International Journal of Decision Support System Technology
Volume 14 • Issue 1

Infgraph: Influential Researcher and Cited Research Analysis Using Citation Network

M. Geetha., Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences Chennai, Tamilnadu, India

K. Suresh Kumar, Department of Information Technology, Saveetha Engineering College, Chennai, Tamilnadu, India

Ch. Vidyadhari, Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

R. Ganeshan, School of Computer Science and Engineering, VIT Bhopal University, Madhya Pradesh, India

ABSTRACT

The understanding of references in research articles is essential for performing effectual research. This paper devises a hybrid model to find the influential cited paper and influential researchers from Web of Science (WOS) data. For determining the influential researcher, a series of steps is performed. Then the co-citation is performed for providing author-author co-relation that predicts the next co-author. Thereafter, visualization of the network is performed for research communication amongst different authors. Then, the network density is computed. Finally, the cluster coefficient is adapted for finding the influential researcher. Concurrently, for discovering influential cited papers, the pre-processing is performed using the stop word removal and stemming process. Then, the word2vec model is utilized for training the model to forecast the suitable word that comes next. Finally, the modified word mover's distance (MWMD) is utilized for determining the semantic similarity in order to discover influential cited papers.

KEYWORDS

Citation Analysis, Citation Network, Influenced Citations, Network Density, Word Movers Distance

1. INTRODUCTION

Citation analysis is a useful technique in the assessment of the impact of an article and discovers the significant articles in a certain field. Also, it has become an integral part of the decision making procedure in scientific and academic life as a foundation for ranking institutions, journals, authors, and even for creating promotion decisions. The speediness and development scope in these areas made it crucial for the researchers to be conscious of data published across various research domains and various organizations (Kajikawa & Takeda, 2009). The citation network of the paper helps to depict the cited relation amongst various papers (Thakur, 2018; Thakur, 2017) which has offered a

DOI: 10.4018/IJDSST.311065 *Corresponding Author

promising way for modeling the correlation amongst the papers (Liu, *et al.,* 2019). In social networks, the actors connect with the interactions for exchanging valued resources. The citation network is an explicit social network wherein the actors indicate articles, journals, and authors. In this method, the valued resources indicate knowledge and ideas and the interactions represent the citation of actors (Pieters, *et al.,* 1999).

The assessment of scholarly communication by citation patterns is widely utilized for determining the scientific teamwork that mapped scholarly disciplines and assessed the influence of investigation outcomes and scrutinized knowledge transport amongst several areas. The citations are utilized for establishing a network that involves author citation networks in which the node indicates papers, journals, and authors and the edge signifies the count of papers that are cited, and co-cited. Even though the easy count of citation remained an imperative indicator, but it's limited to semantics (Ding, *et al.,* 2014). The impact factor of scientific publications is evaluated by considering the count of citations they receive. It depicts how regularly they are being referenced by other publications. This publication has linked authors that originate institutions and venues of publication that involves conference, journals, and proceedings for comparing the scientific impact. For example, one frequently adapted indicator is the impact factor which helps to determine the journal quality. The impact factor is yearly published using Journal Citation Report (JCR) by adapting citations (Rahm & Thor, 2005). The adaption of citation analysis is to provide an evaluation of research, which aimed to evaluate the altered scholarly work contribution considering knowledge advancements. The scientists help to cite the work as they discover it to be most beneficial for perusing the research. The count of citations received by the publication is considered a quantitative measure of resonance and has built a scientific community (Neuhaus & Daniel, 2008).

The citation data are utilized for evaluating several applications like the researcher's credibility and institutions and so on. There are various benchmark citation measures utilized for determining the researcher's strength (Muppidi, *et al.,* 2018). An understanding of references from the research article is essential for performing better research (Muppidi & Thammi Reddy, 2020). The publication of scientific journals is an imperative task in research domains. The scientific articles are utilized for reporting academic goals so they are considered an imperative part of developing the scientific research. Each scientific article comprises references to other articles which are containing relevance and state the contribution of other researchers (Osareh, 1996) (Zhao, 2005). The consideration of reference used in the data is imperative for conducting the victorious study (Muppidi & Reddy, 2020) (Iqbal, et al., 2021) (Zeyun and Dawood, 2017). In (White & Griffith, 1980), a method is devised that considered the count of co-occurrences. In (Boyack & Klavans, 2010), a bibliographic coupling technique was utilized that provided the most precise solutions using co-citation analysis. In (Álvarez, *et al.,* 2016), a method is devised for generating a summary and retrieving information using the annotation corpus considering aspects, and polarity. In (Yu, *et al.,* 2014), a method is devised for citation analysis in order to identify the abnormal journals using coercive self-citation. The method employed11 features for devising the feature space in order to define the behaviours of a journal citation. In (Jurgens, *et al.,* 2016), behavioural analysis of scientific fields is devised with a corpus that annotated citation function and centrality. It revealed how the author is susceptible to communication configuration and venue of publication in the citation. In (Trujillo & Long, 2018), co-citation network was devised using bibliographic data and these data are utilized for recognizing the system and verifying it by testing internal consistency and stability.

The goal is to discover the influential researcher and influential citation from Web of Science (WOS) data, which includes research papers and journals. At first, the parsing of WOS data is carried out wherein the influential researcher and influential cited paper is extracted in a concurrent manner after processing. For determining the influential researcher, the parsing and extraction of the name of the author and Citation Report (CR) link is performed. After performing CR links, co-occurrence frequency of authors is computed, which showed co-citation analysis. Co-citation is carried out using a network graph. The co-citation network graph delivered visualization of communication amongst

different authors. The proposed technique offers author-author co-relation that predicts who is the next co-author. Simultaneously for determining influential cited papers, the research paper, and its cited papers are subjected to pre-processing phase. In pre-processing, keywords are produced that can be utilized in forecasting the next word. Then semantic similarity amongst research article and reference articles are evaluated. From Modified Word Movers Distance (MWMD), an influential cited paper is evaluated.

