Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 06 Issue 02
2024

Research Article
https://doi.org/10.56979/602/2024

Citation Count Prediction of Scholarly Articles

Lubna Zafar[1], Nayyer Masood[2], Fazle Hadi[3], and Sheeraz Ahmed[4*]

[1]University of Poonch Rawalakot, 12350, Azad Kashmir, Pakistan.
[2]Capital University of Science and Technology, Islamabad, 44000, Pakistan.
[3]Higher Education Department, Khyber Pakhtunkhwa, 19110, Pakistan.
[4]Iqra National University, Peshawar, 25000, Pakistan.
[*]Corresponding Author: Sheeraz Ahmed. Email: sheeraz.ahmad@inu.edu.pk

Received: December 11, 2023 **Accepted:** February 09, 2024 **Published:** March 01, 2024

Abstract: Assessing the citation count is crucial for gauging the impact of scientific publications. Predicting future citation counts can assist researchers in discovering references and delineating research areas. In our study, we introduce a novel model called FoS Trend based Citation Count Prediction (FTCCP), which aims to forecast the citation count of scientific articles by leveraging field of study (FoS) trends and early citation counts. By analyzing the citation patterns within the first few years post-publication (specifically 1-3 years and 1-5 years), FTCCP extrapolates the long-term citation impact of an article. Notably, we focus solely on the FoS trend and Early Citation Count without considering other factors such as authorship, publication venue, or journal. While some prior research incorporates a broader range of features for citation prediction, we intentionally keep our model simple to ensure its applicability across diverse research domains.

Our investigation revolves around two feature categories for FTCCP: FoS trend and Early Citation Count. We employ Multiple Linear Regression to develop the citation count prediction model. Results from experiments conducted on the Microsoft Academic Graph (MAG) dataset demonstrate promising outcomes, indicating the effectiveness of FTCCP when utilizing FoS trend and Early Citation Count compared to models relying solely on citation history, as evidenced by higher R^2 scores. Furthermore, our proposed features exhibit superior performance compared to traditional ones.

Keywords: Citation Count Prediction; Scientific Impact; Field of Study Trend; Regression.

1. Introduction

There is a growing amount of research signifying the impact of citation in scholarly articles [1]. As yet, research studies in this area have mainly interested on the citation dynamics of papers [2], collaboration networks [3], and scientific impact prediction [4]. Predicting the future impact of scientific articles, often measured through citation counts, poses a significant challenge. Citation count serves as a common indicator for forecasting a paper's influence. However, accurately predicting citation counts requires identifying the key features that influence them. It's important to recognize that scientific articles exhibit diverse citation patterns. For instance, some articles may remain unnoticed for years before garnering significant attention (referred to as "Sleeping Beauty in Science") [5], while others may gradually lose citations due to the emergence of new methodologies.

As a result, employing a uniform or straightforward method may not adequately predict the future citations of a research article. More advanced models are required to comprehend the intricacies of citation dynamics. Research into predicting citation counts has primarily concentrated on two key areas: employing diverse machine learning techniques and utilizing specific feature sets for prediction objectives.

Several prominent machine learning algorithms include Support Vector Machine (SVM) [6], XGBoost [7], Gradient Boosting Decision Tree [8], among others. On the other hand, some of the features mainly used for this purpose include journal impact factor, journal reputation, venue, early citations, author's authority, age and topic of the paper [9]. Our focus in this paper is to suggest some novel features to be used for the citation count prediction that have been ignored so far. We believe that the domain or area of the paper could be an important feature to be exploited in this regard. However, to the best of our knowledge the scientific community not considered field of study (FoS) trend feature for prediction of citation count.

The Field of Study (FoS) delineates the particular focus area of a scientific article. For example, an article that contrasts different Machine Learning algorithms such as Support Vector Machine and Naïve Bayes would be categorized under the FoS labels of "Artificial Intelligence" or "Machine Learning" [10]. A research trend represents a common direction followed by researchers over a defined period, signifying an area that is gaining attention and growing over time. When a group of researchers publishes scientific works within the same FoS, it may contribute to the emergence of a research trend, thereby increasing the popularity of the FoS within various fields.

