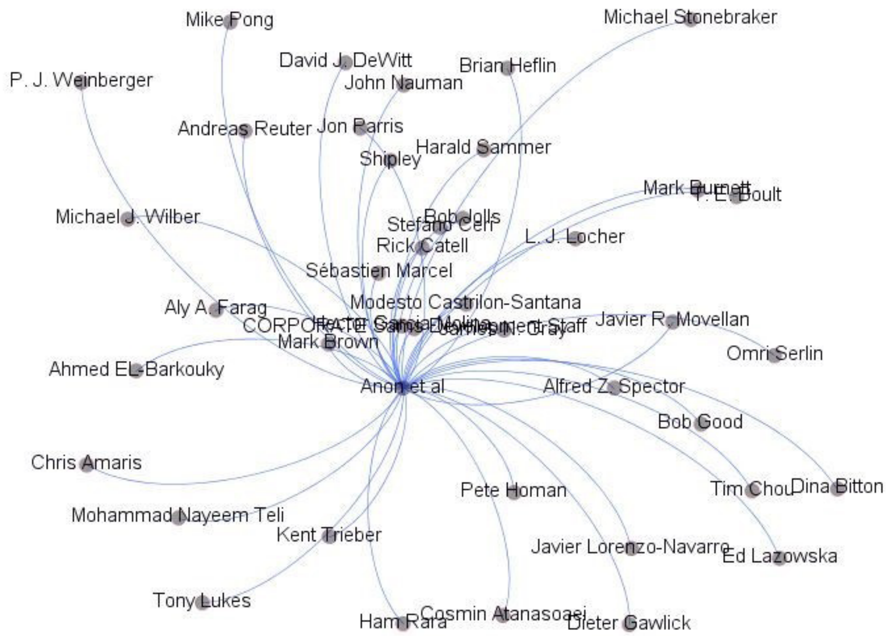



**FIGURE 7. Collaboration network analysis using Gephi where nodes represents author's id and edges represents collaboration relationship between them.**

will translate above example as 1,1,1,0 and 1,1,1, -1 which will impact the result with negative value ultimately reflecting the case of a bad document in the result.

MRR evaluates the recommendation system based on relevant item at the top of a ranked list. It is calculated by using Equation 12.

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank_i} \tag{12}$$

where the $n$ represents the number of users and $rank_i$ shows position of the first correct item in the ranked list.

The third evaluation metric, AP is the proportion of the relevant documents from the set of recommended documents. It is calculated using Equation 13.

$$AP = \frac{1}{m}\sum_{1\leq k\leq n} rel(k).Prec@k \tag{13}$$

where $m$ is the number of relevant articles in the list, $n$ is the total number of recommended articles and *rel(k)* is the relevance information. Prec@k of each relevant item from the ranked result set is the proportion of top $K$ relevant documents of recommender system. In this work, the value of $K$ resides in the range 1 to 10 [38].

The sample research papers used for experimentation in this work are shown in Table 3. These papers represent papers of interest and their selection is based upon different features of research articles that can impact the results of different recommendation approaches. For example, papers 1 and 3 belong to the category of eminent papers due to their high citation count. Paper 4 is selected when results are to be compared for a recent paper chosen as the paper of interest. Similarly paper 2 represents the scenario when an old paper is selected as the paper of interest.

The ranked results provided by CNRN algorithm along with those from Google Scholar and MSCN are presented to experts for evaluation. A total of twenty researchers evaluated recommended papers and title, authors and year of publication are provided to experts. The researches were to evaluate the results from each of the CNRN, Google Scholar and MSCN approaches based on their satisfaction from each of these approaches.

