Received May 28, 2020, accepted June 27, 2020, date of publication July 6, 2020, date of current version July 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007558

Impact of Field of Study Trend on Scientific Articles

LUBNA ZAFAR AND NAYYER MASOOD
Department of Computer Science, Capital University of Science and Technology, Islamabad 75400, Pakistan
Corresponding author: Lubna Zafar (lubbnaa@gmail.com)

ABSTRACT The volume and diversity of scientific literature are escalating every day and millions of new scientific articles are published every year. Researchers work and publish in their respective fields of interest. A major portion of the scientific community publishing in the same field of interest forms a trend in the field which could be deemed as relatively more popular than other trends. A pioneering researcher picks a field of interest by depending upon its popularity. This may have a positive impact on the acceptance of a study or high count of citations in future. This study identifies how significant it is to follow a research trend and the impact of the field of study (FoS) trend on research paper citations. For this purpose, we have chosen the field of Computer Science and Microsoft Academic Graph dataset from the 2007-2015 time period. In the dataset, every paper has a list of fields of study. The FoS provided in MAG is systematized hierarchically into 4 levels; level-0 – level-3. In this study, we apply the clustering technique to the FoS and citations pattern separately. Likewise, we also analyze how papers following a FoS trend, gain citations over the time. We also introduce a novel method Field of Study Multigraph (FoM) using graph centrality measures degree, betweenness and closeness to analyze the FoS trend, citation trend, and the relation between research areas in scientific articles from the domain of Computer Science. The experimental results show that the FoS has a certain impact on citation count. Furthermore, the results depicts that if papers belong to the same FoS, then there are 66% of the chances of having a similar citation pattern and that they have the same citation trend as they also achieved a high correlation value. This proves that a FoS has a certain impact on the citation count of a paper and researchers should contemplate the FoS trend before selecting a particular research area.

INDEX TERMS Field of study trend, citation trend, clustering, computer science.

I. INTRODUCTION

The volume and diversity of scientific literature are increasing at an exponential rate due to enormous inventions in science. Almost 2.5 million scientific articles are published every year and the amount gets doubled after every five-year [1], [2]. The articles are published by different venues, such as conferences and journals and released to the wider community by digital libraries such as Google Scholar, Citeseer, DBLP etc. These libraries index the publications in a hierarchical manner wherein each node of a hierarchy corresponds to a particular field of study (FoS) [10]. The dynamic increase in the research plethora has made it difficult for the scientific community to discover hidden patterns from a particular field of study (FoS). The FoS determines the area of focus of a particular scientific article. For instance, a paper focusing on comparison between different machine learning algorithms like Naïve Bayes, Support Vector Machine etc. will belong to the FoS, ‘‘Machine Learning’’ or ‘‘Artificial Intelligence’’ [10]. Typically, the inclination of the scientific community towards certain fields of study (FoS) is more among other fields due to emerging of trends in the field. Due to the dynamic increase in the research plethora, it becomes difficult for the scientific community to detect trends in a particular FoS. A research trend is a the research general direction followed by researchers during a specified period of time and is defined as, an area that is evolving and grabbing importance over time [1]. Publications by a large group of researchers in the same FoS may form a trend, resulting in increased popularity of the FoS among other fields.

A pioneering researcher typically opts for a field that is more popular or its trends are being followed by the wider scientific community. This is done based on an assumption that contemplation of these aspects may increase the acceptance probability of the piece of work done in the trendy FoS, and further lead towards the rapid gain of citations

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

in future [3]. In scientific literature, following the research trends and dynamics can hold noteworthy benefits and this is significant to specify the interest of researchers. Following a research trend does not mean traditionalism to plan at great or yielding to symmetry; rather it indicates having dynamic awareness regarding innovative intuition [3].

To date, the scientific community has presented different studies to determine flow or trend of scientific literature. The traditional trend detection-based studies rely on bibliometric indices involving key parameters like publication count and citation count [3]. The prime purpose of these studies is to discover the topic areas that are growing in interest over the time; trend analysis is performed to collect information and discover a pattern from data [2]. In literature, the centrality measures has been extensively studied in the field of social network analysis [15]. Surprisingly, a very few attempts have been made to apply such measures for keywords extraction, wherein degree centrality is used [25]. The citation count is considered as one of the potential bibliometric indices to discover trend or flow of a research [3]. These studies have overlooked a few important aspects which are the focus of our study.

To the best of our knowledge: (i) there are no published experiments on: (i) the significance of FoS trend following, (ii) centrality measures to analyze FoS trend and their relationships and if (iii) researchers follow FoS trend, this creates a high impact on research paper citations. This study uses the scientific articles published in the domain of Computer Science and analyses:

RQ1: *What is the impact of FoS on citation count by evaluating how similar is the citation trend of papers belonging to the same fields?*

RQ2: *Can we use any measure other than citation count to detect the trend of FoS?*

RQ3: *Is there any relationship between different fields of study?*

One major difficulty in addressing these issues is the availability of relevant data, that is, a worthy source of dataset is required. Since bibliographic datasets having features like title, authors, conference, and journal information in the field of Computer Science are not so hard to acquire as DBLP [4] freely provides this metadata in a structured manner. However, features like citations, keywords, and FoS are harder to acquire as they are not available in the form as DBLP provides other features.

Discovering the FoS of a research paper is itself a research problem. Microsoft Academic Graph (MAG) is provides a rich source of dataset making it easier to acquire such a dataset [5]. Precisely, MAG has a study that depicts a relationship between research papers and their corresponding field of study (FoS) in a hierarchical manner [6]. In MAG, every paper has a list of FoS. The FoS in MAG are systematized hierarchically into 4 levels; level-0 – level-3 with level-0 being the most general FoS, e.g., Computer Science and level-3 being the most specific e.g., cluster analysis.

In this research, we use the MAG dataset of conference papers to analyze FoS trend and their impact on research paper citations. We perform clustering on FoS and citations pattern separately. We present a novel method Field of Study Multigraph (FoM), formed by using centrality measures, degree, betweenness and closeness to analyze the field of study trend, citation trend, and the relation between research areas in Computer Science scientific articles. The frequency of FoS in papers is also calculated to detect FoS trend. The study calculates a Rand Index to find the similarity between two data clustering's to analyze the impact of FoS on citation count.

Finally, we use the correlation coefficient to find the nature of a relation between FoS and citation patterns. The outcomes of the study revealed that the papers belonging to the same FoS have similar citations pattern. Furthermore, citations pattern can also be estimated against a particular FoS and if a paper belongs to the same field of study, then there are 66% of the chances that they have the same citation trend as they also achieved high correlation value. This proves that a field of study has a certain impact on the citation count of a paper and researchers should also contemplate the trend of a field of study while selecting a particular research area.

This paper is structured as follows; Section II discusses the related work and section III discusses the proposed methodology. Section IV examines experimental results and in section V we conclude the paper.

II. RELATED WORK

In literature, researchers have proposed different techniques for trend detection and analysis. A citation network is proposed for temporal ordering [7] of documents to detect topic evolution and embryonic trends from data and formerly use citations to calculate the loads for the key terms in papers. Research papers' data is synchronized to a classification of areas built on the important words from the titles and abstracts and is studied to capture the variations in the number of publications linked with such topics using a citation network [16]. However, as [8] pointed out, in citation network, keywords of research papers are not pre-processed and do not show the significance of research topic areas in various scenarios, different keywords of papers even present similar topics.

