IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Received 1 December 2017; revised 3 September 2018; accepted 4 September 2018.
Date of publication 17 September 2018; date of current version 3 March 2021.

Digital Object Identifier 10.1109/TETC.2018.2870179

On the Expressive Power of Scientific Manuscripts

G. S. MAHALAKSHMI, R. SIVA, AND S. SENDHILKUMAR

G. S. Mahalakhsmi is with the Department of Computer Science & Engineering, College of Engineering Guindy (campus), Anna University, Chennai, TN 600025, India
R. Siva. is with the Department of Computer Science and Engineering, KCG College of Technology, Chennai, TN 600097, India
S. Sendhilkumar is with the Department of Information Science & Technology, College of Engineering Guindy (campus), Anna University, Chennai, TN 600025, India
CORRESPONDING AUTHOR: G. S. MAHALAKSHMI (gsmaha@annauniv.edu)

ABSTRACT Every research manuscript is appreciated in the form of citations. Citations are expected to carry the essence of the underlying base paper by some rhetorical means. However, this is not true in reality. Citation manipulations are equally possible which shall be identified using research semantics. This paper discusses machine learning based approaches for analyzing research citations with the aim of finding quality research citations. On analyzing the semantics of the research manuscript and the respective citations, this paper proposes various metrics for citation quality analysis including deep cite, raw expressive power, expressive power and normalized expressive power.

INDEX TERMS Citation analysis, semantic analysis, citation quality, machine learning, text mining, availability index, article metrics, deep learning, expressive power

I. INTRODUCTION

Measuring scientific influence using statistical and empirical approaches have been well studied in the recent past [1]. However, despite the rising of text mining and machine learning research, there has not been noted progress in application of machine in bibliometrics until recent past. Over last two years, there are sincere approaches on applying text mining techniques for providing new research findings in bibliometrics. Specialised machine learning algorithms are getting developed towards research quality analysis. The major delay in applying text mining to bibliometric is the closed or restricted access of research manuscripts. However, this issue is also wiped off completely with the huge blow of open access journals which eventually compelled the restricted access to semi-restricted access or that carrying open access fee. Since publishing research articles carries article processing charges right from less than 10 US $, the dream of article publishing is no more a dream coming true. However, to discuss about the quality of such research articles is always a question.

In the other side, this research boom resulted in storming citation counts which resulted in higher and higher journal impact factors. Various researchers have argued about the correctness of Journal Impact Factors [2]–[4]. The fundamental measure governing Journal Impact Factor is citation count. Therefore, there is a strong necessity to analyse the motive behindfetching research citations. There are manipulations quite possible in this aspect as well that an ordinary article when citing a most popular research manuscript gets higher visibility. Again this depends on the availability of the research article as well. This paper proposes machine learning approaches to analyse the rhetorical sentiments against research citations with respect to the base research article. Through the analysis this paper also proposes various article metrics concerned with citation quality.

II. RELATED WORK

Citations are an integral part of quality research. Citation counts are evergreen factors of research prestige. Citation indicators have received wide acclaim in the bibliometric literature. However it is very much essential that these indicators have to be accurate, robust and not biased [5]. H-index [6], IF, 5-year IF, SNIP, SJR, Eigen factor, Article Influence Score [7] are to name a few. Delayed citations [8], [9] are also a sign of interest in citation analysis. It is a measure of citation durability of articles which reflect the information content said in the cited article for their entire citation life-cycle [10], [11].

With remarkable progress on text mining and machine learning research disciplines, there is not much research dedicated into analyzing scholarly literature. Shifting the bibliometric research focus into article structure analysis [12] and article content analysis [13] has been started only in the recent past. Right from keyword analysis [14] to Content and proximity based approaches for analyzing research co-citations have been examined [15], [16]. Citation context extraction and expansion of citation contexts using various external word sources were attempted to bring more meaningful interpretations into context analysis. Citation context rhetorics have been explored widely in the literature. All these approaches have initated semantic analysis using popular similarity metrics like cosine similarity [17], [18].

Starting from cue phrase based rhetoric analysis to critic analysis, there lies enough scope for text mining incorporated with supervised and unsupervised learning methods to reveal the most out of citation contexts [19]. The order of cites appearing in the citing article [20] is also checked for plagiarism with that of the cited article. There is handful of budding literature on application of machine learning to bibliometric analysis. Sincere efforts to employ deep learning techniques to quantify author contributions [21] identify highly cited articles [22] and article topics [23] etc. are being considered in the recent past.

Application of topic models has been a welcome approach in citation quality analysis. Topic Models are useful in defining a probabilistic representation of the latent factors of corpora called topics. They are typically used for extracting the representative contents from text corpora [24]–[27]. Context sensitive topic models [28] are employed to measure author influence. From completely unsupervised approaches [29] to topic modeling,to Probabilistic generative models [30], neural networks and their variants are utilised to obtain the deep unfolding for generating document topics [30].

This paper proposes techniques for measuring the citation context of a research manuscript from respective citations and projects the rhetorical quality as a measure of article's expressive power by employing deep learning techniques. The articles impact is also analyzed and the proposed metrics are normalized to establish the actual quality of the research manuscript from the perspective of its citations.

III. BACKGROUND: AUTOMATED CLUSTERING OF RESEARCH CITATIONS

The problem taken is of high semantic orientation related to citation context and therefore it is essential that a manuscript which has high citation count has to be assumed for experimental analysis.

A. DATASET

The research article assumed for experimental purposes is the famous article that has proposed the 'h-index' [6]. To date, the article has 7153 research citations. For experiments, citations upto 2016 are considered. Further details about the dataset collection are provided in Table 1.

In the total of 1854 'not downloaded' category, 46 were books, 1670 were non-English citations, and 138 were not downloadable at the time of data collection. We employed specialized crawlers for downloading the title of all citation articles and then further downloaded the full citation articles. There wasn't a great success in automated citation downloading from google scholar and therefore, we collected the missed ones manually.

TABLE 1. Citation summary of Hirsch [2006] until 2016.

Year	Downloaded	Not downloaded	Total citations
2005	11	5	16
2006	55	23	78
2007	108	54	162
2008	198	64	262
2009	253	86	339
2010	388	175	563
2011	407	226	633
2012	418	191	609
2013	521	237	758
2014	547	283	830
2015	512	249	761
2016	532	252	784
Total	**3951**	**1854**	**5796**

B. AUTOMATED CLUSTERING

The research citations of Hirsch [6] were examined for interdisciplinary research discussion. The objective is to measure the impact of the seed article and therefore, the interdisciplinary research citations were identified using automated clustering. The automated clustering is approached via split and merge technique (refer Algorithm 1). For initializing cluster size, we assumed 0.1 percent of the available corpus.