The key contribution of the paper is:

- **Proposed Hybrid model for finding influential researcher, and influential cited paper:** A novel hybrid model named Infgraph is devised, which helps to discover the influential cited paper and influential researchers using WOS data. Also, this method offered co-relation among authors that predict who is the next co-author.

Other sections are structured as: Section 2 presents classical influential research and citation analysis techniques. The proposed influential research and citation analysis is examined in Section 3. The efficiency of the developed model is presented in Section 4 and Section 5 presents the conclusion.

2. MOTIVATIONS

The major part of connecting research is citations. The comprehension of references used in the document is imperative for conducting successful research. The citation using a paper helps to reinforce parameters and connect them with sophisticated links which increases the quality of the citation. The illustration of existing techniques is done along with the limitations.

2.1 Literature Survey

The citations to the paper can help to authorize communities for computing the academic role and work quality. Understanding the usage of references in the document is crucial for conducting successful research. The eight classical models on the basis of citation analysis are depicted below. Angelou, *et al.* (2020) devised a technique for citation analysis using networks. The model was developed on the basis of preferential attachment and random networks. This method was analyzed by comparing the outcomes with another configuration model. The hybrid model provided an improved patent citation network and helped to illustrate the preferential attachment in later years. This technique facilitated the configuration of plausible theory considering real data using structural alterations of patent citation networks. However, this method was difficult when the cluster was big. In the proposed method, the clusters are created mutually so that all the clusters are equal in size. Lefebvre, *et al.* (2020) devised a technique for understanding and reviewing the research papers to advance the knowledge in this framework. The model helped to amalgamate and compute the mentoring for bridging classical knowledge. The citation network analysis was performed for locating influential articles across the disciplines. The method was analyzed with a citation network for identifying a powerful career. This method illustrates the transfer knowledge amongst disciplines and evaluated mentoring in this domain. The recognition of seminal career mentoring helped to offer sports mentoring scholars using a roadmap for promoting mentoring knowledge. However, the similarity measures failed, when there is no similar word between two sentences. In the proposed method, the word embedding space is used to solve this issue.

Tabak, *et al.* (2021) devised a method by analyzing the patterns and trends of papers considering references. It was observed that literature was growing distinctively in recent years. The method helped to discover single-authored and keyword ‘‘liquidity’’ for predicting the citation. In addition, the method helped to discover the most excellent papers, authors, and journals with respect to time. Anyhow, the categorization process of this method was difficult for large size data. In the proposed

method, this issue is solved by removing the redundant words and phrases present in the text data. Scherbakova and Bredikhin (2021) devised the citation network of scientific articles using a hypothesis, which helped to predict the likelihood of node using the node's degree. The method showed the rate at which the article attained citation based on citation count and determined the functional form. The method examined article age while receiving the citation. However, this method required more processing time. In the proposed method, the unwanted terms in the text document are removed so that the processing time is reduced.

Liu, *et al.* (2021) devised a technique that utilized adaptive topological co-evolution. The method devised a concise model using the co-evolution phenomenon between inter-reliant networks. The employment of structural evolution using counterparts is an essential topological evolutionary method for the complicated network. The method offered a perspective for improved understanding. Anyhow, the joint citation of different authors' identification was difficult. In the proposed method, the frequency of joint citation of different authors from the scientific graphs is evaluated to overcome this problem. Bai, *et al.* (2020) devised a model that performed citation analysis using the heterogeneous institution-citation network. The IPRank was utilized for computing the impact of paper and institution in a concurrent manner. The model utilized a heterogeneous structural technique for unveiling the impact of paper and institution that reflect the impact of citation, structural measure, and institution. At last, the impact of a similar institution was combined and the ranking of institutions and papers are computed. Anyhow, the performance of this method was low. In the proposed method, the performance is increased by determining the unethical citations and influenced citations from the scientific data.

Sugishita and Asakura (2020) devised a model for citation analysis using a citation network comprising vulnerability using two domains. The publication records were accumulated in from WOS and citation network using 1181 nodes and 4601 links but the processing time was high. The proposed method removes the unwanted terms in the text document so that the processing time is reduced. dellaBriottaParolo, *et al*. (2020) devised a model for citation analysis by examining the cumulative knowledge creation process considering 35 million publications. The method studied stylized models of persistent influence and diffusion which takes possible chains of citation. The method studied the publication which is based on Nobel prizes that had high ranks. The method suggested that analyzing cumulative knowledge can be beneficial in evaluating cumulative knowledge in evaluating scale and type of entire research. Anyhow, the determination of the semantic similarity in processing the natural language was difficult. The skip-gram model converted words to vectors in the proposed method and defended the relationship between words.

(Connor and Joffe, 2020) developed a model, named intercoder reliability (ICR) for the qualitative analysis. It enhanced the transparency, communicability, and systematicity of the coding procedure, which provided trustworthiness to the audience. However, it had less trustworthiness in reporting and generating the theme and preparation and collection of the data. The proposed model helped to offer author-author co-relation that predicts the next co-author and works together to predict the community of authors. (Xu, et al., 2020) presented a bibliometric analysis, which was done on the basis of 1,310 publications collected from the WOS. In this method, the keywords of supply chain disruption (SCD), organizations, and authors were analyzed in detail. Also, it was collect the recent hotspots of research and offered better research choices. However, some of the objective results were not explained clearly. In the proposed method, the optimization function in MWMD is used to attain better results.

3. PROPOSED INFGRAPH FOR FINDING INFLUENTIAL CITATIONS AND RESEARCHER

Citation data are utilized for evaluating several goals such as research credibility and institutions and so on. There have been standard citation measures utilized for finding the researcher's strength. The citation using a paper can reinforce its opinions and secure it with the educational association. The

citations to the paper can help to authorize communities for computing the academic role and work quality. The fundamental attribute that connects the study is a citation. Understanding the usage of references in the document is imperative for conducting triumphant research. Acknowledging the usage of references from the data is imperative for conducting a victorious study.