## V. RESULTS AND ANALYSIS

The system generates NDCG, MRR and AP metrics for each of the CNRN, Google Scholar and base method MSCN and Figures 8 and 10 present the results of each technique respectively. Their x-axis represent each of the recommended papers whereas y-axis is the NDCG and MRR metrics. The results revealed that the proposed CNRN approach outperformed Google Scholar and MSCN when eminent papers 1 and 3 are selected as papers of interest. For each of the ten papers recommended by all the approaches, NDCG metric for CNRN is better than both of the existing approaches Google Scholar and MSCN. The reason being that Google scholar recommends paper purely based on the citation count which makes its recommendations less relevant to the paper of interest and hence evaluation by the experts declared these recommendations as insignificant. On the other hand, MSCN approach generates citation network of papers and hence recommends papers which are more similar to the paper of interest. As a result, experts considered recommendations by the MSCN approach to be more useful

**TABLE 3.** Papers used for evaluation and experimentation.

| PaperID | Title | Author | Year | Source | Cited by |
|---|---|---|---|---|---|
| 1 | Wireless insecurity: examining user security behavior on public networks | Tim Chenoweth, Robert Minch, Sharon Tabor | 2010 | Communications of the ACM | 19 |
| 2 | Software engineering project management, estimation and metrics: discussion summary and recommendations | J. Paynter | 1996 | Proceedings 1996 International Conference Software Engineering: Education and Practice | 5 |
| 3 | A Transaction Cost Model of Electronic Trust: Transactional Return, Incentives for Network Security and Optimal Risk in the Digital Economy | VF Kleist | 2004 | Electronic Commerce Research, Springer | 34 |
| 4 | Research Paper Recommender Systems on Big Scholarly Data | Tsung Teng Chen, Maria Lee | 2018 | Knowledge Management and Acquisition for Intelligent Systems | 0 |

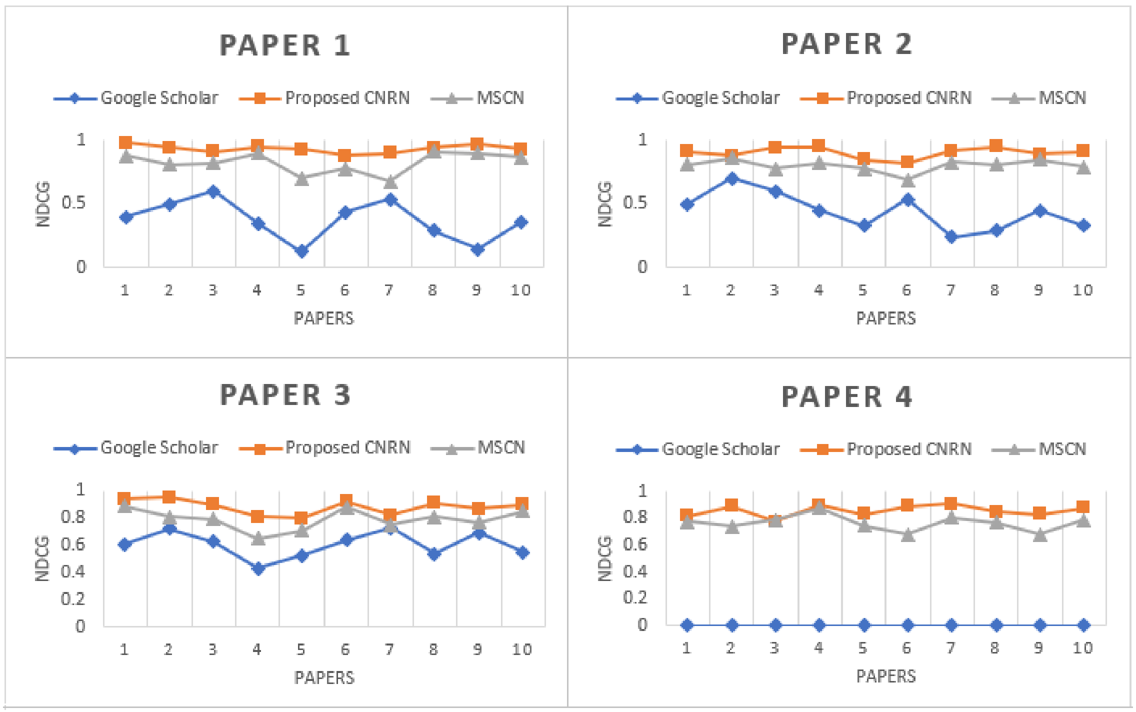



**FIGURE 8.** Performance comparison of proposed CNRN with GOOGLE SCHOLAR and MSCN using NDCG metric.

as compared to the Google Scholar. When the results of our proposed approach CNRN are compared with those from the MSCN approach, experts declared them of even better quality than the MSCN approach. This is due to incorporating author network in addition to the citation network which makes its recommendations more suitable than those from the MSCN approach.