Algorithm 1: Automated Clustering of Research Articles

Input: Research articles in text format without references
Output: Research articles categorized as multiple cluster and Outlier detection
Methodology: Automated clustering using split and merge technique

```
1. Assume K (0.1 percent of Corpus size) as number of clusters.
2. Assign each cluster a document from the corpus.
3. Assign the remaining documents to the closest cluster.
4. While the cluster does not converge either with the previous 2 iterations
   a. Assign the documents from the cluster to the closest cluster.
   b. Recalculate the centroid of each clusters.
   c. Split the clusters if the intra cluster similarity is less than threshold (Split logic is mentioned below as separate code flow)
   d. Merge two clusters if the inter cluster similarity is greater than or equal to the intra-cluster similarity
5. Merge the clusters if the inter-cluster similarity is greater than or equal to the intra-cluster similarity.
6. Merge all the clusters having cluster size as 1 and name it as Outlier
Split (Cluster):
1. Assume n = (cluster size/20 percent of cluster size)
2. Initialize nFold&nFoldStart to 0.
3. While nFold< n,
   a. Split the cluster into n new clusters with each new cluster size to be 20 percent of the old cluster size.
```

TABLE 2. Results of automated clustering of scientific articles for inter-domain filtration.

S. No.	Article title	Validation
1.	The Clinical Relevance of Information Index (CRII): assessing the relevance of health information to the clinical practice	correctly detected
2.	A survey of DEA applications	correctly detected
3.	A concept for inferring 'frontier research' in grant proposals	correctly detected
4.	On the Use and Abuse of Economic Journal Rankings	incorrectly detected
5.	The objectivity of national research foundation peer review in South Africa assessed against bibliometric indexes	incorrectly detected

```
    b. Merge two clusters if the inter cluster similarity is
       greater than or equal to the intra-cluster similarity.
    c. Increment nFoldStart by newClustersize/n and
       nFold by 1.
    d. Check the value of the k (new clusters formed) clusters
       with the value of the k clusters formed in previous fold.
        i. Find the value of the clusters in current and
           clusters retained in the previous generations.
        ii. Find the maximum number of intersection of each
            k clusters in the one generation with each of the k
            clusters in another generation.
        iii. Find the value of each k clusters by dividing the
             maximum with size of the k cluster.
        iv. Find the value of the clusters formed in one
            generation by taking the average value of the k
            clusters in that generation.
        v. Retain the clusters in the generation whichever is
           having the highest value.
4.  Return k new clusters.
```

There were 5 clusters with single citation articles, which were assumed as outliers. Among the 5 articles, 3 were correctly detected as inter-disciplinary where the other 2 were not very accurate. This may be due to various reasons like the semantics of idea discussed within the research article. We did not care much at this point on bringing back the 2 as the validations were only examined by fellow researchers (Table 2). We evaluated the automated clustering with that of other two standard clustering and the results are tabulated in Table 3.

TABLE 3. Evaluation of automated clustering with other state-of-art techniques.

Clustering technique	Purity	Information gain	Normalised mutual information
Automated Clustering - Proposed	0.75	0.07	0.06
gsdmm [31]	0.63	0.004	0.007

As seen from Table 3, the proposed automated clustering results in improved purity and is recommended for inter-disciplinary article identification (refer Figure 1). One might argue that the articles of varied domains might have clustered together and could have escaped the filtration. However, semantic analysis at the later part of the proposed citation quality analysis would drop such articles as they would be semantically far from the seed article.

IV. RHETORICAL CITATION QUALITY ANALYSIS

The underlying idea behind analyzing the rhetorical nature of citation contexts are presented in Figure 2. The proposed work utilizes two approaches of topic models for semantic analysis, namely, extended topic modeling [21] and deep topic modeling [21]. Both the topic models employ Hierarchical Dirichlet approach (HDP) [32] for generation of topics.

Initially, the citations checked for inter-disciplinary research discussions are subjected to citation context extraction. This process attempts to extract the cite contexts of every citations to the seed paper, roughly around 100 words. However, this is not fixed and is assumed with flexibility including the start and end of the cite context (refer Algorithm 2).

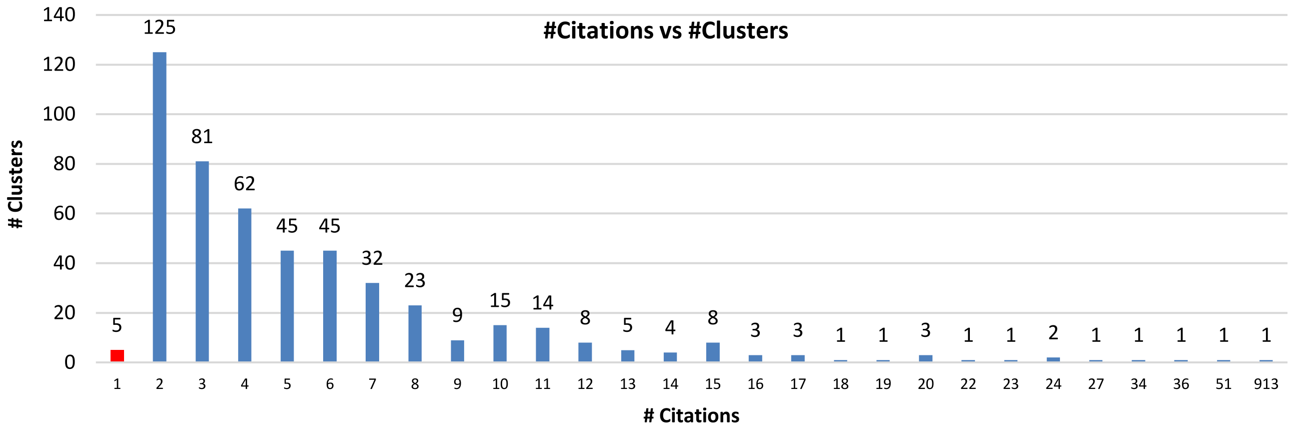

FIGURE. 1. Results of automated clustering.