Figure 1 represents the block diagram of infgraph. At first, the parsing of WOS has performed wherein influential researchers and influential cited paper are extracted simultaneously. For determining the influential researcher, the parsing and extraction of author name and CR link are done wherein, the discovery of co-occurrence frequency is performed. Initially, the co-occurrence frequency of the first author is performed, which indicates the co-citation analysis and visualized in the co-citation network graph. The network graph and network density are used in the co-citation analysis. The co-citation network graph performs visualization of research communication amongst different authors. In this graph, every circle represents citation and the complete graph indicates citation amongst different articles. The network density is termed as an imperative metric for computing the potentiality of network connection. The cluster coefficient node is described as a possibility of two randomly chosen nodes of each other. The proposed technique offers the author-author co-relation, predicting, who is most likely to be the next co-author and author work together on which community of authors. For determining influential cited papers, the research paper, and its cited papers are

Figure 1. Schematic structure of Infgraph model for citation analysis

subjected to pre-processing phase. In pre-processing, keywords are produced that are utilized for predicting matching words. The word2vec model is utilized for training the model for predicting the suitable word that comes next. Then semantic similarity of a research article and its reference articles are evaluated using MWMD.

3.1 Influential Researcher: A Computation Model

The preliminary step in this model is the extraction of name of the first author and CR link. The imperative part of extraction is the co-citation of all author pairs. The different scientific data are arranged in co-occurrence matrix form. In this process, the WOS data has been acquired as input for experimentation for determining the co-citation produced from the co-citation graph. The citation network graphs distribute the research visualization communication amongst different authors. It computes the frequency of joint citation of different authors from the scientific graphs. The network comprises edges and nodes that represent the co-occurrence matrix. In this matrix, the names indicate nodes of the graph and edge indicates the weights of the graph.

3.1.1 Parsing and Extraction of First Author Name and CR Link

The extraction of citation references like names of author, title of article, name of journal, date of publication, and other entities is a beneficial and imperative process. The references are discovered for identifying the reference section.

3.1.2 Co-Citation

The next step is Co-citation, which is carried out using network graph and its property that represents network density. The density of the network is computed using a co-citation network graph using WOS data. In this process, the co-citation network graph helps to perform visualization of research amongst different authors. Each circle indicates citation and complete graph indicates citation amongst different articles. The WOS data is utilized for analysis which is indicated as the association of different scientific documents and its citation.

3.1.3 Visualization of Network

The visualization of the network is done using python wherein the WOS dataset is considered as an input for determining the co-citation assessment producing a co-citation graph.

3.1.4 Network Density

The density of the network is an imperative metric for computing the aptitude of network connection. The potentiality of a connection is a connection, which might exist amongst two nodes in a network.

3.1.4.1 Calculation of Network Density

It is computed based on the potential connection amongst nodes in the network. It is utilized for determining potential connection which is formulated as:

$$Potential\ \ correction = \frac{n \times (n-1)}{2} \tag{1}$$

where, n symbolizes the count of nodes in the network for determining network potentiality.

The network density is formulated as:

$$Network\ \ density = \frac{Actual\ \ connections}{Potential\ \ connections} \tag{2}$$

3.1.5 Cluster Coefficient

It is a metric of the degree in which the nodes in the graph are likely to cluster mutually. The cluster coefficient amongst node is described as a probability of two arbitrarily chosen friend's node and is formulated as:

$$Cluster\ \ coefficient(V) = \frac{2Nv}{Kv(Kv-1)} \tag{3}$$

where, V symbolizes node whose cluster coefficient is evaluated, Kv signifies the degree of V, and Nv indicates the count of links amongst neighbour of V.

3.1.6 Influential Researcher

The co-citation analysis is utilized for determining the occurrence analysis, which is utilized as an effectual technique for determining the analysis of citation pattern, but it failed to employ co-occurrence analysis amongst authors. It can be utilized for determining co-occurrence analysis amongst the authors. Thus, the model helped to offer author-author co-relation that predicts the next co-author and works together to predict the community of authors.

3.2. Influential Cited Paper: A Computation Model

A model is devised for determining the unethical citations, and influenced citations from the scientific data. The devised model helps to carry out various activities for determining relevancies amongst scientific data and is known as another document as the reference. The technique employs technical documents as input and prepares keywords for removing stop words from the main document and its reference document in parallel. A briefer illustration of each step is given below:

- **Research paper and its cited papers:** The goal is to determine the unethical citations and influenced citations from the research papers. The model helped to perform various tasks for determining the relevancy of the scientific document which is termed references. This model generates base papers and reference papers.
- **Pre-processing:** In this phase, the base and reference papers are taken as input wherein stop words and stemming process is employed for processing. The pre-processing of the scientific document is performed for eliminating superfluous words from the text dataset. The consequence of pre-processing is to provide smoother data processing. The text data are generally large in size that comprises redundant words and phrases, which affect the categorization process. Thus, it is imperative for removing the redundant and conflicting words by considering pre-processing phase.

3.2.1 Stop Word Removal

The stop words indicate words that are generally utilized in the sentence that involves prepositions, articles, and pronouns. The filtering of stop words is done prior to data processing. The stop word removal is a procedure for eliminating stop words using large data. The non-information behavior words are removed for minimizing the noise present in the data. The stop words removal is used to avoid large space accretion and enable rapid processing to attain effectual outcomes. The stop words like verbs, and nouns are eliminated with data. The stop words indicate words that involve "a" "as" "to", "the" etc. This word does not hold imperative meanings in sentences and is frequently detached using a research paper in pre-processing step. Another imperative step in pre-processing is stemming, which is a procedure to discover the root or base word with the stemming technique.

3.2.2 Stemming

The stemming is a procedure utilized for transforming the words to their stem form. In huge documents, several words are used which depicted a similar concept. The imperative method utilized for minimizing the word to its roots is known as stemming. The stemming procedure is utilized for transforming the words to the stem. In huge data, several works are used which helped to convey the same concept. The imperative step utilized for minimizing words to their root word is known as stemming. Various words can be processed to reduce words to the base form. The stemming process is compact, easy to utilize, and precise which removes the unwanted terms.