When the results are compared for paper 4, which belongs to the category of recent papers, it can be seen that Google Scholar performed worst and its evaluation by experts is completely unsatisfactory. The reason being that Google Scholar tends to recommend irrelevant papers because paper 4 has citation count of zero. On the other hand, MSCN performs better than Google Scholar because it considers citation

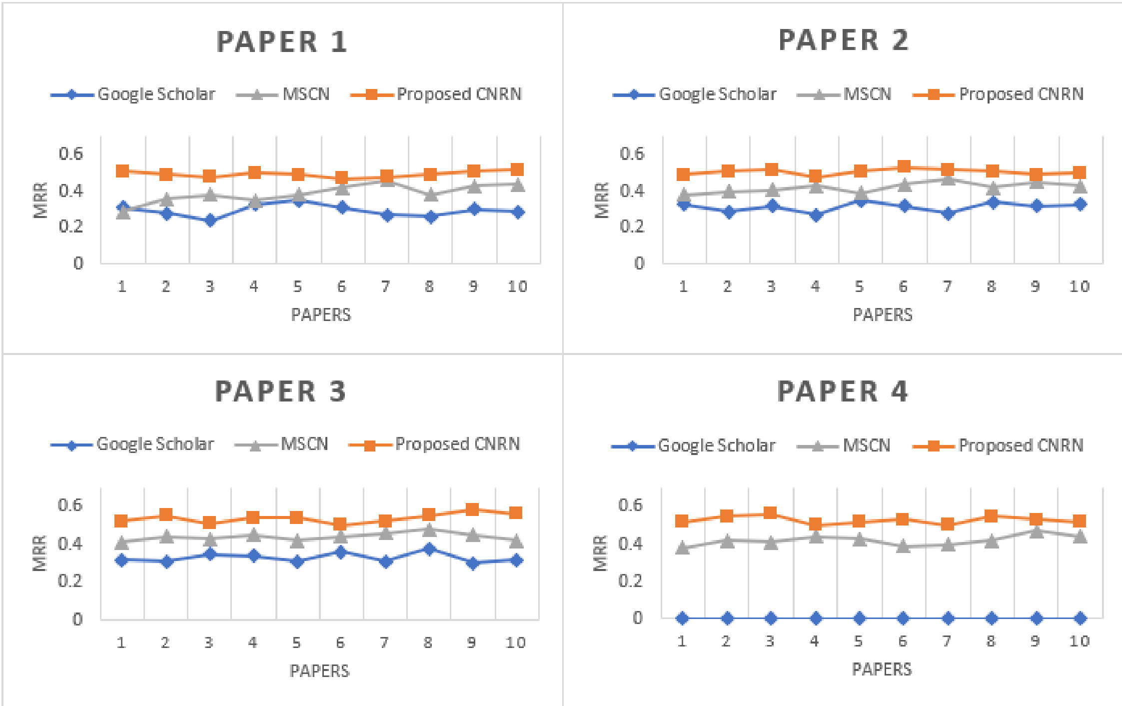



**FIGURE 9.** Performance comparison of proposed CNRN with GOOGLE SCHOLAR and MSCN using MRR metric.

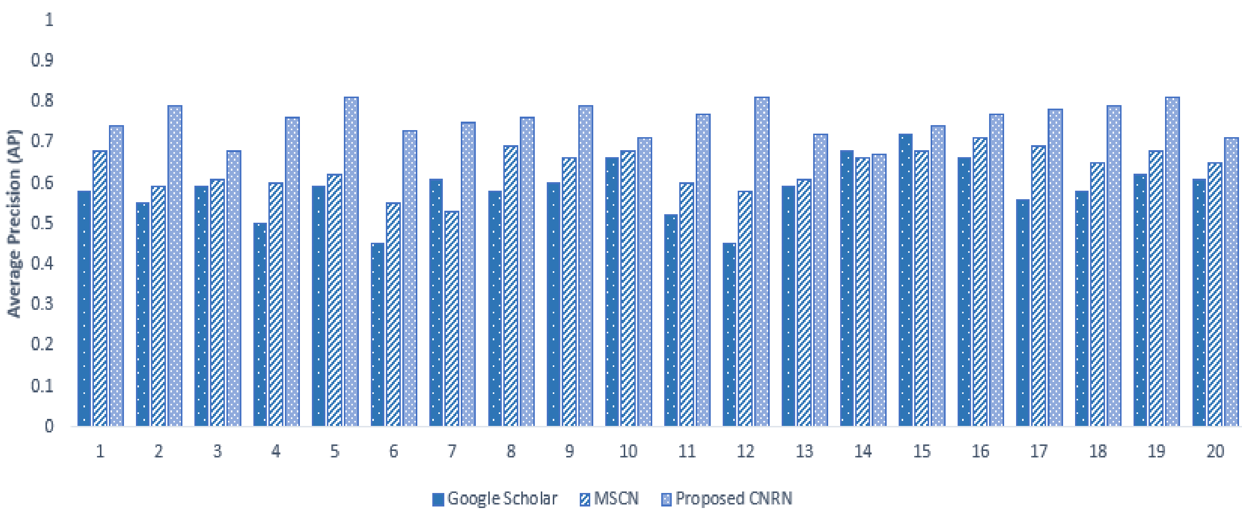



**FIGURE 10.** Performance comparison of proposed CNRN with GOOGLE SCHOLAR and MSCN using average precision metric.

network which makes it to recommend relevant papers even for the recent paper. The proposed CNRN approach outperformed both of the approaches and its recommendations for recent paper are also best than the other two approaches. The high quality recommendations by CNRN approach are due to considering author network in addition to the citation network which further improves the results. When an old paper is selected as paper of interest such as paper 2, Google Scholar again fails to recommend high quality papers due to its inherent problem of considering only citation count. As a result, it recommends outdated papers which are declared by expert as unsatisfactory.

Figure 10 presents the comparison of average precision (AP) metric for proposed CNRN approach with Google Scholar and MSCN techniques. X-axis shows each of the twenty experts and y-axis is their AP measure for the three approaches selected for comparison. Unlike NDCG and MRR which provide measure of graded relevance for each of the