Algorithm 2: Citation Context Extraction

```
Input: Research articles categorized as multiple cluster
Output: Extracted Context of all research articles.
1. Identify the reference number or detail about the citation
   context.
2. Identify the occurrence of the citation or co-citation by
   using the author name, reference number or by using
   some specific keywords (introduced by, discusses,
   proposed by) followed by or before the author name.
3. Retrieve 100 words before and after the citation
   occurrence.
4. Store the details of the co-citation research papers for
   future co-citation analysis.
```

Algorithm 3: Extended Topic Context Modeling

```
Step 1: HDP Input: Extracted context of research articles
Output: Topics Generated for all research context
1. 1st Mapper: Count the number of words in each file and
   emit file name as key and count as value
2. 1st Reducer: Emit the key, value pair as it is and output
   to the intermediate file.
3. 2nd Mapper:
   a. Get the total word count from the intermediate file.
   b. The no. of topics is found out by the total words
      by the no of words in each topic.
   c. Override IsSplittable function of FileInputFormat
      class to get the entire single file in one map tasks.
   d. Assume $\boldsymbol{\alpha} = 0.1$, $\boldsymbol{\beta} = 0.01$, a scalable parameter and
      it is updated based on the number of lines and total
      number of words in the file.
   e. The index of each word is determined. The words
      appearing multiple times are mapped to the lowest index.
   f. Form the base distribution H and get the initial topic
      assignments for all the topics using the random generation method.
   g. Form the next distribution using the previous distribution
      and the scalable parameter and get the
      updated topic assignments
   h. Continue step g for maximum number of iterations.
   i. Write the topic assignments into a file.
   j. Output to reducer the name of the document and its
      corresponding topic file.
4. 2nd Reducer – Emit the key, value pair as it is and output
   each input's topic assignments to separate file.
Step 2: WordNet Input: Topics Generated for all research
context
Output: Extended topics for all research context
1. For all the words in the input file, proceed the below steps.
2. Get the probability of the word from the input file.
3. Find the set of synonym words using WordNet in nltk
   corpus.
4. Assign probability for each word with the probability
   found in Step 2.
```

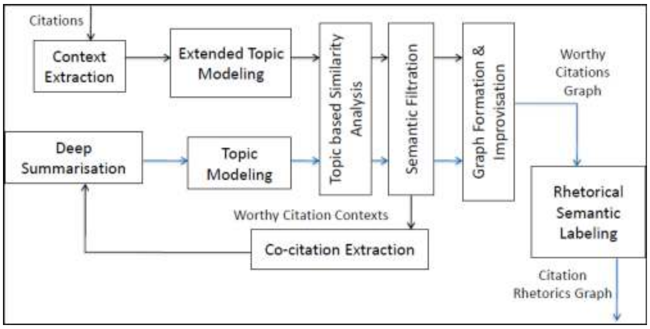

FIGURE 2. **Overall idea of rhetorical citation quality analysis.**

Extended topic models are topic models with WordNet [33] suggestions for word extensions (Algorithm 3). Here, the extracted citation contexts are subjected to extended topic modeling. Initially the contexts are processed for HDP topics and the results are fed to WordNet for adding better word suggestions. The low perplexity of 0.23 proves that WordNet additional suggestions provide more strength to interpreting the citation context (refer Figure 3). There might be more than one context in which the base article is cited within the text. All contexts are treated separately since rhetorical difference in textual narration is equally possible.

The topics obtained are checked for topic similarity and are filtered based on a dynamic adaptive threshold. This dynamic adaptive threshold involves applying adaptive differential evolution (ADE) [34]. The contexts need to be thoroughly checked for intended and implied meaning. Therefore, the contexts are subjected to various classes of textual and statistical similarity metrics [35]. The results of similarity metrics are subjected to adaptive differential evolution (refer Algorithm 4).

Algorithm 4: Adaptive Differential Evolution for Optimized Similarity Threshold

```
Step 1: Similarity Score Optimization
Input: Various Similarity Measures of all research context
Output: Optimised score for all research context
1. Represent each individual as a vector of metrics.
2. Form the initial population.
3. Initialize the mutation strategies, crossover vector and
   scaling factor.
4. Chose a mutation strategy based on the scaling factor
   and a random number.
5. Generate a trial vector from the initial population using
   crossover vector.
6. Evaluate the trial vector and if the trial is better than the
   original vector, replace with the original vector in the
   population.
7. Continue this process for maximum number of
   generations.
8. Return the maximum value from all the generations as
   optimized score
```

We have also attempted at performing the optimized score computation based on classic differential evolution. The

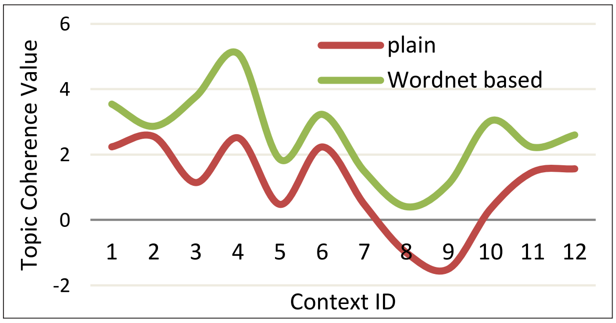

FIGURE 3. WordNet impact over citation context topic coherence.

convergence of adaptive DE is better than classic DE (refer Figure 4). Equally, adaptive DE produced higher optimum threshold when compared to classic DE, on execution of all citation contexts. The citations thus obtained are quality citations and are processed to form the first level research citation graph. With these set of citations, respective co-citations are obtained and is subjected to deep topic modeling.

This paper utilises Deep topic models which are topic models constructed over deep stacked auto encoders [21]. Deep stacked auto encoders are chosen for their ease and fastness to converge on handling large text corpora. The co-citations extracted for seed article might have overlap with the citation set of seed article and would already be part of the constructed citation graph. Therefore, utmost care is observed to remove those in common and the remaining unique co citations are assumed for further processing (Algorithm 5).

Algorithm 5: Co-citations identification based on seed article citations

```
Input: Filtered citations (corpus of research article)
Output: Co-citations research article.
1. The co-citation from the citations are identified and
   downloaded.
2. Check if the co-citation research paper is after seed
   paper's publication date.
3. If step 2 is No, exclude that co-citation from the co-
   citation list.
4. If Step 2 is Yes, Check if the co-citation research paper
   is already a direct citation
5. If step 3 is yes, exclude that co-citation from the co-
   citation list.
6. If step 3 is no, get the summarized document using the
   stacked auto encoder.
7. Perform the topic modeling for the summarized
   document.
8. Find the similarity between seed paper topics and the
   co-citation summarized documents topic
9. Filter the relevant citations using DE.
```

The co-citations would not have a direct citation connection with the seed article; additionally, there would be more than one co-citee present in a single context; therefore, it is quite impractical to obtain the co-citation context as well.

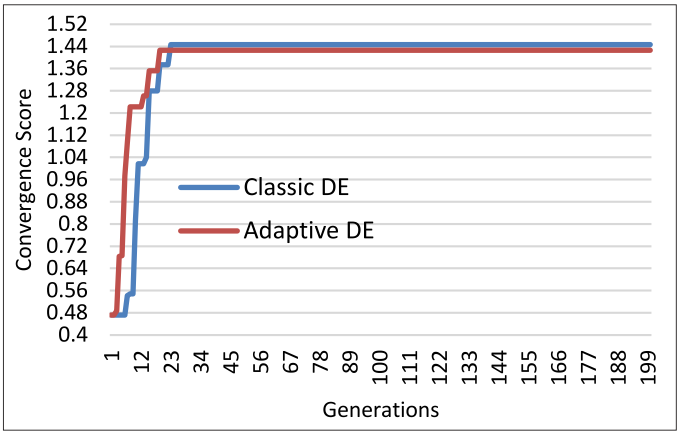

FIGURE 4. Convergence of adaptive DE with classic DE on finding optimum context similarity threshold.