A network of co-occurring [9] keywords in scientific data and detected the growth in period of the link weights is used to identify trends and emergent research topic areas. Patent analysis, bibliometric study, and text-mining analysis techniques [10] are used to identify research trends. A method proposed [5] compares the scattering of keywords extracted from the research data using citation graphs associated with publications encompassing these keywords. The method assumes that if a keyword term is suitable for a topic area then the research papers encompassing the keyword will have a strong link to paper content. However, the technique is not well suited for areas which are in their early stage of the process.

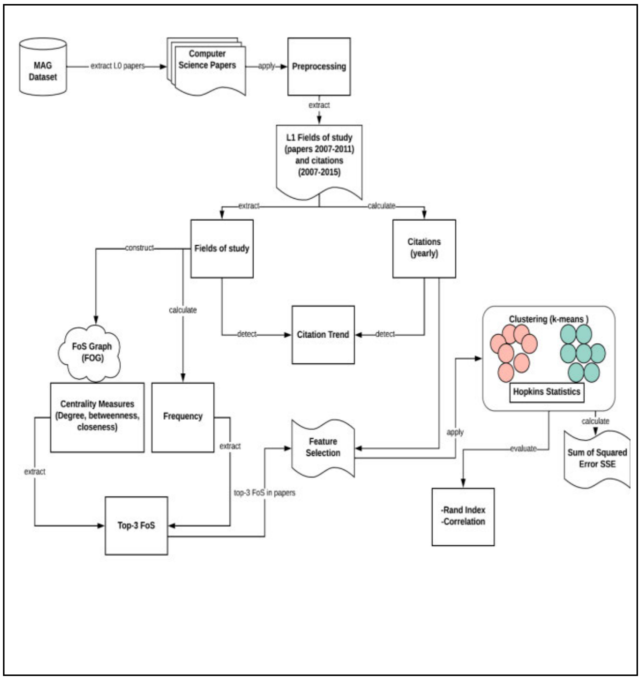

FIGURE 1. The proposed methodology.

A common technique is created on the usage of keywords as substitutes for research topics. In this situation, each keyword typically signifies a particular topic. This technique can be defined as a keyword-based topic model. A technique proposed by [11] analyzes keywords to detect trends in the scientific literature. Two-dimensional text mining approaches including clustering and bibliometric analysis of keywords is used to analyze the knowledge structure of scientific research of the journal. Similarly, there exists another approach that creates paper-topic relationships by using keywords and words mined from the abstract to study the trends of topics on diverse time scales [21].

To identify topic trends, it is promising to define a topic state according to features such as the number of associated publications/citations [12] the number of authors energetic in it [13] and accordingly observe their evolution over the time. Also, the relational topic modeling which chains network structure and LDA of papers to model topic areas unit citation networks and LDA [14] is used to discourse the issue of topic evolution.

The method identifies topics in autonomous subsets of data and influences citations to link topics in altered periods. A hybrid approach [15] detects the growth and decline in trends of research topics, however, it does not detect early research trends. The hybrid approach combines the PLSA for topic modeling in a window that slides through the stream of paper to study the topic growth. For a researcher, only selecting the location for its publication is not a good choice for paper acceptance, numerous aspects must also be considered, such as the listeners he/she is inscription for, the research topic area, and likewise the venue/location strategies [16]. Though, new procedures of self-archiving, like blogs, describe an exciting substitute, which is gradually castoff in a few research groups. Research articles are generally related to a titles of research topics, which are normally contingent by the keywords identified by the researchers [17] or mined from the manuscript with automatic approaches. Research topics are explored and examined by authors and their groups for several aims, such as determining innovative information and producing innovative approaches [18].

A research study identifies trends in CS especially its relationship with research funding [19] using ACM and IEEE papers with research fields based on ACM and IEEE classifications. In literature, keyword-based, graph-based, and bibliometric approaches are used for trend detection and analysis in scientific articles. The most common way to study the research trend is citation count [3]. However, analysis of FoS trends in the field of Computer Science has not been given adequate attention by the scientific community.

Our study is closely related to [19] which identifies FoS scores to investigate general publication trends, citation trends, the evolution of research areas in Computer Science. Based on the critical analysis of the literature review, the identified research gap is a lack of study that investigates the association between the trend of FoS and the new papers being written in that FoS. This may be a useful consideration, especially for new researchers in terms of making a decision to pick a particular FoS to conduct research. The gap has led us to formulate the following research questions (RQs):

RQ1: How similar is the citation trend of papers belonging to the same fields?

RQ2: Can we use any measure other than citation count to detect the trend of FoS?

RQ3: Is there any relationship between different fields of study?

III. METHODOLOGY

This section encompasses details about the proposed methodology. We have proposed two different methods to address the research questions. The details about the data set are discussed in sections 3.1 and 3.2. The clustering technique, FoS clusters, and citations trend are discussed in 3.3, 3.4, and 3.5.

Details about Field of Study Multigraph (FoM), formed with the help of centrality measures is discussed in sections 4 and 4.1. Figure 1 is a graphical representation of modules of the proposed methodology.

A. DATASET DESCRIPTION

The dataset employed for this study is taken from Microsoft academic[1] [6] and is known as Microsoft Academic Graph (MAG) dataset which contains information about different academic articles, fields of study, and the association between academic articles and fields of study. The academic articles include conference papers, journal papers, and books. The

[1] http://academic.research.microsoft.com

TABLE 1. MAG dataset count of multidiscipline and Computer Science entities.

Entity	Total Count	Computer Science Count
Papers	228,956,810	1,354,603
Authors	231.969,837	2,324,591
Conferences	4,414	1,277
Fields of Study(FoS)	50,007	9,800

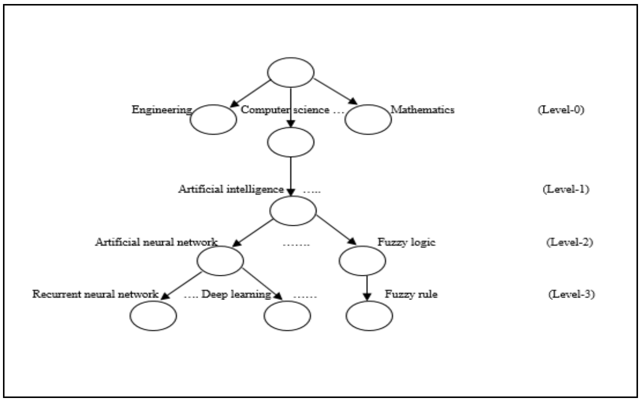

FIGURE 2. MAG different levels.

data about these articles include paper id, paper title, authors, etc.

The academic articles in MAG are from multiple fields of study (FoS) such as Physics, Computer Science,

Engineering, Chemistry, and many others. The statistics about overall data and data specific to Computer Science are thus, we can separate the topic of each paper without analyzing the abstract of the paper or the paper content itself.

Figure 2 above shows a snippet of the MAG hierarchy from level-0 to level-3. Level-0 contains FoS at a more generic level, like Engineering, Computer Science, etc. The lower levels contain more specific FoS as shown in the figure.

Every paper in MAG has a unique ID and is mapped to one or more associated FoS in the multiple levels of MAG hierarchy i.e. level-0 to level-3. An example of mapping is shown in figure 3 where a paper from the domain of Computer Science is mapped to different FoS from level 3 to level 0. In general, the hierarchy of the FoS is in the form of a directed acyclic graph, i.e. an FoS may have more than one parent FoS. For example, Cluster Analysis (level-3), belongs to Feature Selection (level-2) and Classification (level-2) which belongs to Machine Learning (level-1) and Computer Science (level-0). The level-1 FoS of CS has been shown in appendix A.