3.2.3 Word2vec Representation

The goal of the word2Vec is to produce a vector that exemplifies words. Each word vector is of a few dimensions and each distinctive word in the repository is allocated with a vector from the space. The training of the model is the imperative phase in word2vec. The layered model contained input layer, output layer and projected layers for predicting close words from the research papers. Each word vector is trained for making the possibility of the nearest words using corpus, which is given by:

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{j\in nb(n)}\log\ p\left(\frac{w_j}{w_n}\right) \tag{4}$$

where, $nb(n)$ symbolizes a group of neighbouring words of word w_n and $p(w_j / w_n)$. The associated vector is represented as V_{wj} and V_{wn}.

3.2.4 Skip-Gram Model

It is crucial for determining the semantic similarity in processing the natural language. It is utilized as a neural network using a solitary hidden layer for performing particular tasks. The technique converted words to vectors and defend relation between words. Figure 3 depicts the Skip gram model to discover contextual words.

3.2.5 MWMD for Similarity Matching

The technique concentrated on sentence semantic similarity for computing similarity on the basis of word frequency. The semantic similarity is computed using MWMD, which discovers semantic similarity amongst the scientific research papers. This computed the distance amongst research

Figure 2. Word2Vec embedding training model

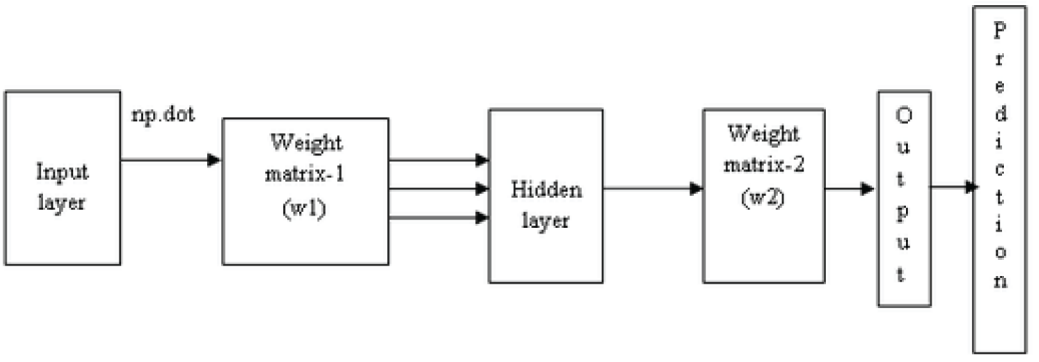
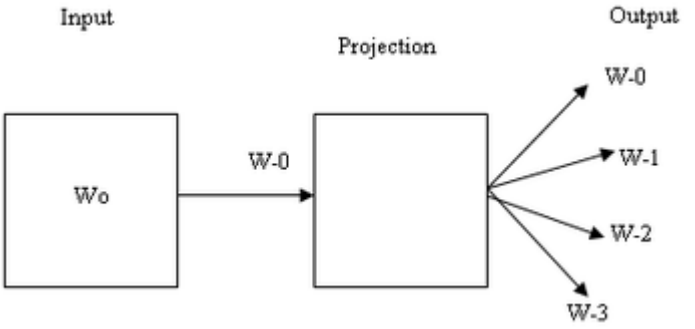


Figure 3. Skip gram model to discover contextual words



papers considering semantic distance amongst words wherein words are accumulated in the lexical dataset. After determining semantic distance amongst words, the method computes document likeness considering many-to-many similarities amongst words. The optimization function in MWMD worked effectively and is expressed as:

$$X \in R^{dxn} \tag{5}$$

where, d signifies word embedding dimension, and n symbolizes count of words:

$$d \in R^{n} \tag{6}$$

$$d_i = \frac{c_i}{\sum_i c_i} \tag{7}$$

where, i expresses word tokens. The distance amongst words represents Minkowski distance which is expressed as:

$$\left(\sum_{i=1}^{n} | X_i - Y_i |^p \right)^{1/p} \tag{8}$$

where, i and j symbolize word tokens. The research paper distance is given as:

$$\sum_{ij} T_{ij} c(i,j) \tag{9}$$

where, T symbolizes n time and n matrix. Each element $T_{ij} \geq 0$ and is modelled as:

$$\min \sum_{i,j=1}^{m} D(i,j) \tag{10}$$

It assists to discover how much of the word i in base paper travels to word j in the reference document. D symbolizes m dimensional distance matrix. It indicates reference documents and the source document. MWMD is considered the least cumulative cost that affects base paper to reference paper and offers an improved solution. The MWMD is expressed as:

$$\sum_{i,j=1}^{m} TD_{ij} d(i,j) \geq \sum_{i,j=1}^{m} TD_{ij} \left[\sum_{i=1}^{n} | x_i - y_i |^p \right]^{1/p} \tag{11}$$

3.2.6 Influential Cited Paper

Conventional similarity calculations failed wherein two sentences pose no ordinary words but unusual words had a similar meaning. This model solved the issue by utilizing word embedding space. The technique utilized word embedding using two different articles for evaluating minimal distance amongst them. The model utilized closet space without employing numerous words to obtain an influential cited paper.

4. RESULT AND DISCUSSION

The assessment with classical strategies using a publicly avaiable dataset is depicted. The analysis is done by considering various research papers.

4.1 Experimental Setup

The implementation of the proposed strategy is carried out in python considering PC has Windows 10 OS, 4 GB RAM, and Intel core processor.

4.2 Dataset Description

The dataset used for the performance evaluation is Co-citation Analysis (https://data.mendeley.com/datasets/4n8ns8vzvz/3). The contributors of this dataset are Dmitriy Korobskiy, George Chacko, Sitaram Devarakonda, Avon Davey, James Bradley, and Siyu Liu. This dataset contains the details about the publication year, unique publication count, unique reference count, and total reference count.

4.3 Evaluation Measures

The efficiency of the developed model is computed on the basis of certain parameters that involves precision, recall, and F-measure.

4.3.1 F1-Score

It is defined as a measure of the accuracy of test and is weighted through the harmonic mean value of recall and precision of test. The F-measure is defined as:

$$F1 - Score = \frac{2 * precisiom * recall}{precision + recall} \tag{12}$$

4.3.2 Precision

It is termed as the proximity of more than two dimensions to each other, and is complex from that of accuracy. It is expressed as:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

where, FP is false positive and TP is true positive.