This study presents a new method termed FoS Trend-based Citation Count Prediction (FTCCP) aimed at predicting the citation count of a scientific article. It relies on the trend of the article's field of study and the number of citations it accrues during its early publication years. Essentially, this approach utilizes the field of study and citation count of a paper during its initial years after publication (particularly within 1-3 years and 1-5 years) to anticipate its citation trajectory over an extended duration.

In this study, we do not study other features such as author, venue and journal features. Although some studies in literature comprise more features for predicting citation count, however, We constrain our analysis to the Field of Study (FoS) trend and early citation patterns of publications to maintain research simplicity and applicability across various domains. We utilize Multiple Linear Regression (MLR) to forecast the citation count of each scientific article based on chosen features.

Our research aims to address the following questions:

RQ1: Can we accurately forecast a paper's future impact based on its Field of Study trend and early citation counts after publication?

RQ2: What are the primary features that influence whether a paper will receive a substantial num ber of citations in the future?

Our contributions can be summarized as follows, encapsulated within the proposed model FTCCP:

1. We forecast future citation counts utilizing the characteristics of FoS trend and Early Citation Count.
2. To gauge the importance of each feature, specifically FoS trend and Early Citation Count, we utilize MLR. Initially, we gauge the predictive capability of the model solely based on Early Citation Count. Following that, we evaluate the FoS trend feature, and ultimately, we integrate both features for prediction.
3. Our experimental results illustrate the effectiveness of our proposed approach.

The structure of the paper is as follows: Section 2 provides a review of pertinent literature. Section 3 elaborates on the features and delineates our prediction methodology. Section 4 showcases the outcomes of our prediction approach. Lastly, Section 5 offers conclusions drawn from our findings.

2. Literature Review and Background Study

In literature, numerous studies exist concerning the prediction of the influence of scientific works. Eugene Garfield introduced the concept of an impact factor to quantify the significance of a journal [11]. This factor represents the average number of citations received by research articles published within the preceding two years [12]. However, it fails to assess the impact of individual research papers [13]. J. E. Hirsch proposed the h-index as a metric to evaluate the scientific productivity of a researcher, which is defined as the number of their research articles with citation numbers greater than or equal to h [14].

However, the h-index is limited in its ability to measure the specific scientific impact as it overlooks the varying contributions of co-authors. To address this limitation, Stallings et al. introduced Pin-dexes, A-, and C-indexes, which calculate weighted sums of peer-reviewed papers, collaboration, productivity, and journal impact factors, respectively [15]. These approaches are all centered around citation counts. In recent years, numerous researchers have delved into predicting the citation counts of research articles. S. Bethard et al. investigated citation behavior and made noteworthy observations [16].

Brody et al. analyzed the short-term citation impact of web usage to predict medium-term citation impact [17]. Castillo et al. explored how the authors of a research paper can aid in predicting its future citations [17]. Lokker et al. utilized journal features and a sample of 20 research papers to demonstrate that citation counts can be consistently predicted up to two years after publication using data available within three weeks of publication [18]. Fu et al. predicted citation counts of biomedical research articles using predictive data available at the time of publication up to ten years into the future [19].

Alfonso et al. employed predictive models such as logistic regression, Bayesian networks, and decision trees to forecast the citation count of a research paper within four years after publication [20]. Various researchers employ diverse prediction algorithms and features to forecast the citation counts of papers using larger datasets. Callaham et al. utilized a decision tree algorithm to predict citation counts for 204 papers from a medical specialty meeting [21]. Kulkarni et al. employed linear regression to examine the citation count of 328 medical research papers during 1999-2000 [22].

McGovern et al. undertook a citation count prediction task using 30,199 research papers and their citations [23]. Yan et al. investigated various predictive models and numerous features correlated with citation count to forecast the citation count of papers [24]. Notably, Yan's study does not utilize any features from the citation network, and their experiment does not ideally predict the short-term impact of papers.

Livne et al. utilize the Microsoft Academic Search (MAS) dataset, encompassing 38 million papers [25]. MAS categorizes the dataset into 18 primary academic fields and subsequently predicts the citation count of publications. The eigenvalues employed in those studies are not accessible post-publication, posing difficulty in accessing such features. In this research, we focus on using Field of Study (FoS) trend and Early Citation Count to forecast the citation count of scientific articles. While some recent studies have addressed the same issue with early citation count as a feature [9] [26], our study specifically considers FoS trend and Early Citation Count as features. This presents a notable challenge as it excludes other features such as author details, journal, and venue information.