1) DATA PRE-PROCESSING

As explained earlier, the MAG dataset contains articles from different domains. For this study, we have selected the research papers from the field of Computer Science published during 2007 to 2015. Even though the MAG contains the papers that are published in journals and conferences. However, we have considered only conference papers as significant outcomes are frequently published initially in conferences [20]. This approach works on FoS of level-1 because it is the earliest and most generic distribution of FoS of a particular domain of knowledge [3]. The FoS in MAG becomes more specific when we move down in the hierarchy. After getting the level-1 FoS of CS papers, we store the paper id, publication year, paper title, FoS, level-0, and level-1 FoS associated with the paper in a separate file named as FoS dataset as shown in table 2.

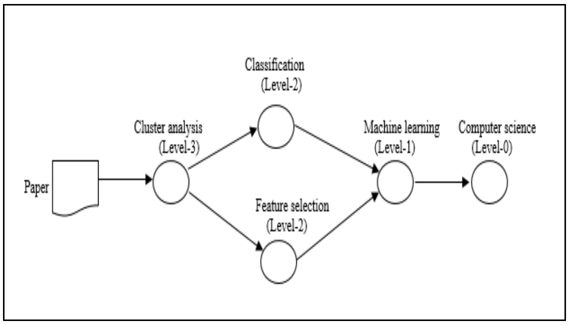

FIGURE 3. An example of Computer Science FoS levels.

TABLE 2. FoS of a sampled paper.

Paper ID	Year	Title	FoS	Level-0 (FoS)	Level-1 (FoS)
P1	2007	Joint optimization of relay strategies and resource allocations in cooperative cellular networks	cellular network, telecommunications, computer science, base station, resource management , operating system, wireless network, relay channel, computer network.	computer science	telecommunication,operating system, computer network.

To find out the association between the citations trend of papers and their corresponding FoS, we need to process our dataset to collect the yearly citation count of each paper and the number of publications for each FoS over the years. The MAG dataset does not contain the year-wise count of citations. For this purpose, we have selected those papers that have publication year between 2007 and 2011, and calculated the yearly citation count of each paper for the next five years, as shown in table 3 below.

In the above table, the first column shows the paper number and its publication year, the second column illustrates the

TABLE 3. Yearly citation count of five sampled papers.

ID	Level-1 FoS	Publication Year (PY) – Total Papers	Yearly Citation Count				
			PY +1	PY +2	PY +3	PY +4	PY +5
P^1_{2007}	telecommunications, operating system, computer networks	2007-5863	8	76	104	120	112
P^1_{2008}	world wide web, computer security, computer networks	2008-6599	1	19	21	22	21
P^1_{2009}	machine learning, data mining, artificial intelligence, simulation	2009-7159	0	1	2	0	0
P^1_{2010}	computer vision, simulation, artificial intelligence, machine learning	2010-7070	0	4	9	12	15
P^1_{2011}	data mining, database, machine learning, information retrieval	2011-6315	4	10	20	25	16

TABLE 4. FoS citation count (frequency).

	Yearly Citation Count of Different Level-1 FoS of CS				
Level-1 FoS	2007	2008	2009	2010	2011
machine learning	1283	844	1214	1412	1733
data mining	979	1039	1775	1836	1144
computer vision	970	550	1023	1131	1108
artificial intelligence	919	887	1084	1084	1554
operating system	885	534	663	992	748
theoretical computer science	820	433	551	992	644

level-1 FoS associated with the paper. The third column contains the publishing year, the next five columns contain the citation count of papers over the next five years and the last column contains the papers of the data set in a year-wise manner. After calculating the citations pattern of an individual paper, we have calculated the citations pattern of each of 34 level-1 FoS of CS. For this purpose, we have summed the citation count of papers belonging to different FoS. Table 4 below shows the citations pattern of some of FoS over five years.

Finally, we have replaced the FoS associated with each paper with the citation count of FoS for the publication year of the paper. Out of those citation counts, we have picked the top three ones. The example of pre-processed data used to perform experiments is shown in table 5 below. In this table, five papers published in the year 2007, the citation count of the top three associated FoS for 2007, and the citation count of each paper for the next five years, are shown as an example. The prepared data set contains the papers published from 2007 to 2011. In the next section, we present our approach to investigate the similarity between FoS and citations pattern.

TABLE 5. Papers with their citation counts and those of associated FoS.

Paper ID	Yearly Citations of FoS			Yearly Citations of Papers				
	Top 1	Top 2	Top3	2007	2008	2009	2010	2011
P^1_{2007}	885	696	530	1	1	2	2	4
P^2_{2007}	884	854	696	5	7	8	12	14
P^3_{2007}	1283	979	919	4	1	3	5	5
P^4_{2007}	1283	970	919	1	3	1	1	3
P^5_{2007}	1283	979	745	3	7	12	15	20

B. CLUSTERING

We have applied the clustering technique to analyze the impact of FoS on citation count of papers. Clustering is a method of grouping similar patterns (commonly signified as a vector of measurements) into different clusters based on similarity. Clustering analysis is one of the key analytical methods in data mining. The clustering technique is mainly appropriate for the studies focusing on capturing inter-relationships amongst the data items [21]. This study forms two different sets of clusters to address the RQ1. In one set of clusters, a 5-year count of citations of papers is considered as the feature set and in the other set, we have used the citation count of top three level-1 FoS associated with papers. Thereafter, similarity between two sets of clusters is calculated using Rand Index and Correlation.

Before applying clustering, we first analyzed the clustering tendency of our dataset. For this purpose, **Hopkins Statistic** H is picked. This is a spatial statistic that tests the spatial randomness of a variable as distributed in a space [22]. This test is conducted iteratively using 0.5 as a threshold. If the value of H is less than 0.5, it means that data does not have statistically significant clusters. If the value of H is close to 1, this means that the data can significantly form clusters. We have computed H for our dataset separately on the citation pattern of papers and also the citation count of FoS. This has been computed year-wise for all the papers. All the values of H were more than 0.5 suggesting that our dataset tends to form meaningful clusters. Table 6 shows the value of H calculations.

As indicated by the values of H, our dataset has a reasonable tendency for clustering. We have applied k-means clustering on Computer Science papers for five different years with two different selected feature sets, which are yearly citation counts of corresponding FoS and papers' citation

TABLE 6. Hopkins statistic values for two feature sets.

Year	Citation Count of FoS	Citation Count of Papers
2007	0.7	0.6
2008	0.7	0.6
2009	0.7	0.6
2010	0.7	0.7
2011	0.7	0.7

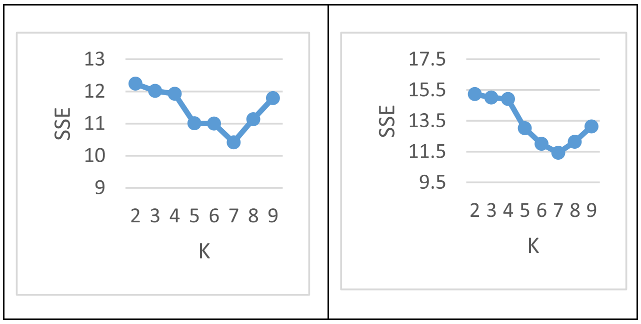

FIGURE 4. The relationship between SSE and the value of k for citation count(left), FoS(right).

counts as shown in table 6 (above). Afterward, the similarity between the two sets of clusters is calculated for evaluation.

K-means clustering [23] is a partition-based cluster analysis method. According to this algorithm, first, we have randomly selected k data values as initial cluster centers or centroids, then calculated a proximity metric (generally Euclidean distance) between each data value and each centroid and assigned it to the closest cluster, updated the averages of all clusters, repeated this process until the criterion is not matched.