Algorithm 6: Deep Topic Modeling of Co-citations

```
Input: Co-Citations research paper
Intermediate Output: Summarized text for all co-citation
research articles
Output: Topics generated for summarized text.
Methodology: Stacked Auto encoder with Greedy Layer
wise Training.
Pseudocode:
1. Remove stop words using the nltktokener 'token-
   izers/punkt/english.pickle'nltk.data.load('tokenizers/
   punkt/english.pickle')
2. Get the tf-idf document –term matrix using TfidfVector-
   izer of sklearn.feature_extraction.text
3. Training Phase:
   a. Train the outer most encoder using sparseAutoen-
      coder with the input shape and hidden layer.
   b. Output of this layer is the dot product of the weight
      and the input.
   c. Train the second outer most encoder with the hidden
      layer 1 and hidden layer 2 size.
   d. Output of this layer is the dot product of the weight
      and the previous layer.
   e. Train the second and third most encoder with the hid-
      den layer 2 and 3.
   f. Output of this layer is the dot product of the weight
      and the previous layer.
   g. Return the concept space as output.
4. Obtain the cosine similarity between the data and
   model.
5. Sort and get the top few sentences as the deep sentences.
6. Run the HDP topic modeling for summarized text.
```

Yet, there lies an indirect method of computing the same. All the co-citeearticles would have some message in common with the seed article, as mentioned by the co-citing research articles. This idea inspired us to obtain the full text of the co-citees excluding references. Since references were not part of the main theme of discussion we chose to ignore them at present. The co-citee full texts are summarized using Deep

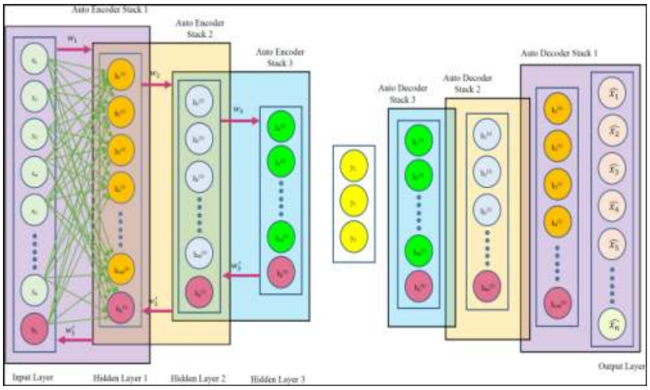

FIGURE 5. **Three-Layer deep sparse auto encoder [21].**

Stacked Auto Encoder [21] which use unlabeled data in a complete unsupervised environment in order to build a compressed representation of input data. The compressed and ranked sentences at the output are subjected to HDP topic modeling [21]. In other words, DSA-H topic modeling or Deep topic modeling is performed over co-citee articles and further subjected to topic based similarity analysis.

DSA-H topic model [21] handles meagre input yet provide more interpretable topics as compared to traditional LDA. The reason is that instead of LDA, generative model HDP is used for topic modeling. Since HDP is very generative in nature and is capable of producing large volume of topics, care is taken by the auto encoder to feed only the important segments of the research article as input. DSA-H stacked auto encoder utilizes three layers of hidden layer stack before arriving at the output layer (refer Figure 5) [21].

The auto encoder pre-trained with restricted Boltzmann machine (RBM) [36], learns the inputs and further aims at considerable reduction of input dimension at every hidden layer thereby learning generative models of data. Provision of hidden layers is what contributes towards the sparseness of deep auto encoder. The improvement in average topic coherence of DSA-H for a sample journal full-text articles is given in Figure 6. The topic coherence for assumed dataset on DSA-H model is 4.41. Therefore it is clear that the topics are identified with DSA-H in a more appreciable manner.

With the citations and respective co-citations filtered in a more intelligent manner, the entire corpus of citations as well as co-citations are represented in the form of a citation network. Nodes are research articles and edges are citation/co-citation relations. Further, the graph is repetitively mined for cross-citation links existing among the citation/co-citation nodes. Special links are maintained for labeling citations as well as co-citations. Several papers were both in the citing and co-citing segments of the seed paper, which is an indication that those articles were of utmost importance with respect to content lineage of the seed article. Therefore, utmost care is taken to retain the multitude importance of edges across the articles.

The citation network graph is constructed in a stepwise fashion handling three different article relations: citation, co-citation and cross-citation, one at a time. The graph formed possessed 3700 nodes and 3699 primary edges. These primary edges are formed with only citation relations. The sample graph obtained is shown in Figure 6. The graph is improvised with embedding the co-citations at the next layer. The co-citation edges are superimposed over the citation graph obtained in previous step (refer Figure 6). At this level there were 105 nodes suggested by the co-citation identification process. Among these, 52 nodes already existed as part of citation graph, and 53 nodes were new suggestions. Therefore, newly suggested 53 nodes were added to the network along with respective edges. The cross-citation analysis resulted in 10844 additional cross-citing edges among the nodes of citation network. These 10844 edges were mined over both citing and co-citing nodes and the multitude of relationships between pair of nodes is retained for further processing (refer Figure 6).

The cross-citation analysis further revealed interesting insights into the parallel work progressed so far and the visualization is obtained and presented yearwise (Figure 7 and See Figure A1, Table 4. It is exciting to see that the parallel work in the theme of h-index or related metrics increases every year. Most important revelation obtained is that, in 2005, at the same time of the seed article [6] that proposes h-index was published, there were other competing articles just published or in the pipeline (refer Table 4) but which failed to attract enough research attention as like the seed article.