4.3.3 Recall

It estimates the whole amount of the actual positives, in which the system captures with the label of it as true positive and it is represented as:

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

where, FN is false negative.

4.4 Influential Researcher: Result

In this analysis, co-citation graph analysis is done on the basis of network properties using network density. The outcomes of network properties based on WOS are illustrated in table 1. The properties of networks are illustrated. The outcomes helped to precede further study on the citation of scientific documents.

4.5 Influential Cited Paper: Result

The semantic similarity measure amongst sentences is computed with some sample similarities depicted in table 2. The analysis is done based on the publications from the Institute of Electrical and Electronics Engineers (IEEE) journals and conferences, Springer, and the Association for Computing Machinery (ACM) publisher. The results are categorized as irrelevance, non-domain relevance, and domain relevance.

In this analysis, the publications from IEEE journals and conferences are considered for evaluation. The results are categorized as irrelevance, non domain relevance and domain relevance. The outcomes

Table 1. Assessment of developed model using properties of network

Properties of Network	Result
Average degree	2.553
Average weighted degree	13.864
Density of Graph	0.019
HITS modularity	0.543
Length of Average path	2.304
Page rank associated components	3
Average clustering coefficient	0.171
Diameter of Network	6

Figure 4. Priorities of the network using network density

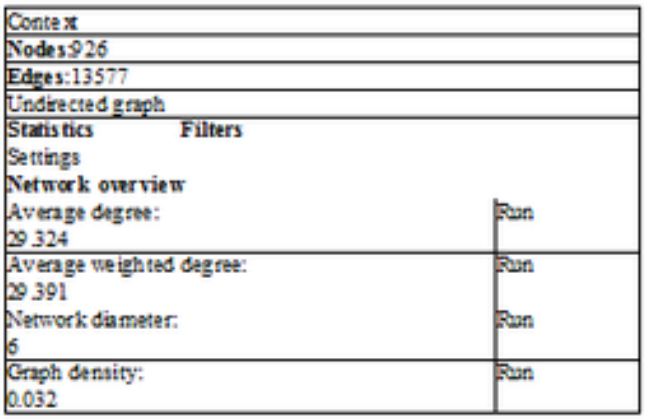


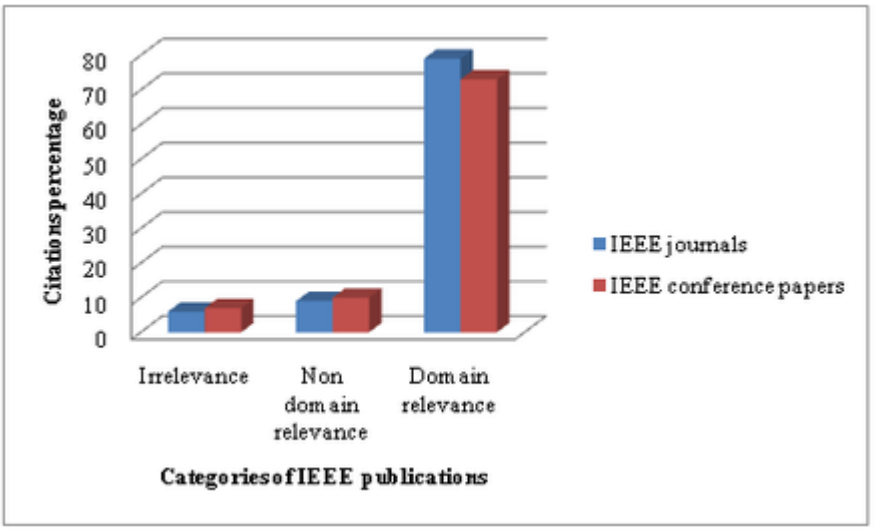
Figure 5. Co-citation edge and weight depiction of WOS data

Data table		Graph					
Nodes Edges	Configuration	Add node	Add edge	Search/ Replace	Import spread sheet	Export table	More actions
Nodes	Id	Label		Degree	Weight	Eccess	Closeness
WANG X	WANG X	WANG X		61	61	4	2.353
RATHNASAMY S	RATHNASAMY S	RATHNASAMY S		60	60	4	2.514
GLOVER F	GLOVER F	GLOVER F		60	60	4	2.433
PICRO M	PICRO M	PICRO M		59	60	4	2.93
HUELSERMANN R	HUELSERMANN R	HUELSERMANN R		59	60	4	2.93
AHUJA R K	AHUJA R K	AHUJA R K		56	57	4	2.298
RENAIS O	RENAIS O	RENAIS O		56	57	4	2.298
GUERIN R	GUERIN R	GUERIN R		55	55	5	3.026
THE INTERNET	THE INTERNET	THE INTERNET		54	54	5	3.539
LEE J	LEE J	LEE J		53	53	4	2.386
WANG Y	WANG Y	WANG Y		53	53	4	2.392
YU W	YU W	YU W		51	51	5	3.017
KARAGIANNIST	KARAGIANNIST	KARAGIANNIST		49	50	5	3.058
MUKHERJEE B	MUKHERJEE B	MUKHERJEE B		49	49	4	2.97
ZANDER S	ZANDER S	ZANDER S		49	50	5	3.058
KAWAIDA H	KAWAIDA H	KAWAIDA H		47	47	1	1
WEI S	WEI S	WEI S		47	47	1	1
DINGLEDEIN R	DINGLEDEIN R	DINGLEDEIN R		47	47	4	2.323
GROOM JR	GROOM JR	GROOM JR		47	47	1	1
YANG JM	YANG JM	YANG JM		47	47	1	1
LI XY	LI XY	LI XY		47	47	1	1

Table 2. Example of sentence wise semantic similarity amongst base paper and one of the reference papers

	A1	A2	A3	A4	A5	A6	A7	A8
A1	1	0.928	0.563	0.365	0.657	0.644	0.682	0.528
A2	0.928	1	0.542	0.351	0.622	0.615	0.646	0.507
A3	0.563	0.542	1	0.878	0.909	0.959	0.926	0.885
A4	0.365	0.351	0.878	1	0.880	0.864	0.853	0.921
A5	0.657	0.622	0.909	0.880	1	0.945	0.953	0.944
A6	0.644	0.615	0.959	0.864	0.945	1	0.959	0.909
A7	0.682	0.646	0.926	0.853	0.953	0.959	1	0.915
A8	0.528	0.507	0.885	0.921	0.944	0.909	0.915	1

Figure 6. Citation proportion of IEEE journal, and conference



revealed that the majority of research papers of each publisher have influence citation using marginal difference. The domain significance of articles is high and at the same time, there is less percent of influenced citation exits. Several papers are considered wherein it is observed that the majority of citations are domain relevance, some percent are non domain relevance and some are irrelevance.