K-means clustering aims to partition data into k clusters in which each data value belongs to the cluster with the nearest mean. The equation used for Euclidean distance is:

$$\mathrm{d} = \sum\nolimits_{k=1}^{k} \sum\nolimits_{i=1}^{n} ||x_i - u_k||^2 \tag{1}$$

where k signifies k cluster centers, u_k signifies the k^{th} center, and x_i represents the i^{th} the point in the dataset. The value of k, in K-means, is set by evaluating Sum of Squared Error (SSE) with different values of k generally starting from 2 and moving onwards. For our experiments, the graph between the value of k and corresponding SSE is shown in figure 4.

As per this diagram, the value of SSE falls with an increase in the value of k and it rises at 8. Therefore, we set the value k as 7. After applying k-Means clustering on citation counts of FoS with k equals to 7, a total of seven clusters has been formed.

C. FIELD OF STUDY CLUSTERS

The clustering results show the interaction of certain FoS with each other. We can see this with the interaction such as co-appearance of FoS in a research paper and similar FoS shows similar citation trends as they are clustered in the same group. We can see this with the interaction such as co-appearance of FoS in a research paper as shown in table 7. In particular, in research fields interdisciplinary interactions such as Machine Learning, Data Mining, Data Science, FoS may co-exist within one article, and the relationship between FoS may be important information. Therefore, it is essential to analyze the FoS that has a great influence on other FoS, such as the relationship between FoS, and the FoS that co-exists in articles.

TABLE 7. Grouping of different FoS based on the similarity of their citation count patterns.

Clusters	Field of Study (FoS)
Cluster0	Distributed Computing, Real-time Computing, Operating System, Parallel Computing.
Cluster1	Artificial Intelligence, Machine Learning, Computer Vision, Simulation.
Cluster2	Computer Security, Computer Networks, World Wide Web, Telecommunications.
Cluster3	Data Mining, Data Science, Database, Machine Learning.
Cluster4	Theoretical Computer Science, Algorithm, Computer Vision, Computer Graphics.
Cluster5	Operating System, Telecommunications, Computer Networks.
Cluster6	Machine Learning, Data Mining, Database, Information Retrieval.

As it can be seen from the above table that cluster0 comprises following FoS of level-1: "Distributed Computing, Real-time Computing, Parallel Computing, Operating System". It indicates that the citation pattern of these four FoS is common. These FoS usually occur together in the majority of research publications as Top-3 FoS. We can also observe that similar FoS shows similar citation trends of papers as they are clustered in the same group. Cluster1 comprises these FoS: "Computer Networks, Real-Time Computing, Operating System, Telecommunications" with the same interpretation and likewise the other groups.

The clustering results show that cluster0 comprises following level-1 FoS: "Distributed Computing, Real-time Computing, Parallel Computing, Operating System". These combinations look very natural, e.g., naturally, there is a possible relationship between the Distributed Computing, Real-time Computing, and Parallel Computing. These FoS usually occur together in the majority of research publications and both FoS seem to be more equal in terms of influence on each other. We can also observe that similar FoS shows similar citation trends of papers as they are clustered in the same group. Cluster1 comprises these FoS: "Computer Networks, Real-Time Computing, Operating System, Telecommunications". We have also generated 7 clusters based on the citations pattern of the papers as shown in table 7 (above).

D. EVALUATION METRIC

1) RAND INDEX

To find out the similarity between two sets of formed clusters, we have used the Rand Index (*RI*) which is defined as a measure of the percentage of correct decisions made by the

TABLE 8. Similarity between FoS and citation clusters from 2007–2011.

Publication Year of Paper	Duration of Citation Pattern	Value of Rand Index
2007	2007-2011	0.67
2008	2008-2012	0.67
2009	2009-2013	0.68
2010	2010-2014	0.67
2011	2011-2015	0.68

algorithm [24]. Rand Index gives a value between 0 and 1, where 1 means two clustering outcomes match identically. Rand Index can be calculated using the following formula [24];

$$RI = \frac{a + b}{a + b + c + d} \tag{2}$$

where, a: two similar documents to the same clusters, b: two dissimilar documents to different clusters, c: two similar documents to the different clusters, and d: two dissimilar documents to the same clusters. As can be seen from table 8 that there is a certain level of similarity between FoS and the citation pattern of papers. The FoS has a certain level.

2) CORRELATION

We have also computed the correlation coefficient to examine the relationship between FoS citations pattern. Correlation is one of the most common and useful statistics to examine the nature of the relationship between data items [25]. A positive correlation indicates the extent to which two variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - \sum (x)^2][[N\sum y^2 - \sum (y)^2]}} \tag{3}$$

The formula returns a value between -1 and 1, where: 1 indicates a strong positive relationship, -1 indicates a strong negative relationship, and the result of zero indicates no relationship at all. Where, N= number of pairs of scores, $\sum xy$ = sum of products of paired scores, $\sum x$ = sum of x scores, $\sum y$ = sum of y scores, $\sum x^2$ = sum of squared x scores, $\sum y^2$ = sum of squared y scores.

For this experiment, we have considered 5 years' citation counts of papers belonging to a particular FoS cluster. Out of these papers, we have taken a stratified random subset of 80% papers and used them as training data set and remaining 20% as a test set. In this way, 7 different training and test data sets have been formed which comprise of five years' average of citation count of papers belonging to the same cluster. These values are shown in table 9 below.

The values illustrated in the above table reveals that average citation count across multiple FoS is approximately similar. Next, to find the level of similarity among papers belonging to the same FoS, we have performed two steps: (1) we have calculated the correlation coefficient between training dataset of one year with test dataset of every other year and compared them.

(2) Then, we plotted the training dataset against the test dataset of the same year to graphically see the level of similarity between them. Table 10 below shows the correlation coefficient between different clusters' training dataset with each of the other clusters' test dataset. The highlighted values show that every cluster has the highest correlation with the test dataset of its cluster. This proves that the papers belonging to the same FoS have similar citation patterns and if we select a particular FoS to work in, then we can have an estimate of the citation pattern that we may receive on our work.

Figure 5 below shows the plots of training and test datasets of different clusters and citations pattern. The plots also show the similarity between the average citation trend of the same

TABLE 9. Average citation count of papers from 2007–2011.

	Yearly Average of Training Data Set				
Clusters	**2007**	**2008**	**2009**	**2010**	**2011**
cluster0	2.2	1.7	1.3	1.5	4.4
cluster1	2.2	1.5	1.4	1.5	4.4
cluster2	2.2	1.4	1.4	1.5	3.4
cluster3	2.3	2.5	1.2	1.5	4.4
cluster4	3.2	1.4	1.3	1.5	4.4
cluster5	2.2	1.5	1.3	1.5	4.4
cluster6	2.4	1.4	1.3	2.5	4.4

	Yearly Average of Test Data Set				
Clusters	**2007**	**2008**	**2009**	**2010**	**2011**
cluster0	2.2	1.4	1.2	1.5	4.4
cluster1	1.1	1.4	1.3	1.7	3.4
cluster2	1.1	1.4	1.3	1.3	2.5
cluster3	2.1	1.4	1.3	1.5	3.4
cluster4	2.1	1.4	1.3	1.5	4.4
cluster5	2.1	1.3	1.3	1.5	4.4
cluster6	1.1	1.4	1.3	1.5	3.4

TABLE 10. Correlation matrix of FoS citations from 2007–2011.