The network is further embedded with rhetorical citation relations across edges. There might be more than one rhetoric present if the relating nodes are in multiple contexts, which are labelled completely. For enabling rhetoric labeling, we assumed the 12 citation classification categories of Simon Teufel [37] (refer Table 5). The 12 citation categories are

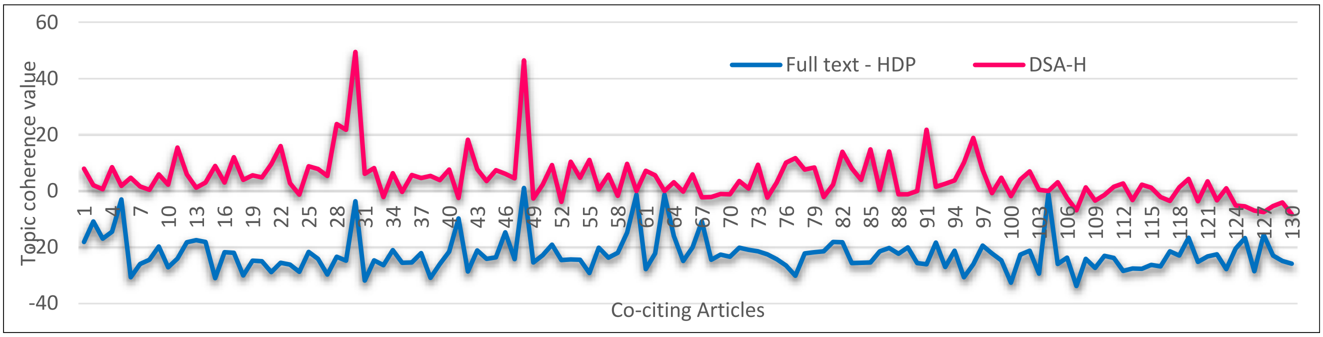

FIGURE 6. **Topic Coherence (DSA-H) for Co-citing articles of rhetoric citation network.**

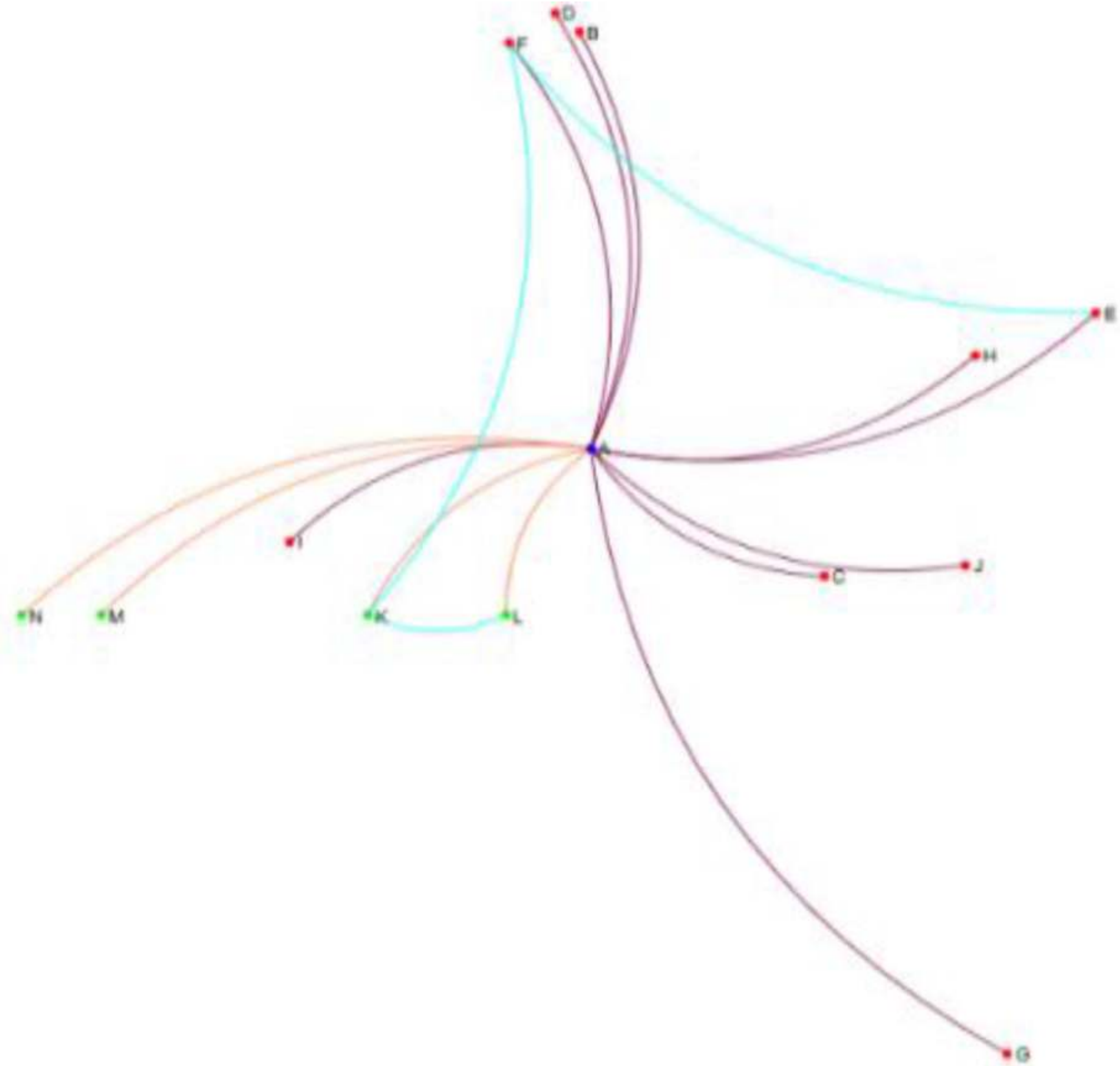

FIGURE 7. Parallel work identified (2005) clockwise–Nodes: blue–seed article; green–co-citing article. Edges: Pink–Citations; Orange–co-citations; Blue–Parallel work.

TABLE 4. Parallel work in 2005.

S.No.	Article Title	Citations
1	Modified index to quantify individual's scientific research output	0
2	Facts from text—is text mining ready to deliver?	168
3	An index to quantify an individual's scientific research valid across disciplines	10
4	Robert Van de Walle	0
5	A Hirsch-type index for journals	289
6	Ten challenges to transform taxonomy	20
7	A parameter to quantify dynamics of a researcher's scientific activity	37
8	Measures and mismeasures of scientific quality	24
9	On the opportunities and limitations of the H-index	231
10	Index aims for fair ranking of scientists	460
11	Biologist Helps Students Get a Leg Up on Scientific Inquiry	4
12	A parameter to quantify dynamics of a researcher's scientific activity	37

TABLE 5. Citation categories of simon teufel [37].

S.No.	Category	Description
1	Weak	Weakness of cited approach
2	CoCoGM	Contrast/Comparison in Goals or Methods(neutral)
3	CoCo-	Author's work is stated to be superior to cited work
4	CoCoR0	Contrast/Comparison in Results (neutral)
5	CoCoXY	Contrast between 2 cited methods
6	PBas	Author uses cited work as basis or starting point
7	PUse	Author uses tools/algorithms/data/definitions
8	PModi	Author adapts or modifies tools/algorithms/data
9	PMot	This citation is positive about approach used or problem addressed (used to motivate work in current paper)
10	PSim	Author's work and cited work are similar
11	PSup	Author's work and cited work are compatible/provide support for each other
12	Neut	Neutral description of cited work, or not enough textual evidence for above categories, or unlisted citation function

TABLE 6. Proposed rhetoric citation sentiment categories.