From figure 7, the Springer papers from original research, review articles, quick communication, and case studies are considered. According to analysis, all kinds of articles, such as original research, rapid, review articles, case studies, and reference papers are discovered and less percentage of irrelevance citations and the maximum citations are domain relevance.

The ACM published articles into four classes, namely proceedings of conference, Magazines, scholarly communications, and ACM Digital library, which is depicted in figure 8. Also, the results of the citations are categorized as irrelevance, non domain relevance and domain relevance. According to outcomes, all four classes have influenced citations, but less percentage of irrelevance citations and the maximum citations are domain relevance.

Figure 9 illustrated F1-Score values which stated that MWMD performed better in contrast to WMD. The comparison between two variants of Word Mover's Distance (WMD) namely Relaxed

Figure 7. Percentage of Citation using different Springer data

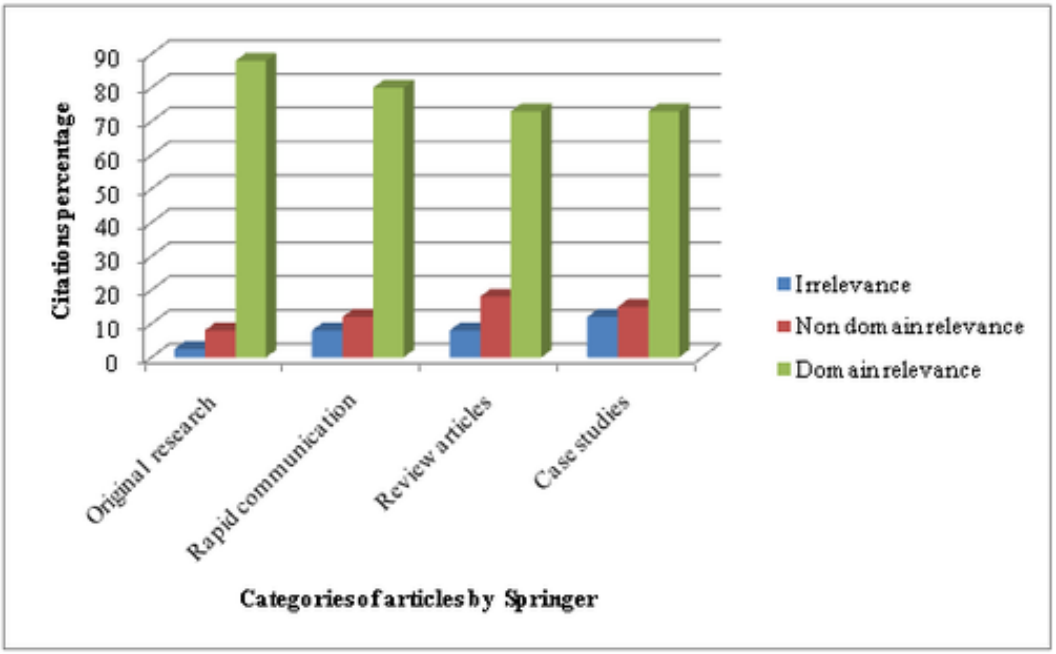


Figure 8. Percentage of Citation considering ACM publisher

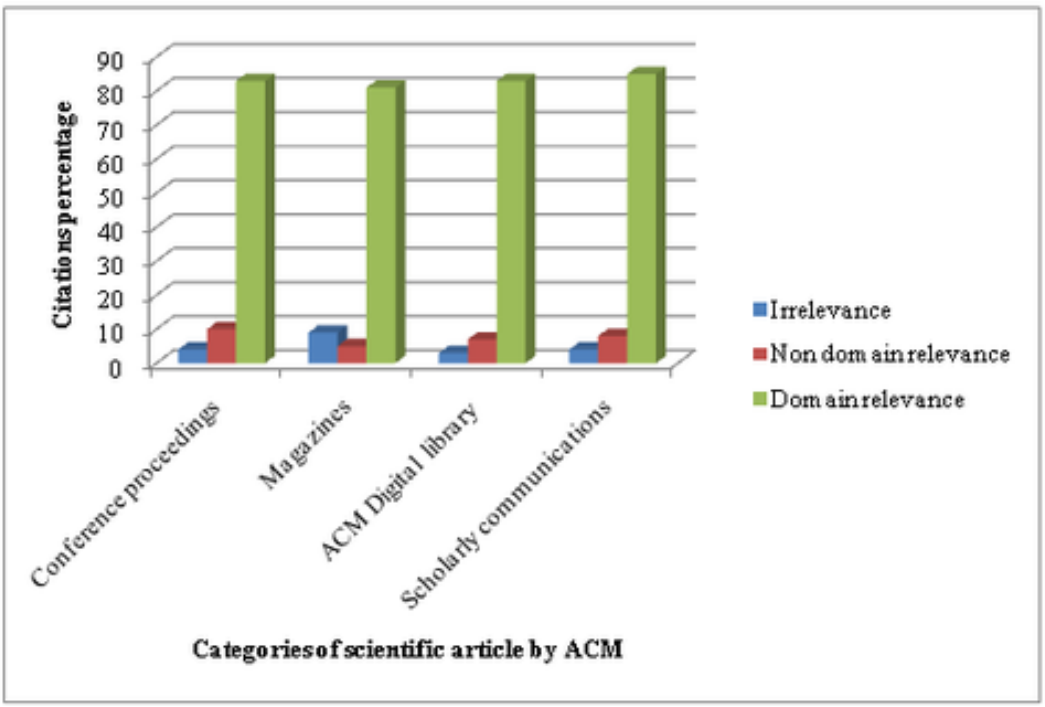
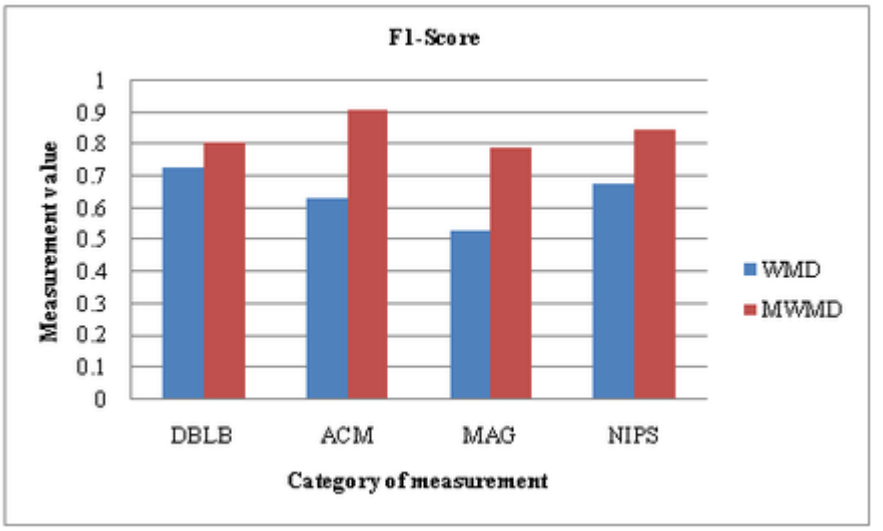