Clusters	Correlation						
	test	test	test	test	test	test	test
Cluster0 (training)	**0.99**	0.65	0.51	0.46	0.73	0.22	0.54
Cluster1 (training)	0.65	**0.92**	0.49	0.54	0.7	0.22	0.37
Cluster2 (training)	0.51	0.49	**0.83**	0.29	0.6	0.35	0.26
Cluster3 (training)	0.46	0.4	0.29	**0.91**	0.71	0.23	0.53
Cluster4 (training)	0.73	0.56	0.6	0.41	**0.93**	0.23	0.13
Cluster5 (training)	0.22	0.22	0.35	0.23	0.23	**0.97**	0.65
Cluster6 (training)	0.54	0.37	0.36	0.53	0.43	0.69	**0.87**

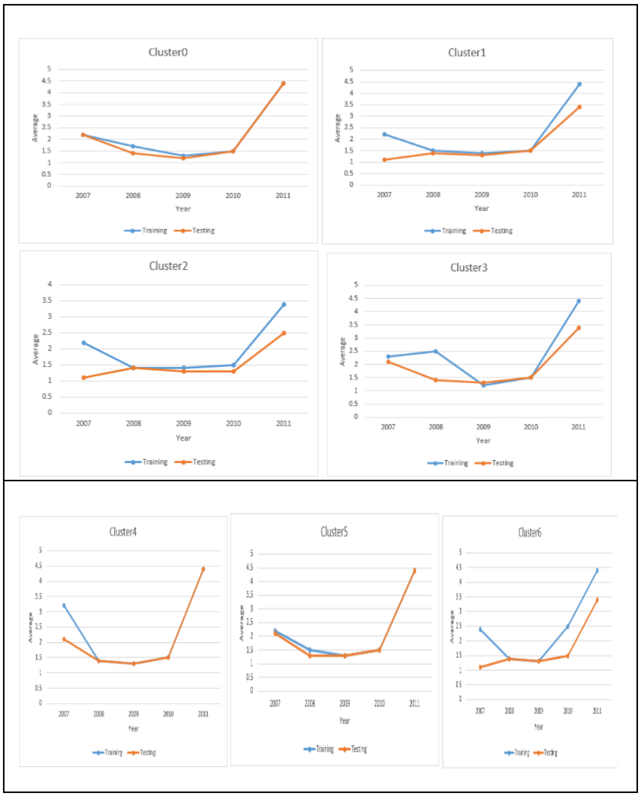

FIGURE 5. Training and test datasets of different clusters and citation patterns.

FoS. Moreover, the level of the correlation coefficient is also clear from the corresponding graph, for example, cluster0 has the maximum value of correlation coefficient which is also evident from the corresponding plot of figure 5, where both lines are almost identical.

The correlation result shows the papers belonging to the same FoS and following the trend have similar increasing or decreasing patterns of citations, as shown in figure 5. The experimental results show that FoS has a certain impact on citation count. Furthermore, a high count of citation depicts that if a paper belongs to the same FoS, then it may have the same citation trend. This proves that a field of study has a certain impact on citation count of a paper and researchers should also contemplate the trend of a field of study while selecting a particular research area.

E. FIELD OF STUDY TREND AND RELATION BETWEEN RESEARCH AREAS

In this paper, we use a multigraph with centrality measures to measure an FoS trend other than citation count (RQ2-3). Since most of the papers in our dataset correspond to more than one FoS, which establishes a link or relation between them. One possible approach to explore the significance or trend of an FoS other than the citation count could be the co-occurrence of an FoS with other FoS. More an FoS co-occurs with other FoS, more significant or trendy it is. The graph is a natural representation of such links between objects providing different centrality measures to measure the significance of objects within the graph.

For this purpose, we propose to construct an FoS multigraph (FoM) from the articles. Next, the trend of each FoS can be determined using graph centrality measures. In this study, we have applied three classic centrality measures (degree centrality, closeness centrality, and betweenness centrality). These centrality measures have been evaluated in the context of FoS. Lastly, these metrics are considered as FoS trend metrics and compared with the results obtained for the citation count (table 5).

1) FIELD OF STUDY MULTIGRAPH (FoM) CONSTRUCTION

A field of study multigraph (FoM) is built from the FoS of Computer Science papers. A multigraph is permitted to have multiple edges (also called parallel edges) between two nodes. Thus, two vertices (nodes) may be connected by more than one edge. A multigraph is a set of vertices, V, a set of edges, E, and a function $f\colon E \to \{\{u, v\}\colon u, v \in V \text{ and } u \neq v\}$. The significance of every FoS is then resolute using graph centrality measures and papers are categorized based on the FoS they comprise. The construction of the FoM graph is principally based on the FoS which are enclosed in a research paper and their vicinity. Each FoS that is enclosed within the research paper is signified by a system of a labeled node. The edges are focused to grab the structure of the FoS as they occur inside the research papers (relationship of FoS in the paper) as is illustrated in Figure 6. The nearness between the FoS is signified by the edges that join the nodes and is defined using an explicit extensive diversity of FoS. As an example, let us suppose three papers with their corresponding FoS, as given below.

Paper1 FoS: Algorithm, Computer Vision.

Paper2 FoS: Algorithm, Computer Vision, Data Mining, Machine Learning.

Paper3 FoS: Data Mining, Machine Learning.

The FoM for the above papers is shown in figure 6. In the above example, $f(e1) = f(e2)$, so we say e1 and e2 are multiple or parallel edges. However, the edges e2 and e7 are not called parallel edges. The FoM shows that Algorithm is connected to Computer Vision, Data Mining, and Machine Learning. Similarly, Computer Vision is connected to Algorithm, Data Mining, and Machine Learning. Data Mining is connected to Algorithm, Computer Vision, and Machine Learning. Likewise, Machine Learning is connected to Algorithm, Data Mining, and Computer Vision. Algorithm and Computer Vision has parallel edges (e1, e2) as these FoS have appeared in paper 1 and paper 2. Similarly, Data Mining and Machine Learning have parallel edges (e7, e8) as they appeared in paper 2 and paper 3. As soon as the FoM graph is constructed, centrality measures including degree, betweenness, and closeness are computed for each node by using the formulas shown in equation 4, 5 and 6 respectively.

2) CENTRALITY MEASURES

Once the FoM is constructed, centrality measures are computed to assign a score to each node. Let $G = (V, E, f)$ be

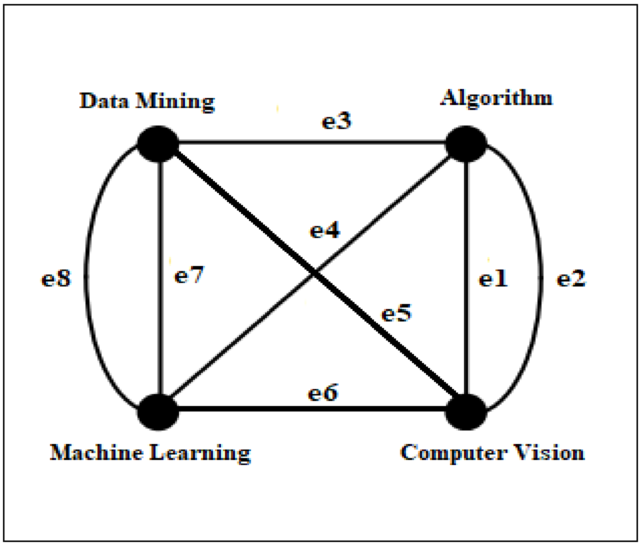

FIGURE 6. FoM construction.

a multigraph with a set of vertices (FoS) V, a set of edges E and f mapping edges between nodes. Starting with degree centrality, this section describes all the centrality measures employed in this study.