In this study, we focus on Field of Study (FoS) trends and Early Citation Counts as predictors for citation counts. While similar problems have been addressed previously, we uniquely emphasize these features, excluding author, journal, and venue information.

Our research questions (RQs) are as follows:

RQ1: Can we accurately forecast a paper's future impact based on its Field of Study trend and early citation counts after publication?

RQ2: What are the primary features that influence whether a paper will receive a substantial num ber of citations in the future?

In response to these inquiries, we introduce a groundbreaking model named FoS Trend-based Citation Prediction (FTCCP). This model is crafted to accurately forecast the long-term citation count of a paper.

3. Research Methodology

In our proposed approach, FTCCP, we commence by retrieving fundamental details from papers within the dataset. This includes titles, authors, abstracts, citations, fields of study, and references. Subsequently, we proceed to extract predictive features, leveraging them to train a predictive model. This model, based on Multiple Linear Regression [27], is then utilized to forecast citation counts for papers. A notable aspect of our method lies in the extraction process of predictive features.

3.1 Data Set Description

In this study, we employ a dataset provided by Microsoft Academic [10], referred to as the Microsoft Academic Graph (MAG) dataset. This dataset encompasses a wide array of information regarding scientific papers, encompassing conference papers, journal papers, and books. MAG comprises diverse information about paper (e.g., paper_id, paper_title, authors, abstract, keywords, field of study, publisher, year, citation, and venue etc).

MAG encompasses papers across many disciplines like Computer Science, Biology, Physics and Mathematics etc. and statistics is given as; papers=228,956,810, authors=231.969,837 and Computer Science dataset statistics are as; papers=1,354,603 and authors=2,324,591 (as this study focuses on Computer Science).

Within the Microsoft Academic Graph (MAG), each paper is classified into its respective Field of Study (FoS), obviating the necessity to scrutinize the paper's abstract or content. These FoS categories in MAG are organized hierarchically across four levels: level-0 to level-3 [10]. This study specifically focuses on the level-1 FoS in Computer Science, as outlined in Appendix A.

In MAG, every paper is allotted a unique paper_id and is linked with multiple FoS across different levels of the MAG hierarchy, spanning from level0-level3. Level-0 denotes broader FoS categories such as Computer Science, while level3 represents more granular topics like Big Data Processing. A FoS may have multiple parent FoS and adheres to a Directed Acyclic Graph (DAG) structure. For example, in Figure 1, Big Data Processing (level3) is nested under Cloud Computing and Data Stream (level2), which in turn is situated under Computer Network (level1), with Computer Science (level0) serving as the primary field.

3.2 Field of Study and Citations Pattern Extraction

We have selected Computer Science Conference papers from the time period 2007-2017. However, selecting Computer Science papers we have to search the level0 FoS of every paper that have Computer Science in its associated FoS. Once the papers categorized under Computer Science, a level0 field in the FoS hierarchy, are identified, the next step involves exploring the FoS of these selected papers to detect the associated level1 FoS. Subsequently, we compile essential information including paper_id, paper_title, publication_year, FoS, level0, and level1 FoS linked with each paper. This data is then stored in a separate file referred to as the FoS table, illustrated in Table 1.

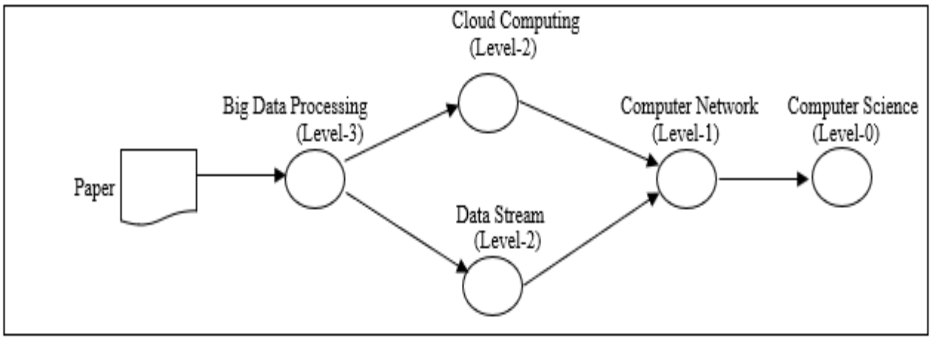

Paper
Big Data Processing (Level-3)
Cloud Computing (Level-2)
Data Stream (Level-2)
Computer Network (Level-1)
Computer Science (Level-0)