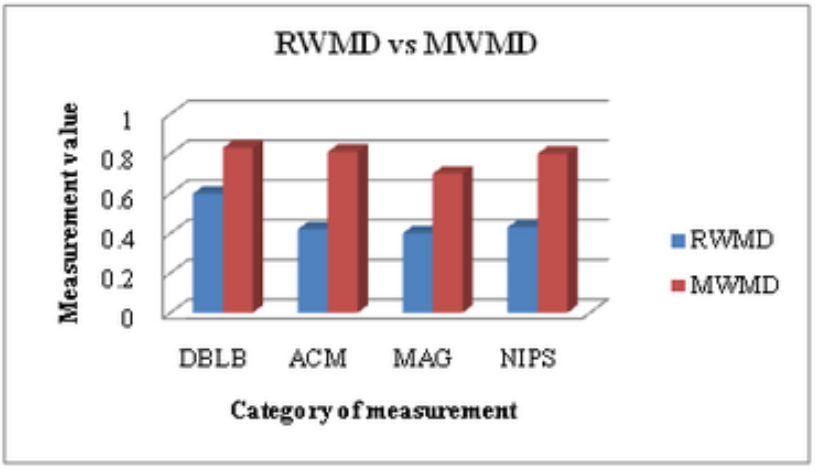
Figure 9. Assessment with F1-score



Word Mover's Distance (RWMD) and MWMD is depicted. The citation dataset is obtained from Neural Information Processing System (NIPS), DataBase systems and Logic Programming (DBLP), ACM, and Microsoft Academic Graph (MAG) that contained conference papers data. These datasets and one synthetic data set are produced and utilized in the analysis. The MWMD performed better compared to WMD in terms of the F1-score metric. The ACM has the highest F1-score than the other citation datasets, such as DBLB, MAG, and NIPS.

Figure 10 depicts the comparison amongst RWMD and MWMD. The citation dataset obtained from NIPS, DBLP, ACM, and MAG is considered for this evaluation. The DBLB has a maximum F1-score than the NIPS, ACM, and MAG. The outcomes revealed that WMD performed better in contrast to RWMD considering the F1-Score measure.

Figure 10. F1-score of relaxed WMD with respect to MWMD



4.6 Comparative Analysis

This section represents the comparative analysis of the proposed infgraph with the existing methods, such as Hybrid network (Angelou, K *et al.* 2020), IPRank (Bai, X *et al.* 2020), Adaptive topological coevolution framework (Liu, X.F *et al.* 2021). The analysis is done by varying the published paper and the performance is evaluated based on the metrics, such as precision, recall, and F1-score. From this analysis, it is clear that the proposed infgraph has the maximum precision, recall, and F1-score than the existing Hybrid network, IPRank, and Adaptive topological coevolution framework. The proposed infgraph has a maximum precision of 0.9029, recall of 0.9259, and F1-score of 0.9142.

In this research, a novel hybrid model named Infgraph is devised, which helps to discover the influential cited paper and influential researcher using WOS data. Also, this method offered co-relation among authors that predict who is the next co-author. The performance of the proposed system is improved by using various techniques, such as determining the influential researcher, Co-citation, network density, cluster coefficient, stop words and stemming process, close words identification, skip-gram model, and MWMD similarity matching.

5. CONCLUSION

In this paper, a novel hybrid model is devised, which helps to discover the influential cited paper and influential researcher using WOS data. For discovering influential researcher, certain step is carried out in which first step is to extract the first author name and citation reference link. Thereafter, the co-citation is carried out to generate author-author co-relation which helps to predict next co-author. Then, the visualization of the network is carried out for research communication among different authors. Then, the network density is evaluated for computing the network connection potentiality. At last, the cluster coefficient is employed for discovering the influential researcher. At the same time, for discovering influential cited papers, the research papers and its cited papers are adapted. In this process,

Table 3. Comparative analysis based on published papers

Methods/ Published papers	Hybrid network	IPRank	Adaptive topological coevolution framework	Proposed infgraph
	Precision			
10	0.7004	0.7212	0.7416	0.7739
100	0.7381	0.7665	0.7815	0.8083
1000	0.7579	0.7837	0.8042	0.8441
10000	0.8178	0.8535	0.8783	0.9029
	Recall			
10	0.7340	0.7611	0.7708	0.7941
100	0.7646	0.7990	0.8023	0.8444
1000	0.7783	0.8171	0.8361	0.8824
10000	0.8540	0.8871	0.9007	0.9259
	F1-Score			
10	0.7196	0.7406	0.7559	0.7839
100	0.7511	0.7824	0.7917	0.8260
1000	0.7680	0.8000	0.8198	0.8629
10000	0.8355	0.8699	0.8894	0.9142

the pre-processing is employed wherein the stop word removal and stemming process is adapted for eliminating the redundant words. Then, the word2vec model is further used for the training model to estimate the suitable word that comes next. At last, the MWMD is used for determining similarity. The proposed method assists the researchers to ensure the new research chances and construct new viewpoints. In the future, publications from various data sources, such as Google Scholar, Theses, and ProQuest Dissertations will be considered for the analysis.