Degree centrality is defined as the number of edges incident upon a node. Applied to FoM, the degree of a node v_i represents the number of FoS that co-occur with the FoS equivalent to v_i. Let $C_D\ (v_i)$ be the degree centrality of a node v_i is given by [15]:

$$C_D(v_i) = \deg(v_i) \tag{4}$$

Generally, vertices with a higher degree or more connections tend to have a greater capacity to influence others. In the context of FoM, the value of degree centrality indicates the co-occurrence of a node (FoS) with other FoS in different papers which may be considered as influence or trend of that FoS.

Closeness centrality (or closeness) of a node is a measure of centrality in a connected graph, calculated as the sum of the length of the shortest paths between the node and all other nodes in the graph. Thus, the more central a node is, the closer it is to all other nodes in the network. Let distance (v_i, v_j) be the shortest distance between nodes v_i and v_j. The closeness centrality of a node v_i is [15]:

$$C_c(v_i) = \frac{1}{\sum_y distance(v_j, v_i)} \tag{5}$$

The degree centrality signifies the importance of a node (FoS) based on its direct connections with other nodes (FoS), whereas the closeness centrality covers both direct and indirect connections of an FoS showing how central a node in the FoM is.

Betweenness centrality is a measure of centrality in a graph based on the shortest paths. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through. This computes the number of times an FoS (node) behaves as a bridge alongside the shortest path between two other FoS (nodes). Here, $\sigma(s_t)$ is the total number of shortest paths from node s to node t and $\sigma(s_t, v)$ is the number of those paths that pass through v [15].

TABLE 11. Top-10 centrality measures.

FoS	Centrality Measures for the year 2007		
	Degree	Closeness	Betweenness
Algorithm	2150	0.9393939	0.01053
Artificial Intelligence	3200	0.9487532	0.0194119
Computer Networks	2925	0.9093939	0.0172436
Computer Vision	2680	0.9257143	0.0183707
Data Mining	2755	0.9211111	0.0182324
Database	2435	0.9193939	0.0100324
Machine Learning	3064	0.9117647	0.0194327
Operating System	2720	0.9293939	0.0105444
Theoretical Computer Science	2387	0.9387543	0.0128119
World Wide Web	2545	0.9193939	0.0191463

$$C_B(v) = \sum s \neq v \neq t \frac{\sigma(s_t, v)}{\sigma(s_t)} \tag{6}$$

Being between means that a node can act as a bridge to provide flow of knowledge between most of the nodes in a network. FoS with high betweenness are the pivots in the network knowledge flowing. The nodes with the highest betweenness also result in the largest increase in typical distance between others when they are removed. After constructing FoM, we calculated the degree centrality measures for all nodes of FoM (representing FoS) starting from the year 2007 till 2011. Table 11 below shows the values of centrality measures of top-ten trendy FoS for the year 2007.

IV. RESULTS AND DISCUSSION

A. TRENDY FoS

This section presents detailed analysis of the FoS that are selected as trendy FoS by FoM method using graph centrality measures. By analyzing the constructed FoM, we found the FoS with the highest degree, closeness, and betweenness to understand the trends of FoS over the time. Figure 7 shows the top-ten trendy FoS with degree centrality. Artificial Intelligence, Machine Learning, and Computer Networks have a maximum degree in 2007. Artificial Intelligence, Machine Learning, and Data Mining have achieved a high degree in 2008.

Artificial Intelligence, Computer Vision and Data Mining have a high degree in 2009. Machine Learning, Computer Vision and Data Mining have a maximum high degree in 2010. However, Artificial Intelligence, Machine Learning, and Data Mining attained a high degree in 2011. Closeness

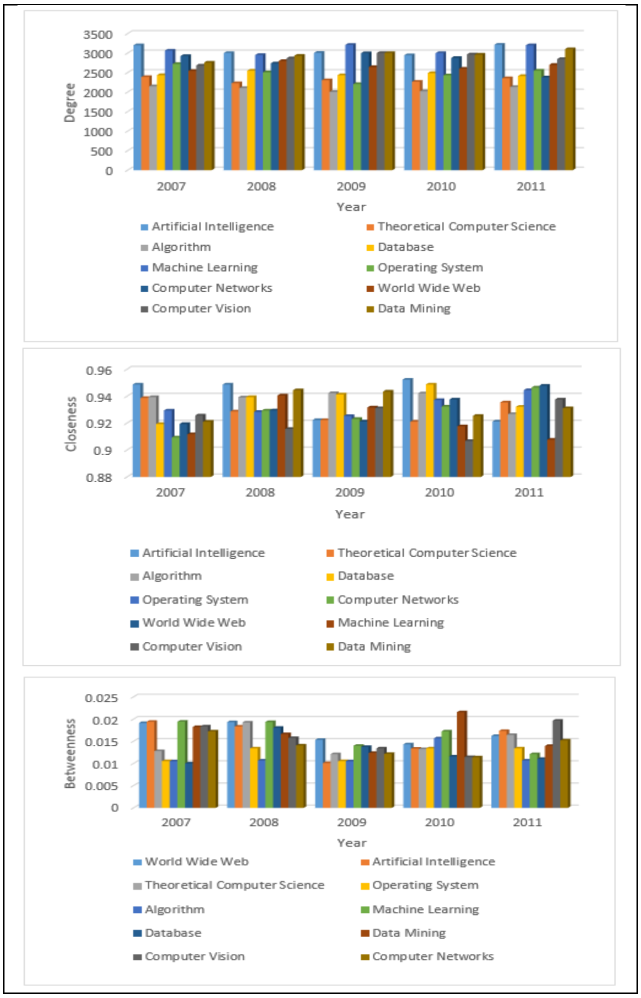

FIGURE 7. Top-10 Trendy FoS degree, closeness, betweenness.

centrality shows the top-10 trendy FoS as shown in Figure 7 Artificial Intelligence, Theoretical Computer Science, and Algorithm has a maximum value in 2007. Artificial Intelligence, Machine Learning, and Data Mining have a maximum value in 2008. Algorithm, Database, and Data Mining in 2009 has revealed the high value. Artificial Intelligence, Algorithm, and Database have achieved a maximum value in 2010. Whereas, Operating System, Computer Networks, and the World Wide Web has a maximum value in 2011.

Betweenness centrality shows the top-10 trendy FoS as shown in Figure 7. Artificial Intelligence and Machine Learning, World Wide Web has maximum betweenness value in 2007. World Wide Web, Theoretical Computer Science and Machine Learning have the highest value in 2008. World Wide Web, Machine Learning, and Database has a maximum value in 2009. Data Mining, Machine Learning, and World Wide Web achieved high value in 2010. Computer Vision, Artificial Intelligence, and Theoretical Computer Science have a maximum value in 2011.

1) TRENDY FoS CITATION TREND

Bibliometric analysis is used to identify citation trends from various aspects. Citation analysis is a bibliometric method used to reveal different patterns of the scientific community. Researchers can measure the significance of their publications with the help of citation analysis. They may gain facts about that paper's effect on its field by calculating the number of times it has been cited in research publications. Additionally, the citation trend is a good measure to analyze the impact of a research publication as high count of citation specifies usefulness and effectiveness.