S.No.	Citation Category	Proposed Citation Sentiments
1	PMot, PUse, PBas, PModi, PSim, PSup	Positive
2	Weak, CoCo-	Negative
3	CoCoGM, CoCoR0, CoCoXY, Neut	Neutral

further classified into positive, negative and neutral (refer Table 6). The rhetoric labels are identified by the matching cue phrases as recommended by Simon Teufel [37]. Upon absence of cue phrases within the context, the rhetorics is assumed to be neutral. Therefore, neutral rhetorical category consists of both citation categorised (as per Table 6) as well as dumb contexts as well. The statistics on rhetorics of the citation network is presented in Table 7. The statistics on rhetoric sentiments of citation network is presented in Table 8. The following section summarises the deliverables obtained from the citation network.

V. PROPOSED INDICES OF RHETORIC CITATION QUALITY

A. AVAILABILITY INDEX

The corpus is assessed for availability of citations which is measured as ‘Availability Index’, denoted as α. It is defined as the ratio of accessible citations to that of total citations of that article.

$$\alpha = \frac{M}{N}, \tag{1}$$

where, M is the no. of accessible citations, and N, the total citations. The availability index for Hirsch [6] is 0.68.

B. DEEP CITEINDEX

The corpus is assessed for deep semantic relevance of citation contexts. It is denoted as the ratio of retained citations with respect to the total citations of the seed article, analysed from context perspective. Deep cite is denoted as τ.

$$\tau = \frac{R}{N}, \tag{2}$$

TABLE 7. Rhetoric citation network statistics.

Year/ Rhetorics	CoCoR0	neutral	Pmot	Psim	Psup	Puse	Weak
2005	3	2	2			1	
2006	5	3	1		4	22	5
2007	10	5	7	1	7	52	7
2008	17	19	6		8	91	14
2009	37	17	10	1	9	112	29
2010	42	34	10		12	169	39
2011	44	25	12	2	15	193	42
2012	44	31	12		14	177	54
2013	58	50	10		14	236	55
2014	59	49	16		16	256	50
2015	37	55	7	2	10	233	48
2016	57	45	15	1	11	247	57
Grand Total	413	335	108	7	120	1789	400

TABLE 8. Rhetoric citation network sentiment statistics.

Year/ Rhetoric Sentiments	Negative	Neutral	Positive
2005		5	3
2006	5	8	27
2007	7	15	67
2008	14	36	105
2009	29	54	132
2010	39	76	191
2011	42	69	222
2012	54	75	203
2013	55	108	260
2014	50	108	288
2015	48	92	252
2016	57	102	274
Grand Total	400	748	2024

where, R is the no. of semantically relevant citations, here, 3699, and N, the total citations, 5796. The semantic relevance is computed as 0.63.

C. MISS INDEX

The miss index (μ) is the ratio of articles available in the rhetoric citation network which are not actually the explicit citations of Hirsch [6]. The miss index is measured as 0.009.

D. EXPRESSIVE POWER

Expressive Power is a proposed metric which indicates the level of how the author’s idea is represented and carried forward via citing/co-citing articles. It is denoted as δ. Raw expressive power is denoted as the component adjusted with the ratio of idea carried by the seed article from its’ references. Normalised expressive power is the normalization of δ with respect to the availability index of seed article [6].

$$\delta = \sum_{j=1}^{M} \frac{cs_j}{M} \tag{3}$$

$$\delta_{raw} = \left(1 - \sum_{i=1}^{N} \frac{rs_i}{N}\right) * \sum_{j=1}^{M} \frac{cs_j}{M} \tag{4}$$

$$\delta_{norm} = \frac{\delta}{\left(\frac{\tau}{\alpha}\right)} \tag{5}$$

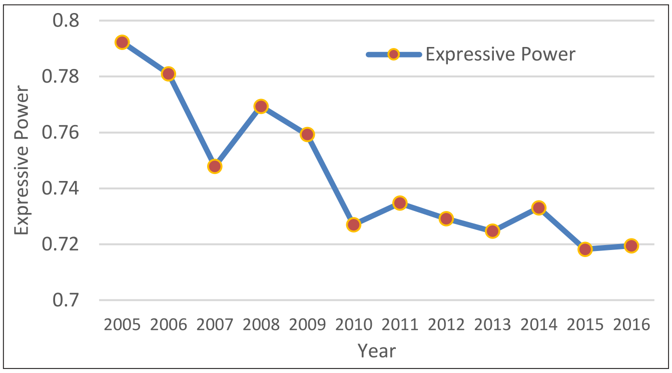

FIGURE 8. Change in expressive power over the years.