**TABLE 4. ArnetMiner Dataset Description [35].**

| FileName | Node | Number |
|---|---|---|
| AMiner-Paper.rar | Paper/citation | 2,092,356/8,024,869 |
| AMiner-Author.zip | Author | 1,712,433 |
| AMiner-Coauthor.zip | Collaboration | 4,258,615 |

recommended article individually, AP is the average measure of rating provided by individual experts for each of the ten recommended articles. The ideal scenario would be if all the recommended papers are relevant. Alternatively, some of the results may be relevant while others are not. The results show that none of the CNRN, Google Scholar and MSCN has AP metric of one. However, the performance of the proposed CNRN is significantly better than existing approaches. The evaluations by almost all the experts stand best for CNRN than Google Scholar and MSCN approaches. For experts 5 and 12, it is widely better whereas for expert 14 Google Scholar is marginally better than CNRN. These results imply that the recommendation using CNRN approach largely satisfied evaluators than both of the existing techniques.

Based on the analysis performed above, we can claim that the proposed CNRN approach outperformed existing approaches. The papers of interest selected for this experiment belonged to different categories. The CNRN approach generated recommendations which satisfied experts for a range of these categories. While existing approaches failed to made any impact on the evaluators for any of the category. The existing approach Google Scholar lacks when either recent or old paper is presented as a paper of interest while MSCN approach also fails to recommend quality papers. The proposed CNRN approach overcomes these drawbacks and performs multi-layer filtering before recommending papers.