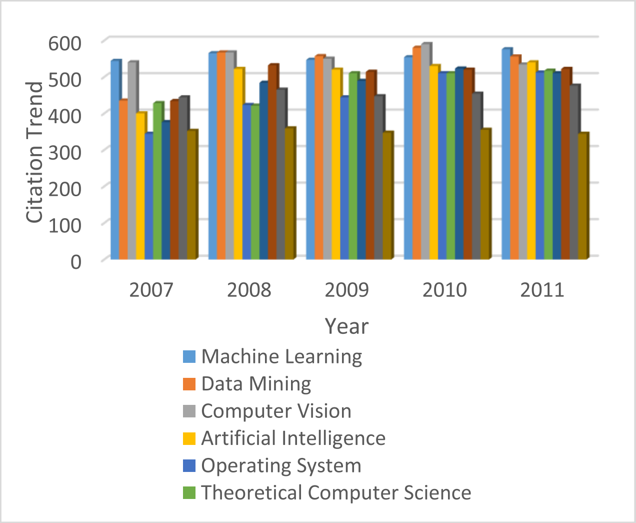

FIGURE 8. Top-10 trendy FoS citation trend.

A citation trend p is the collection of citation sequences sharing a common pattern of evolution of citation count. Citation sequences of various citation trends show different evolutions of citation count [26]. A citation-sequence of a research paper p, indicated as, $s_{\Delta t}(p) = [c_1(p), c_2(p), \ldots c_{\Delta t}(p)]$ is a sequence of citation count $c_i(p)$ over a period of time 1,2,3,…t, where c_i is the citation count of the i^{th} year after p gets published. For a collection of research papers, given a paper $p \in P$, its citation count c (p) is the number of papers that cite p, denoted by, c$(p) = |\{p' \in P : p' cites\ p|$.

An FoS receiving high citation count may be considered the most influential FoS in its discipline [26]. Here, the goal of bibliometric classification is to evaluate the citation trend of top-10 FoS in the Computer Science area. Counting the number of citations for each paper (where top FoS appears) and then calculate total citations of top-10 FoS gives the FoS citation trend, as shown in figure 8. This exposes the impact and worth of the scientific research field. Machine Learning, Data Mining, and Computer Vision have the highest citation trend in 2007, 2008, and 2010. Machine Learning, Artificial Intelligence, and Computer Networks have achieved maximum citation trend in 2009. Whereas, Machine Learning, Data Mining, and Artificial Intelligence have the highest citation trend in 2011.

2) THE SIMILARITY BETWEEN TRENDY FOS AND CITATION CLUSTERS

This section explains the FoS that are selected as trendy FoS by FoM method using graph centrality measures and

TABLE 12. Top-10 FoS order in 2007.

Level-1 FoS	Order of Top 10 FoS w.r.t Different Metrics in 2007			
	Frequency	Degree	Closeness	Betweenness
Machine learning	1	2	9	4
Computer vision	2	6	5	7
Operating system	3	5	4	3
Database	4	8	7	5
World wide web	5	7	8	1
Data mining	6	4	6	6
Artificial intelligence	7	1	1	9
Theoretical computer science	8	9	3	2
Computer networks	9	3	10	8
Algorithm	10	10	2	10

TABLE 13. Similarity between FoS and citation clusters from 2007–2011.

FoS year	Citation	Rand Index			
		Frequency	Degree	Betweenne -ss	Closeness
2007	2007-2011	0.67	0.68	0.63	0.61
2008	2008-2012	0.67	0.68	0.63	0.62
2009	2009-2013	0.68	0.68	0.64	0.62
2010	2010-2014	0.67	0.69	0.63	0.61
2011	2011-2015	0.68	0.69	0.63	0.60

frequency. The measure Rand Index to compute the similarity between two data clustering i.e., FoS, and citations clusters. An interesting fact that has been noticed while analyzing the values of different metrics is that the top-10 FoS across multiple metrics are the same, however, their order among the top-10 values is different. Table 12 below shows the ordering of top-10 trendy FoS across multiple metrics.

After this, we have applied our clustering experiments for each of the three centrality measures as done previously for the frequency of FoS, mentioned in the previous section. Then, we computed RI for each case and compared the resulting values for each other. The RI values of four metrics are illustrated in table 13 below and are shown in the form of a graph in figure 9.

The RI results show a reasonable level of similarity between clustering based on FoS and four different measures, i.e., frequency, degree, betweenness, closeness. Frequency and degree centrality have relatively higher values of RI as compared to the other two and out of these two- degree centralities has the highest RI values across multiple years. As results indicate that degree has achieved the highest RI value 0.69. The results indicate that if the papers belong to the same FoS, then there are 66% of chances, they have the same citation trend. This proves that a field of study has a certain impact on citation count of a paper and researchers should also contemplate on the trend of a field of study while selecting a particular research area. Also, the degree centrality is a more suitable metric to measure the trend of an FoS than a simple citation count.

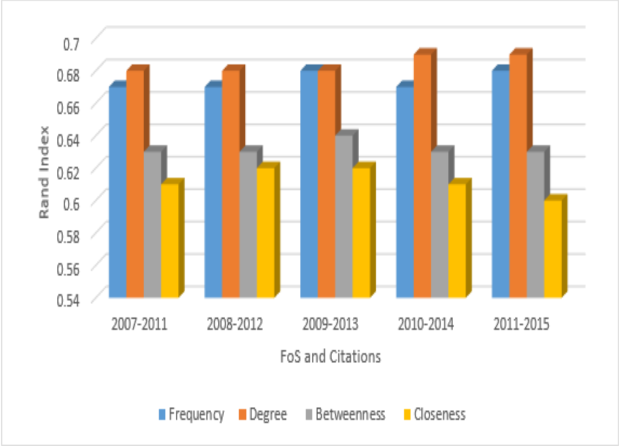

FIGURE 9. Rand index of frequency, degree, betweenness, and closeness.

V. CONCLUSION AND FUTURE WORK

This study has analyzed the effects of following a trend, how significant is to follow a research trend in the field of Computer Science area and the impact of FoS trend on research paper citations. We have employed the Microsoft Academic Graph (MAG) of research papers published during the years 2007-2011. In MAG, every paper has a list of FoS. The study has presented a rigorous analysis of three important aspects pertaining to scientific trend detection: (1) similarity between citation trend of papers belonging to the same fields, (2) An alternate to citation count measure for trend detection in FoS and nature of relation between the FoS that belonging to the same fields.

We have introduced a novel FoS multigraph (FoM) technique to detect the trends in FoS and analyzed the trends with the help of centrality measures and frequency. The trendy FoS over a specific time are discovered by analyzing the constructed FoM and frequency. The FoS in MAG are systematized hierarchically into 4 levels; level-0 – level-3. In this study, we have applied the clustering technique on level-1 FoS and citations pattern separately. The Rand Index has been used to find the similarity between two data clustering, and correlation coefficient has been employed to find the relationship between FoS citations pattern.