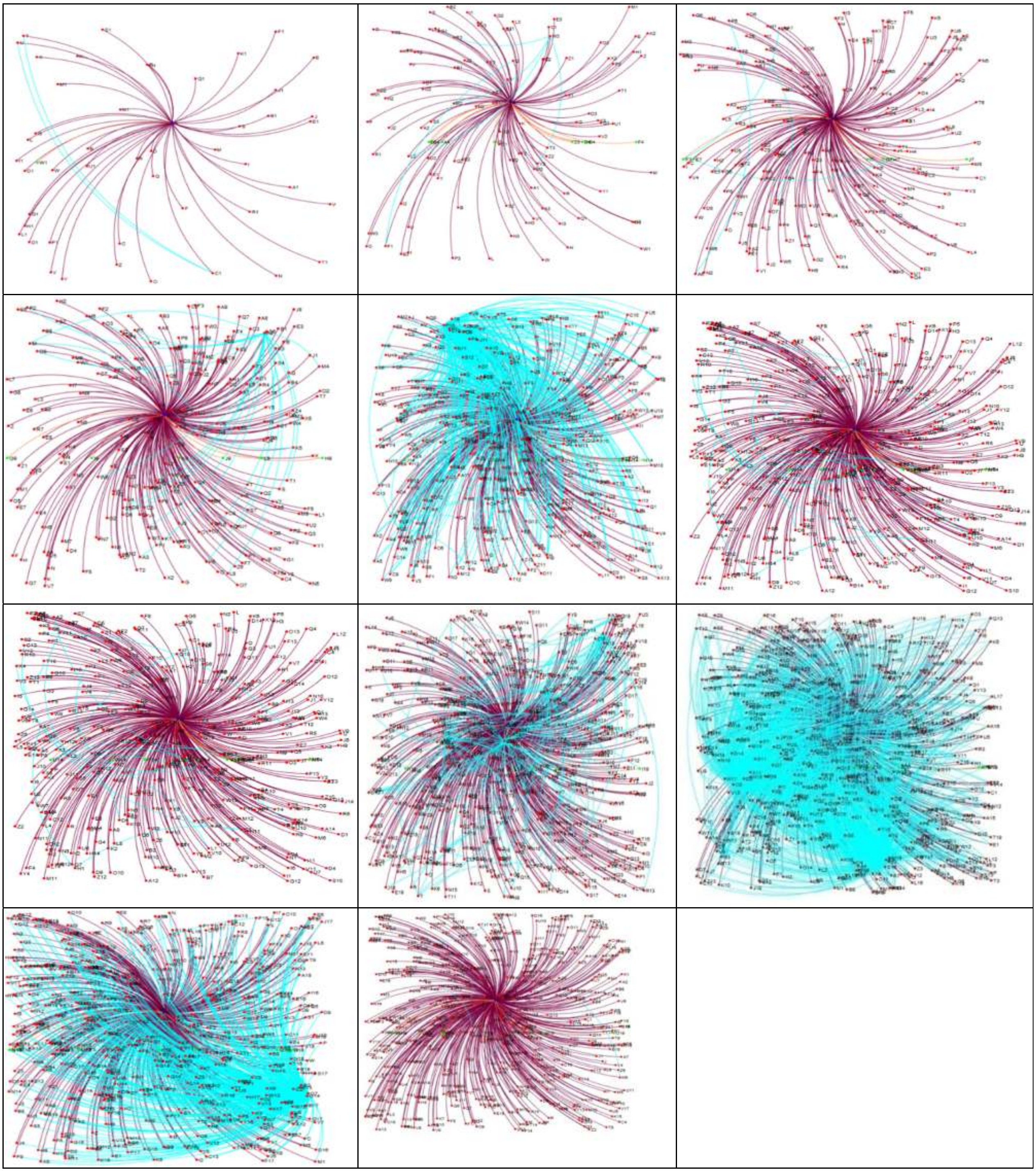

FIGURE A1. Parallel work identified (2006–2016) clockwise–Nodes: blue–seed article; green–co-citing article. Edges: Pink–Citations; Or-ange–co-citations; Blue–Parallel work

where δ–Expressive Power, N–No of References, M–No of Citations, rs–Similarity value of the references of seed article, cs–similarity value of the citations of seed article, α–availability index.

The expressive power is computed as 0.73 and the raw expressive power is 0.27. Normalised expressive power is 0.79. This reveals that inspite of voluminous citations overflowing for Hirsch [6] for more than 12 years, only 79 percent of its content is followed/challenged by the research literature. This indicates an interesting insight that remaining 21 percent of the article is yet to be challenged, which implies the strength of research idea conveyed in the seed article. In other words, δ_{norm} explored is greater than that of the unexplored of the seed research article.

Ideal curve for expressive power has to follow a downward slope as years passby. However, Figure 8 presents the fluctuation in expressive power over the years since seed article is published. The graph follows the ideal drop depicting saturatedness in analyzing/exploring the benefits and shortcomings of the article especially over 2015 and 2016. We postpone the discussion on analyzing the fluctuations as shift in interest on expression of article [6] as depicted by Figure 8 as future work.

This article proposes topic-model based methodologies for quantifying article expressive power in scientific manuscripts, for the first of its kind measurement on article quality analysis. Using expressive power, the semantic based author credits shall be measured in a more accurate manner. Expert author and expert community shall be identified using article expressive power. Using expressive power for measuring article quality would transform the concept of semantic measurement of scientific article into a world of measuring the useful idea communicated by the article into the research arena. Unlike h-index where experienced co-authors get greater credits, using expressive power will transform the semantic strength of the author of underlying research article which had eased the fellow researcher in terms of conception and knowledge dissemination. The entire dataset, meta-data and relevant information can be downloaded here: *https://github.com/gsmahalakshmi/Power-of-Scientific-Manuscripts.* Since the creation of dataset is very time-consuming and there is no such dataset readily available for download, creation of more examples for justifying the claim is quite a length process. However, inclusion of more example datasets from various domains to validate the claim raised in this article would provide additional avenues for further research.

VI. CONCLUSION

The paper proposes interesting analysis of research citations from unsupervised semantic analysis perspective. The idea proposed utilized deep learning techniques for semantic analysis. The paper also proposed various article level metrics like ‘deep cite’ and ‘expressive power’ which shall be used as bench mark metrics for measuring the semantic strength of research article. With deep semantic analysis, the self-citations are indirectly given a fair treatment. In other words, self-citations are neither ignored nor retained, but treated appropriately with respect to their matching deep semantics. Alternatively, the structure of citations like the order of literature being cited could also be copied from that of the base article. To incorporate these cite order semantics will bring interesting insights in the future. Further, constructing the citation graph of all generations of the research article would project an enriched information graph using which the citation lineage and longest research paths shall be analysed.

ACKNOWLEDGMENTS

This Publication is an outcome of the R&D work undertaken in the project under the Visvesvaraya Ph.D. Scheme (awarded to Dr. S. Sendhilkumar, Unique Awardee Number: VISPHD-MEITY-2959) of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

APPENDIX

See Figure A1.

REFERENCES

[1] X. Hu and R. Rousseau, “Scientific influence is not always visible: The phenomenon of under-cited influential publications,” *J. Inf.*, vol. 10, no. 4, pp. 1079–1091, 2016.

[2] P. O. Seglen, “Why the impact factor of journals should not be used for evaluating research,” *BMJ: Brit. Med. J.*, vol. 314, no. 7079, pp. 498–502, 1997.

[3] S. Saha, S. Saint, and D. A. Christakis, “Impact factor: A valid measure of journal quality?,” *J. Med. Library Assoc.*, vol. 91, no. 1, pp. 42–46, 2003.

[4] B. Alberts, “Impact factor distortions,” *Sci.*, vol. 340, 2013, Art. no. 787.

[5] P. Wouters, M. Thelwall, K. Kousha, L. Waltman, S. de Rijcke, A. Rushforth, and T. Franssen, The metric tide: Report of the independent review of the role of metrics in research assessment and management, London: HEFCE, 2015.

[6] J. E. Hirsch, “An index to quantify an individual’s scientific research output,” *Proc. Nat. Acad. Sci. United States America*, vol. 102, no. 46, pp. 16569–16572, 2005.

[7] J. Mingers and L. Yang, “Evaluating journal quality: A review of journal citation indicators and ranking in business and management,” *Eur. J. Oper. Res.*, vol. 257, n. 1, pp. 323–337, 2017.

[8] J. Wang, B. Thijs, and W. Glänzel, “Inter disciplinarity and impact: Distinct effects of variety, balance, and disparity,” *PLoS One*, vol. 10, no. 5, 2015, Art. no. e0127298.

[9] S. S R Siva and G. S. Mahalakshmi, “Discovering citation manipulations via delayed citation recognition,” *Asian J. Inf. Technol.*, vol. 15, no. 12, pp. 1964–1969, 2016.