The analysis conducted in this papers uses AMiner dataset. As part of the future work, it can be repeated on other datasets or can be tested using online methods. Furthermore, it would be interesting to see how incorporating additional features such as journal information can impact the recommendations.

## VI. CONCLUSION

The use of recommender systems for extracting related papers have become vital due to the recent challenge of handling big data. The state-of-the-art Google Scholar and other existing approaches recommend papers, but they have drawbacks when either new or old papers are selected as paper of interest. The quality of recommended articles is also compromised as only citation counts or the relationship among papers is considered as a metric. The proposed CNRN approach overcomes the limitation by including an additional measure of finding key authors other than creating a co-citation network. Candidate score is calculated for each paper by using co-citation, bibliographic coupling and centrality measure metrics from generated graph. The set of papers selected using centrality measures are then fed into the author ranking module which calculates centrality measures for author network. Authors having high eigenvector values are selected and papers published by those authors are recommended to the users.

The proposed CNRN approach was compared with both benchmark approaches Google Scholar and MSCN using NDCG, MRR and AP metrics. The results revealed that the CNRN approach outperformed both of the existing approaches in recommending related papers. It recommended high quality papers irrespective of the citation count or publication date of the paper of interest.

## REFERENCES

[1] S. Mukherjee, D. M. Romero, B. Jones, and B. Uzzi, "The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot," *Sci. Adv.*, vol. 3, no. 4, 2017, Art. no. e1601315.

[2] C. Chen, "Searching for intellectual turning points: Progressive knowledge domain visualization," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. suppl 1, pp. 5303–5310, 2004.

[3] M. A. Domingues, A. M. Jorge, and C. Soares, "Dimensions as virtual items: Improving the predictive ability of top-*N* recommender systems," *Inf. Process. Manage.*, vol. 49, no. 3, pp. 698–720, 2013.

[4] W. Zhao, R. Wu, and H. Liu, "Paper recommendation based on the knowledge gap between a researcher's background knowledge and research target," *Inf. Process. Manage.*, vol. 52, no. 5, pp. 976–988, Sep. 2016.

[5] W. Huang, Z. Wu, P. Mitra, and C. L. Giles, "RefSeer: A citation recommendation system," in *Proc. IEEE/ACM Joint Conf. Digit. Libraries (JCDL)*, Sep. 2014, pp. 371–374.

[6] O. Kassak, M. Kompan, and M. Bielikova, "User preference modeling by global and individual weights for personalized recommendation," *Acta Polytechn. Hungarica*, vol. 12, no. 8, pp. 27–41, 2015.

[7] R. Dong, L. Tokarchuk, and A. Ma, "Digging friendship: Paper recommendation in social network," in *Proc. Netw. Electron. Commerce Res. Conf. (NAEC)*, 2009, pp. 21–28.

[8] C. Bhagavatula, S. Feldman, R. Power, and W. Ammar. (2018). "Content-based citation recommendation." [Online]. Available: https://arxiv.org/abs/1802.08301

[9] N. J. van Eck and L. Waltman, "Citation-based clustering of publications using CitNetExplorer and VOSviewer," *Scientometrics*, vol. 111, no. 2, pp. 1053–1070, 2017.

[10] D. Yu, Z. Xu, W. Pedrycz, and W. Wang, "Information sciences 1968–2016: A retrospective analysis with text mining and bibliometric," *Inf. Sci.*, vols. 418–419, pp. 619–634, Dec. 2017.

[11] B. Gipp, J. Beel, and C. Hentschel, "Scienstein: A research paper recommender system," in *Proc. Int. Conf. Emerg. Trends Comput. (ICETIC)*, 2009, pp. 309–315.

[12] J. Beel and B. Gipp, "Google Scholar's ranking algorithm: An introductory overview," in *Proc. 12th Int. Conf. Scientometrics Informetrics (ISSI)*, Rio de Janeiro, Brazil, vol. 1, 2009, pp. 230–241.

[13] J. Son and S. B. Kim, "Academic paper recommender system using multilevel simultaneous citation networks," *Decision Support Syst.*, vol. 105, pp. 24–33, Jan. 2018.