Figure 1. Computer Science FoS Levels Example

Following this, the citation count for each selected Computer Science paper is computed. Since the MAG dataset does not provide citation counts for each year, we calculate year-wise citations using citation data available (paper_id and references of each citing and cited paper). By scanning the references, we match paper_ids and calculate the count for each year. Subsequently, we record crucial details including paper_id, paper_title, publication_year, and year-wise citations (citation pattern) in a distinct file termed the citation table, depicted in Table 2. Additionally, preprocessing steps are undertaken, involving data cleaning and removal of stop words (such as books from the dataset).

Table 1. FoS of a sampled paper from 2007

Paper_Id	**Publication_Year**	**Paper_Title**	**Field of Study (FoS)**	**Level0 FoS**	**Level1 FoS**
P1	2007	User Security Behavior on Wireless Networks: An Empirical Study.	Wireless WAN, Heterogeneous network, Computer Science, Operating system, Wireless network, Internet security, Database, Security service, Key distribution in wireless sensor networks, Authorization, Internet privacy, Wi-Fi array, Cracking of wireless networks, Empirical	Computer science	Operating system, Database, World Wide Web, Computer network.

	research, Network Access Control, World Wide Web, Transport layer, Computer network.	

3.3 Citation Count Prediction

3.3.1 Problem Definition

Citation Count: In the context of research papers within the set P, denoted as p ∈ P, the citation count c(p) refers to the number of papers that reference or cite p. This is indicated as [28],

$$\text{c}(p)=|\{\, p' \in\ P: p' cites\ p\}|. \qquad (1)$$

Citation Sequence: In the context of a research paper p, a citation sequence refers to a series of citation counts c_i(p) observed over a timeframe extending from year 1 to t, where c_i represents the citation count in the i^th year following the publication of p. This concept is denoted as [28],

$$s_{\Delta t}(p) = [c_1(p), c_2(p), \dots c_{\Delta t}(p)] \qquad (2)$$

Citation Count Prediction: Given a research paper p has received $CC = cc_0, cc_1, \dots, cc_n$, citations after its publication, the aim is to learn a predictive function f to predict the field of study citation counts of a paper $CCP, = cc_{x+1}, cc_{x+2} \dots, cc_n$, for a paper ($x$ < n). after a given time period Δt, denoted as [28],

$$f(p|CC, \Delta t \rightarrow CCP(p|C, \Delta t\ ) \qquad (3)$$

In our study, we possess information regarding the citation count of paper p for the initial x + 1 years following its publication (ranging from the 0^th to the x^th year). We exclusively utilize these citations from the first x + 1 years as input data, without incorporating any additional information such as the authors' backgrounds, journal details, or venue information.

3.4 Feature Definition

3.4.1 Early Citation Count

The Early Citation Count (p) represents the count of citations received by a paper p during the initial years following its publication. Papers of superior quality generally accumulate a higher number of citations shortly after being published. This research concentrates on the early citations of research papers, particularly within the first few years after publication, such as within 1-3 years and 1-5 years, as delineated in the table provided below.

Table 2. Early citation count of 5 example papers

Paper_Id	Publication_Year	cc_0	cc_1	cc_2	cc_3	cc_4
P1	2007	8	76	104	120	112
P2	2008	1	19	21	22	21
P3	2009	0	1	2	0	0

P4	2010	0	4	9	12	15
P5	2011	4	10	20	25	16

3.4.2 Field of Study (FoS) trend

Topic modeling serves as a prevalent method for investigating the content of literature. The identification of a paper's Field of Study (FoS) poses a significant research challenge, and the Microsoft Academic Graph (MAG) offers a valuable dataset that facilitates obtaining such information. MAG incorporates a research study that categorizes the research areas of scientific articles into FoS [10]. Each paper in MAG is linked with a list of FoS, as previously mentioned in the dataset section.