[10] N. Onodera, “Properties of an index of citation durability of an article,” *J. Informetrics*, vol. 10, n. 4, pp. 981–1004, 2016,

[11] C. Min, J. Sun, L. Pei, and Y. Ding, “Measuring delayed recognition for papers: Uneven weighted summation and total citations,” *J. Informetrics*, vol. 10, no. 4, pp. 1153–1165, 2016.

[12] J. O. De Sordi, W. L. de Paulo, M. A. Meireles, M. C. de Azevedo, and L. H. C. Pinochet, “Proposal of indicators for the structural analysis of scientific articles,”*J. Informetrics*, vol. 11, no. 2, pp. 483–497, 2017.

[13] G. S. Mahalakshmi and S. Sendhilkumar, “Automatic reference tracking,” *Handbook of Research on Text and Web Mining Technologies*, 2009, pp. 483–499.

[14] S. Uddin and A. Khan, “The impact of author-selected keywords on citation counts,” *J. Informetrics*, vol. 10, no. 4, pp. 1166–1177, 2016.

[15] H. J. Kim, Y. K. Jeong, and M. Song, “Content- and proximity-based author co-citation analysis using citation sentences,” *J. Informetrics*, vol. 10, no. 4, pp. 95–966, 2016.

[16] A. Balaji, S. Sendhilkumar, and G. S. Mahalakshmi, “Finding related research papers using semantic and co-citation proximity analysis,” *J. Comput. Theoretical Nanoscience*, vol. 14, no. 6, pp. 2905–2909, Jun. 2017.

[17] R. Rousseau, R. Guns, A. I. M. J. Rahman, and T. C. E. Engels, “Measuring cognitive distance between publication portfolios,” *J. Informetrics*, vol. 11, no. 2, pp. 583–594, 2017.

[18] Z. Taşkın and U. Al, “A content-based citation analysis study based on text categorization,” *Scientometrics*, vol. 114, no. 1, pp. 335–357, 2018.

[19] D. Pride and P. Knoth, “Incidental or influential?–A decade of using text-mining for citation function classification,” in *Proc. 16th Int. Society Scientometrics Informetrics Conf.*, Oct. 16-20, 2017.

[20] R. Siva, G. S. Mahalakshmi, and S. Sendhilkumar, “1-hop greedy cite order plagiarism detection,” *Int. J. Control Theory Appl.*, vol. 10, no. 8, pp. 585–588, 2017.

[21] G. S. Mahalakshmi, G. MuthuSelvi, S. Sendhilkumar, P. Vijayakumar, Y. Zhu, and V. Chang, “Sustainable computing based deep learning framework for writing research manuscripts,” *IEEE Trans. Sustainable Comput.*, pp. 1–16, 2018, doi: 10.1109/TSUSC.2018.2829196.

[22] C. Lu, Y. Ding, and C. Zhang, “Understanding the impact change of a highly cited article: A content-based citation analysis,” *Scientometrics*, vol. 112, 2017, Art. no. 927.

[23] G. S. Mahalakshmi, G. MuthuSelvi, and S. Sendhilkumar, “Generation of author topic models using LDA,” *Computational Vision and Bio Inspired Computing*. Berlin, Germany: Springer, 2017.

[24] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, “Learning author-topic models from text corpora,” *ACM Trans. Inf. Syst.*, vol. 28, no. 1, Jan. 2010, Art. no. 4, 38 pages.

[25] D. M. Blei, “Probabilistic topic models,” *Commun. ACM*, vol. 55, vol. 4, pp. 77–84, 2012.

[26] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 113–120.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[28] S. Kataria, P. Mitra, C. Caragea, and C. L. Giles, “Context sensitive topic models for author influence in document networks,” in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, vol. 22, no. 3, p. 2274, 2011.

[29] D. Card, C. Tan, and N. A. Smith, “A neural framework for generalized topic models,” in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, arXiv preprint arXiv:1705.09296, vol. 1, 2017.

[30] J.-T. Chien and C.-H. Lee, “Deep unfolding for topic models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 318–331, Feb. 1, 2018.

[31] J. Yin and J. Wang, “A dirichlet multinomial mixture model-based approach for short text clustering,” in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 233–242.

[32] G. S. Mahalakshmi, G. MuthuSelvi, and S. Sendhilkumar, “Hierarchical modeling approaches for generating author blueprints,” in *Proc. Int. Conf. Smart Innovations Commun. Comput. Sci.*, Jun. 2017, pp. 411–422.

[33] A. Klahold, P. Uhr, F. Ansari, and M. Fathi, “Using word association to detect multi-topic structures in text documents,” *IEEE Intell. Syst.*, vol. 29, no. 5, pp. 40–46, Sep./Oct. 2013.

[34] S. M. Islam, S. Das, S. Ghosh, S. Roy, and P. N. Suganthan, “An adaptive differential evolution algorithm with novel mutation and crossover strategies for global numerical optimization,” *IEEE Trans. Syst. Man Cybern. B Cybern.*, vol. 42, no. 2, pp. 482–500, Apr. 2012.

[35] S. H. Cha, “Comprehensive survey on distance/similarity measures between probability density functions,” *Int. J. Math. Models Methods Appl. Sci.*, vol. 4, no. 1, pp. 300–307, 2007.

[36] M. Wong, B. Farooq, and G.-A. Bilodeau, “Discriminative conditional restricted Boltzmann machine for discrete choice and latent variable modelling,” *Journal of choice modelling*, vol. 29, pp. 152–168, 2018.

[37] S. Teufel, A. Siddharthan, and D. Tidhar, “Automatic classification of citation function,” in *Proc. Conf. Empirical Methods Natural Language Process.*, 2006, pp. 103–110.

G. S. MAHALAKSHMI received the Master’s degree in CSE from Anna University, Chennai, and the PhD degree in the field of Artificial Intelligence, in 2009. She has authored numerous research articles in Reputed Journals and International Conferences. Presently she is an assistant professor (Senior Grade) in the Department of Computer Science & Engineering, Anna University, Chennai. Her research interests include machine learning, social networks, text mining and big data analytics.

R. SIVA received the Master’s degree in engineering in the discipline of computer science and engineering from Manonmaniam Sundaranar University, in 2004. He is currently working toward the PhD degree under the same discipline from Anna University. He has published a number of research works in National and International Conferences. He is currently working as an Associate Professor with the KCG College of Technology, Chennai, Tamilnadu, India. His areas of research include text mining, bibliometrics and data mining.

S. SENDHILKUMAR received the PhD degree in web search personalisation, in 2009. Currently he is an assistant professor (Senior Grade) in the Department of Information Science & Technology, Anna University, Chennai. He has authored numerous research articles in Reputed Journals and International Conferences. His research interests include data mining, social networks, text mining and big data analytics.