[14] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: A literature survey," *Int. J. Digit. Libraries*, vol. 17, no. 4, pp. 305–338, 2016. doi: 10.1007/s00799-015-0156-0.

[15] S. M. McNee *et al.*, "On the recommending of citations for research papers," in *Proc. ACM Conf. Comput. Supported Cooperat. Work*, 2002, pp. 116–125.

[16] S. Doerfel, R. Jäschke, A. Hotho, and G. Stumme, "Leveraging publication metadata and social data into FolkRank for scientific publication recommendation," in *Proc. 4th ACM RecSys workshop Recommender Syst. Social Web*, 2012, pp. 9–16.

[17] M. T. Afzal, N. Kulathuramaiyer, and H. A. Maurer, "Creating links into the future," *J. Universal Comput. Sci.*, vol. 13, no. 9, pp. 1234–1245, 2007.

[18] N. Ratprasartporn and G. Ozsoyoglu, "Finding related papers in literature digital libraries," in *Proc. Int. Conf. Theory Pract. Digit. Libraries*. Berlin, Germany: Springer, 2007, pp. 271–284.

[19] Y. Ding, G. Zhang, T. Chambers, M. Song, X. Wang, and C. Zhai, "Content-based citation analysis: The next generation of citation analysis," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 9, pp. 1820–1833, 2014.

[20] X. Y. Leung, J. Sun, and B. Bai, "Bibliometrics of social media research: A co-citation and co-word analysis," *Int. J. Hospitality Manage.*, vol. 66, pp. 35–45, Sep. 2017.

[21] C. Biscaro and C. Giupponi, "Co-authorship and bibliographic coupling network effects on citations," *PLoS ONE*, vol. 9, no. 6, p. e99502, 2014.

[22] B. Kaya, "User profile based paper recommendation system," *Int. J. Intell. Syst. Appl. Eng.*, vol. 6, no. 2, pp. 151–157, 2018.

[23] K. Sugiyama and M.-Y. Kan, "Scholarly paper recommendation via user's recent research interests," in *Proc. 10th Annu. Joint Conf. Digit. Libraries*, 2010, pp. 29–38.

[24] K. D. Bollacker, S. Lawrence, and C. L. Giles, "CiteSeer: An autonomous Web agent for automatic retrieval and identification of interesting publications," in *Proc. ACM 2nd Int. Conf. Auto. Agents (AGENTS)*, New York, NY, USA, 1998, pp. 116–123. doi: 10.1145/280765.280786.

[25] H. Liu, X. Kong, X. Bai, W. Wang, T. M. Bekele, and F. Xia, "Context-based collaborative filtering for citation recommendation," *IEEE Access*, vol. 3, pp. 1695–1703, 2015.

[26] K. Haruna, M. A. Ismail, D. Damiasih, J. Sutopo, and T. Herawan, "A collaborative approach for research paper recommender system," *PLoS ONE*, vol. 12, no. 10, 2017, Art. no. e0184516.

[27] J. Beel and B. Gipp, "Google Scholar's ranking algorithm: The impact of citation counts (an empirical study)," in *Proc. IEEE 3rd Int. Conf. Res. Challenges Inf. Sci. (RCIS)*, Apr. 2009, pp. 439–446.

[28] L. Bornmann and L. Leydesdorff, "Topical connections between the institutions within an organisation (institutional co-authorships, direct citation links and co-citations)," *Scientometrics*, vol. 102, no. 1, pp. 455–463, 2015.

[29] H. Zaugg, R. E. West, I. Tateishi, and D. L. Randall, "Mendeley: Creating communities of scholarly inquiry through research collaboration," *TechTrends*, vol. 55, no. 1, pp. 32–36, 2011.

[30] T. Opsahl, F. Agneessens, and J. Skvoretzc, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social Netw.*, vol. 32, no. 3, pp. 245–251, 2010.

[31] A. Bihari and M. K. Pandia, "Key author analysis in research professionals' relationship network using citation indices and centrality," *Procedia Comput. Sci.*, vol. 57, pp. 606–613, 2015.