REFERENCES

Álvarez, M.H., Gómez, J.M., & Martínez-Barco, P. (2016). Annotated corpus for citation context analysis, *Latin-American Journal of Computing, 3*, 35–41.

Angelou, K., Maragakis, M., Kosmidis, K., & Argyrakis, P. (2020). A hybrid model for the patent citation network structure. *Physica A*, *541*, 123363.

Bai, X., Zhang, F., Ni, J., Shi, L., & Lee, I. (2020). Measure the Impact of Institution and Paper Via Institution-Citation Network. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 17548–17555.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately. *Journal of the American Society for Information Science and Technology*, *5*, 1–26.

Connor, C. O., & Joffe, H. (2020). Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. *International Journal of Qualitative Methods*.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, *65*(9), 1820–1833.

Iqbal, Q., Ahmad, N. H., & Li, Z. (2021). *Frugal-based innovation model for sustainable development: technological and market turbulence. Leadership & Organization Development Journal*.

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2016). *Citation Classification for Behavioral Analysis of a Scientific Field*. arXiv:1609.00435.

Kajikawa, Y., & Takeda, Y. (2009). Citation network analysis of organic LEDs. *Technological Forecasting and Social Change*, *76*(8), 1115–1123.

Lefebvre, J. S., Bloom, G. A., & Loughead, T. M. (2020). A citation network analysis of career mentoring across disciplines: A roadmap for mentoring research in sport. *Psychology of Sport and Exercise*, 101676.

Liu, H., Kou, H., Yan, C., & Qi, L. (2019). Link prediction in paper citation network to construct paper correlation graph. *EURASIP Journal on Wireless Communications and Networking*, *1*, 1–12.

Liu, X. F., Chen, H. J., & Sun, W. J. (2021). Adaptive topological coevolution of interdependent networks: Scientific collaboration-citation networks as an example. *Physica A*, *564*, 125518.

Muppidi, S., GMRIT, G., & Reddy, K.T. (2018). Challenges in Citation Analysis. *International Journal of Pure and Applied Mathematics*, *118*(7), 27–31.

Muppidi, S., & Thammi Reddy, K. (2020). *Influenced Citation Analysis using Modified Word Movers Distance (MWMD).* IEEE.

Muppidi, S., & Reddy, K. T. (2020). Co-occurrence analysis of scientific documents in citation networks. *International Journal of Knowledge-based and Intelligent Engineering Systems*, *24*(1), 19–25.

Neuhaus, C., & Daniel, H. D. (2008). Data sources for performing citation analysis: An overview. *The Journal of Documentation*.

Osareh, F. (1996). Bibliometrics, citation analysis and co-citation analysis: A review of literature I. *Libri*, *46*(3), 149–158.

Parolo, P. D. B., Kujala, R., Kaski, K., & Kivelä, M. (2020). Tracking the cumulative knowledge spreading in a comprehensive citation network. *Physical Review Research*, *2*(1), 013181.

Pieters, R., Baumgartner, H., Vermunt, J., & Bijmolt, T. (1999). Importance and similarity in the evolving citation network of the International Journal of Research in Marketing. *International Journal of Research in Marketing*, *16*(2), 113–127.

Rahm, E., & Thor, A. (2005). Citation analysis of database publications. *SIGMOD Record*, *34*(4), 48–53.

Scherbakova, N. G., & Bredikhin, S. V. (2021). Preferential attachment in the citation network of scientific articles. *Journal of Physics: Conference Series*, *1715*(1), 012055.

Sugishita, K., & Asakura, Y. (2020). Citation network analysis of vulnerability studies in the fields of transportation and complex networks. *Transportation Research Procedia*, *47*, 369–376.

Tabak, B. M., Silva, T. C., Fiche, M. E., & Braz, T. (2021). Citation likelihood analysis of the interbank financial networks literature: A machine learning and bibliometric approach. *Physica A*, *562*, 125363.

Thakur, N. (2017). *Increased soil-microbial-eco-physiological interactions and microbial food safety in tomato under organic strategies. In Probiotics and Plant Health*. Springer.

Thakur, N. (2018). *In silico modulation techniques for upgrading sustainability and competitiveness in agri-food sector. In Silico Approach for Sustainable Agriculture*. Springer.

Trujillo, C. M., & Long, T. M. (2018). Document co-citation analysis to enhance transdisciplinary research. *Science Advances*, *4*, 1–9.

White, H. D., & Griffith, B. C. (1980). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, *32*, 163–170.

Xu, S., Zhang, X., Feng, L., & Yang, W. (2020). Disruption risks in supply chain management: A literature review based on bibliometric analysis. *International Journal of Production Research*, *58*(11), 3508–3526.

Yu, T., Yu, G., & Wang, M. Y. (2014). Classification method for detecting coercive self-citation in journals. *Journal of Informetrics*, *8*, 123–135.

Zeyun, L., & Dawood, S. R. S. (2017). Examining Inter-city connections in Southeast Asia based upon interlocking city network model. *International Journal of Advanced and Applied Sciences*, *4*(1), 110–115.

Zhao, D. (2005). Challenges of scholarly publications on the Web to the evaluation of science—A comparison of author visibility on the Web and in print journals. *Information Processing & Management*, *41*(6), 1403–1418.