It is understood that commonly used FoS tend to attract more attention, making papers associated with such FoS relatively easier to cite. To discern the trend of an FoS, we calculate the frequency of FoS occurrences for each paper, as depicted in Table 3. Additionally, we define the trend of an FoS as the count of its occurrences in papers over a specified period of time, denoted as t. The distribution T(d) of the field of study across all FoS in paper d can then be represented as:

Top of Form

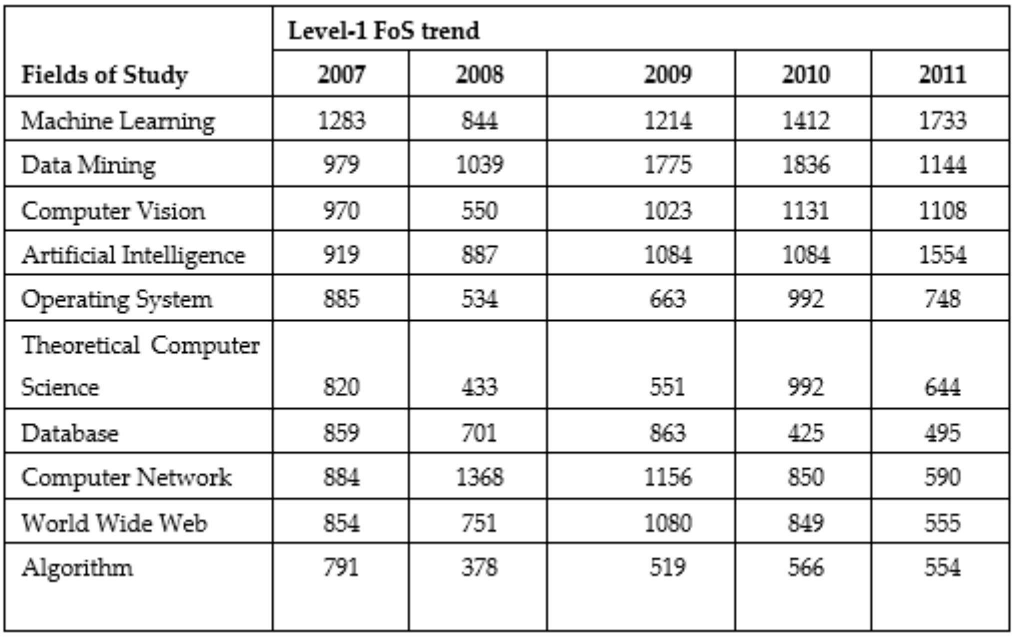
$$T(d) = \{\ (\text{field of study}_1|d),\ \ (\text{field of study}_2|d), \dots ,\ \ (\text{field of study }_z|d)\} \qquad (4)$$

and then we calculate the trend of an FoS in a paper d:

$$trend(field\ of\ study_i|d) \sum_{d\in C} count_T(\text{field of study}_i|d) \qquad (5)$$

(where $count_T(d)$ is the number of count of paper (d) and C is the complete corpus collection)

Table 3. Top-10 FoS trend from 2007-2011



Fields of Study	Level-1 FoS trend				
	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
Database	859	701	863	425	495
Computer Network	884	1368	1156	850	590
World Wide Web	854	751	1080	849	555
Algorithm	791	378	519	566	554

3.5 Feature Selection

To assess the significance of each feature (FoS trend and Early Citation Count) and understand their contributions, we employ Multiple Linear Regression to analyze each feature individually. Initially, we assess the predictive capacity of the model using only the Early Citation Count feature. Subsequently, we assess the FoS trend feature independently, and finally, we incorporate sets of features including FoS trend and Early Citation Count for prediction. This approach enables us to distinguish the specific impact of each feature on the prediction task.

In the FoS trend feature, we transform each FoS into its corresponding trend value and then sort them to identify the top-3 FoS. Consequently, the top-3 FoS trends (T1-T2-T3) are utilized in the experiments. Subsequently, both FoS trend (T1-T2-T3) and Early Citation Count are selected as real-valued features for citation count prediction, resulting in FoS represented as numeric data alongside citation counts, as depicted in Table 4-5.

Table 4. Sampled paper of level1 FoS trend

Paper _ Id	**Publication_ Year**	**Fields of Study(FoS)**	**Level0 FoS**	**Level1 FoS**	**Level1 FoS replaced by trend value (T1-T2-T3-T4)**	**Sort by top-value (T1-T2-T3)**
P1	2007	Wireless WAN, Heterogeneous network, Computer Science, Operating system, Wireless network, Internet security, Database, Security service, Key distribution in wireless sensor networks, Authorization, Internet privacy, Wi-Fi array, Cracking of wireless networks, Empirical research, Network Access Control, World Wide Web, Transport layer, Computer network.	Computer Science	Operating system, Database, World Wide Web, Computer network.	885-859-854-884	885-884-859.