[32] M. Deshpande and G. Karypis, "Item-based top-n recommendation algorithms," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 143–177, 2004.

[33] J. Beel, S. Langer, M. Genzmehr, B. Gipp, C. Breitinger, and A. Nürnberger, "Research paper recommender system evaluation: A quantitative literature survey," in *Proc. Workshop Reproducibility Replication Recommender Syst. Eval. (RepSys) ACM Recommender Syst. Conf. (RecSys)*, Hong Kong, 2013, pp. 15–22.

[34] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: Introduction and challenges," in *Recommender Systems Handbook*. Boston, MA, USA: Springer, 2015, pp. 1–34.

[35] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: Extraction and mining of academic social networks," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 990–998.

[36] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in *Proc. 2nd ACM Int. Conf. Web Search Data Mining*, 2009, pp. 5–14.

[37] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.

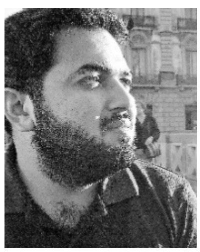
[38] A. Otsuka, K. Nishida, K. Bessho, H. Asano, and J. Tomita, "Query expansion with neural question-to-answer translation for FAQ-based question answering," in *Proc. Companion Web Conf. Web Conf.* Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2018, pp. 1063–1068.

**WALEED WAHEED** was born in Rawalpindi, Pakistan. He received the B.S. degree in software engineering from International Islamic University, Islamabad, Pakistan. He is currently pursuing the master's degree in computer science with COMSATS University Islamabad, Islamabad, with a specialization in computer science. His research interests include information retrieval, databases, data warehouses, semantic Web technologies, and machine learning.

**MUHAMMAD IMRAN** received the bachelor's degree in software engineering from the University of Engineering and Technology at Taxila, Taxila, Pakistan, in 2006, and the master's degree in software engineering and the Ph.D. degree in computer science from the University of Southampton, U.K., in 2009 and 2015, respectively, after securing the Faculty Development Scholarship from CIIT. He was a Lecturer with CIIT, Islamabad, Pakistan, from 2007 to 2008. He is currently an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad, Islamabad. His research interests include social networks, data mining and artificial intelligence, developing trust solutions for online social networks, semantic Web technologies, and the Internet of Things.

**BASIT RAZA** received the Ph.D. degree in computer science, in 2014. He is currently an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad, Islamabad, Pakistan. He has published a number of conference and journal papers of internal repute. His research interests are database management systems, security and privacy, data mining, data warehousing, machine learning, and artificial intelligence.

**AHMAD KAMRAN MALIK** received the M.S. degree in computer science from Muhammad Ali Jinnah University, Islamabad, and the Ph.D. degree from the Vienna University of Technology, Austria. He has been teaching at Quaid-I-Azam University. He is currently an Assistant Professor with COMSATS University Islamabad, Islamabad, Pakistan. His current research interests are social network analysis, access control, and collaborative systems.



**HASAN ALI KHATTAK** (SM'19) received the B.C.S. degree in computer science from the University of Peshawar, Peshawar, Pakistan, in 2006, the master's degree in information engineering from the Politecnico di Torino, Turin, Italy, in 2011, and the Ph.D. degree in electrical and computer engineering from the Politecnico di Bari, Bari, Italy, in 2015. He has been an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad, since 2016. His current research interests include the Web of Things, data sciences, and social engineering for future smart cities. His perspective research areas are the application of machine learning and data sciences for improving and enhancing quality of life in Smart Urban Spaces through predictive analysis and visualization. Along with publishing in good research venues and completing successful funded national and international funded projects, he is also serving as a Reviewer in reputed venues, such as the IEEE Access, the *IEEE Network Magazine*, the *IEEE Consumer Electronics*, Hindawi, SAI, IET, and a few national publishers. He is currently involved in several funded research projects in various domains, such as semantic Web of Things and fog computing while exploring ontologies, Web technologies using Contiki OS, NS 2/3, and Omnet++ frameworks. He is an active member of the IEEE ComSoc, the IEEE VTS, and the Internet Society.