The experimental results show that there is a similarity between clusters formed on the basis of FoS and citations pattern and there also exists a relationship between FoS cita-

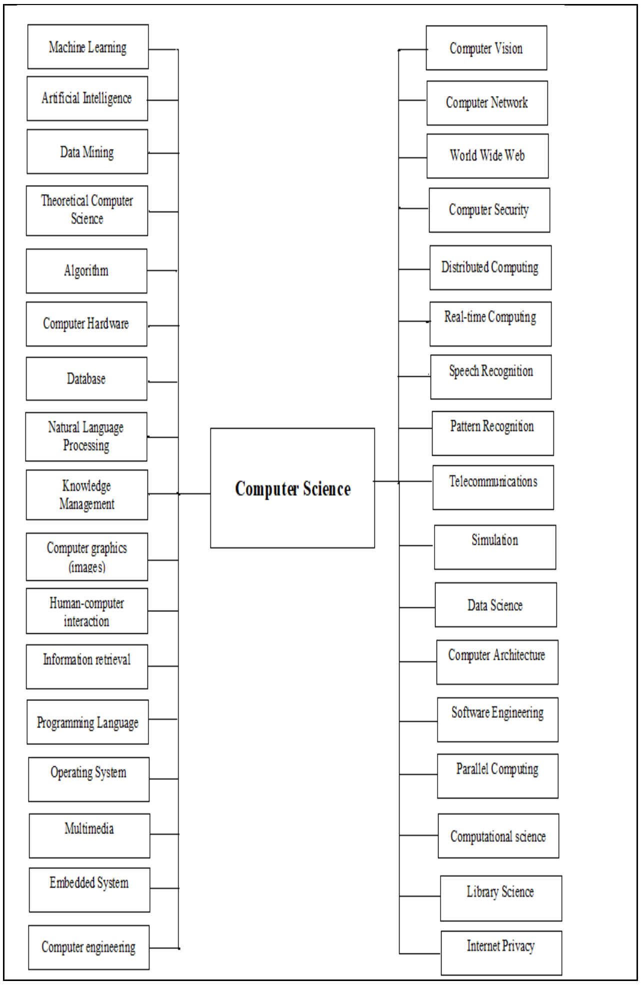

FIGURE 10. **MAG level-1 FoS.**

tions pattern that belong to the same FoS. The results indicate that FoS holds a certain impact on the citation count. Further, if the papers belong to the same FoS, then there are 66% of chances that they hold a same citation trend as they also achieved high correlation value.

This proves that an FoS has a certain impact on the citation count of a research paper and researchers need to consider the trend of an FoS while selecting a particular research area. The study shows that the established approach is general and might be practical to achieve knowledge of different research fields. For future studies, we will apply the author-topic model, a probabilistic model to connect authors to detected FoS in the scientific literature, which will show the common structure for study.

APPENDIX A

See Figure 10.

REFERENCES

[1] A. Hoonlor, B. K. Szymanski, and M. J. Zaki, "Trends in computer science research," *Commun. ACM*, vol. 56, no. 10, pp. 74–83, Oct. 2013.

[2] D. Hicks, P. Wouters, L. Waltman, S. De Rijcke, and I. Rafols, "The Leiden Manifesto for research metrics," *Nature*, vol. 520, no. 7548, p. 429, 2015.

[3] J.-C. Valderrama-Zurian, C. Navarro-Molina, R. Aguilar-Moya, D. Melero-Fuentes, and R. Aleixandre-Benavent, "Trends in scientific research in online information Review. Part 2. Mapping the scientific knowledge through bibliometric and social network analyses," 2017, *arXiv:1709.07817*. [Online]. Available: http://arxiv.org/abs/1709.07817

[4] M. Ley. (2008). *DBLP Computer Science Bibliography*. [Online]. Available: http//dblp.uni-trier.de/

[5] S. Effendy and R. H. C. Yap, "The problem of categorizing conferences in computer science," in *Proc. Int. Conf. Theory Pract. Digit. Lib.*, 2016, pp. 447–450.

[6] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J.- Hsu, and K. Wang, "An overview of microsoft academic service (MAS) and applications," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 243–246.

[7] N. Pobiedina and R. Ichise, "Predicting citation counts for academic literature using graph pattern mining," in *Proc. Int. Conf. Ind., Eng. Appl. Appl. Intell. Syst.*, 2014, pp. 109–119.

[8] L. Bolelli, E. rtekin. ; and C. L. Giles, "Topic and trend detection in text collections using latent Dirichlet allocation," in *Proc. Eur. Conf. Inf. Retr.*, 2009, pp. 776–780.

[9] F. Osborne and E. Motta, "Mining semantic relations between research areas," in *Proc. Int. Semantic Web Conf.*, 2012, pp. 410–426.

[10] B. Nie and S. Sun, "Using text mining techniques to identify research trends: A case study of design research," *Appl. Sci.*, vol. 7, no. 4, p. 401, Apr. 2017.

[11] F. Monaghan, G. Bordea, K. Samp, and P. Buitelaar, "Exploring your research: Sprinkling some saffron on semantic Web dog food," in *Proc. Semantic Web Challenge Int. Semantic Web Conf.*, vol. 117, 2010, pp. 420–435.

[12] D. Hicks, P. Wouters, L. Waltman, S. De Rijcke, and I. Rafols, "Bibliometrics: The Leiden Manifesto for research metrics," *Nat. News*, vol. 520, no. 7548, p. 429, 2015.

[13] F. Osborne, E. Motta, and P. Mulholland, "Exploring scholarly data with rexplore," in *Proc. Int. Semantic Web Conf.*, 2013, pp. 460–477.

[14] M. Mathioudakis and N. Koudas, "Twittermonitor: Trend detection over the Twitter stream," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2010, pp. 1155–1158.

[15] F. Boudin, "A comparison of centrality measures for graph-based keyphrase extraction," in *Proc. Int. Joint Conf. Natural Lang. Process. (IJCNLP)*, 2013, pp. 834–838.

[16] R. Sitarz, "Identification of research trends in the field of separation processes. Application of epidemiological model, citation analysis, text mining, and technical analysis of the financial markets," M.S. thesis, Lappeenranta Univ. Technol., Lappeenranta, Finland, Oct. 2013.

[17] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proc. 13th Conf. World Wide Web*, 2004, pp. 491–501.

[18] S. Yan and D. Lee, "Toward alternative measures for ranking venues: A case of database research community," in *Proc. Conf. Digit. Libraries*, 2007, pp. 235–244.

[19] S. Effendy, I. Jahja, and R. H. C. Yap, "Relatedness measures between conferences in computer science: A preliminary study based on DBLP," in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 1215–1220.

[20] J. A. Konstan and J. W. Davidson, "Should conferences meet journals and where?: A proposal for 'PACM,"' *Commun. ACM*, vol. 58, no. 9, p. 5, Aug. 2015.

[21] S. Mythili and E. Madhiya, "An analysis on clustering algorithms in data mining," *Int. J. Comput. Sci. Mob. Comput.*, vol. 3, no. 1, pp. 334–340, 2014.

[22] A. Adolfsson, M. Ackerman, and N. C. Brownstein, "To cluster, or not to cluster: An analysis of clusterability methods," *Pattern Recognit.*, vol. 88, pp. 13–26, Apr. 2019.

[23] S. Shinde and B. Tidke, "Improved K-means Algorithm for Searching Research Papers," *Int. J. Comput. Sci. Commun. Netw.*, vol. 4, no. 6, pp. 197–202, 2014.

[24] K. Y. Yeung and W. L. Ruzzo, "Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, May 2001.

[25] I. H. Wenno, "The correlation study of interest at physics and knowledge of mathematics basic concepts towards the ability to solve physics problems of 7th grade students at junior high school in ambon maluku province, indonesia," *Educ. Res. Int.*, vol. 2015, Apr. 2015, Art. no. 396750.

[26] C.-T. Li, Y.-J. Lin, R. Yan, and M.-Y. Yeh, "Trend-based citation count prediction for research articles," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, 2015, pp. 659–671.

LUBNA ZAFAR received the B.S. degree in computer science from Azad Jammu and Kashmir University, Rawalakot, in 2007, and the M.S. degree in computer science from International Islamic University, Islamabad, Pakistan. She is currently pursuing the Ph.D. degree in computer science with the Capital University of Science and Technology, Islamabad, Pakistan. Her research interests are data mining, graph theory, and machine learning.

NAYYER MASOOD received the Ph.D. degree from the University of Bradford, U.K., in 1999. He is currently serving as the HoD of the Capital University of Science and Technology, Islamabad, Pakistan. His current research interests are related to multidatabase systems, schema translation, schema evolution, schema integration, data integration, and data mining.