Table 5. Sampled paper citation count

Year	**Paper Title**	**Total Citations**	**2007**	**2008**	**2009**	**2010**	**2011**
2007	User Security Behavior on Wireless Networks: An Empirical Study.	5	1	0	0	2	0

We have devised four feature schemes with the aim of identifying the most effective feature group for citation prediction:

Scheme 1 incorporates the top-3 FoS and Early Citation Count of 3-years (T1-T2-T3-2007-2008-2009) to predict citation counts up to 8 years.

Scheme 2 consists solely of the Early Citation Count of 3-years (2007-2008-2009) for predicting citation counts up to 8 years.

Scheme 3 integrates the top-3 FoS and Early Citation Count of 5-years (T1-T2-T3-2007-2008-2009-2010-2011) to forecast citation counts up to 6 years.

Scheme 4 utilizes only the Early Citation Count of 5-years (2007-2008-2009-2010-2011) to predict citation counts up to 6 years.

These schemes allow for the exploration of various combinations of features to determine their efficacy in predicting citation counts over different timeframes.

In our proposed model FTCCP, we train a process using FoS trend and the early citation history of an article, aiming to predict its future citation counts. The model predicts future citations (cc)ˆ_(x+1),(cc)ˆ_(x+2), ..., (cc)ˆ_n based on its initial citations $cc_0, cc_1,$, ... , cc_x and FoS trend. Initially, the model is trained using a dataset comprising papers with FoS trends and citation count histories. Subsequently, it can be employed to predict citation counts for future years, leveraging the previous citation counts and FoS trend of a specific article.

We employ published articles with FoS trends and Early Citation Counts as both the training and testing datasets. A portion of the articles is assigned to the training set, while the rest are designated for the testing set to evaluate the accuracy of the trained method. FoS trend and Early Citation Count are employed as features in this procedure. The performance of the FTCCP model is assessed by comparing its predictions with the actual citation counts from the testing dataset.

3.6 Regression Top of Form

We employ Multiple Linear Regression (MLR) to construct an FTCCP model designed to predict the citation count of an article using its FoS trend and Early Citation Count. In this model, the prediction of citation count is treated as a regression problem, aiming to estimate a non-negative integer value. This approach is consistent with machine learning methodologies, which approach the research problem through a regression perspective.

3.7 Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR), a statistical method, expands upon the simple linear regression model to accommodate datasets with multiple predictor variables [27]. It facilitates the exploration of relationships among two or more explanatory variables and a response variable by fitting a linear equation to the observed data. In our research, we apply a multiple linear regression model to determine a linear equation that effectively characterizes the relationships within our dataset.

3.8 The Coefficient of Determination

The coefficient of determination R2, also known as the Multiple Correlation Coefficient, is a frequently used measure in univariate Multiple Linear Regression [29]. It represents the proportion of variability in the dataset explained by the linear model. Mathematically, it can be defined as:

$$R^{2} = 1 - \frac{\Sigma(y_{i},- \hat{y}_{i})^{2}}{\Sigma(y_{i},- \hat{Y})^{2}} \quad (6)$$

where

$\hat{Y}$ = the mean value of the samples,
y_i = the actual value,

$\hat{y}_i$= the predicted value

$R^2 \in [0,1]$, we assume to attain a larger R^2that indicates improved performance

4. Results and Discussions

We performed several experiments to evaluate the efficiency of our proposed approach, FTCCP. Initially, following scheme 1, we applied the proposed model to forecast the citation count of 5863 papers from 2007, 6599 papers from 2008, 7159 papers from 2009, 7070 papers from 2010, and 6315 papers from 2011 within our dataset. Figure 2 depicts the actual citation counts contrasted with the forecasts generated by our proposed FTCCP model, employing the top-3 FoS trends and Early Citation Count history up to the third year after publication as input features. Additionally, Table 6 displays the R2 scores obtained by FTCCP.

As shown in the figures, there is no discernible or straightforward pattern in the citation counts across the sampled papers, indicating the complexity involved in predicting the citation count of a paper. Nonetheless, the FTCCP model closely approximates the actual citations for sampled papers, as demonstrated by Figure 2.

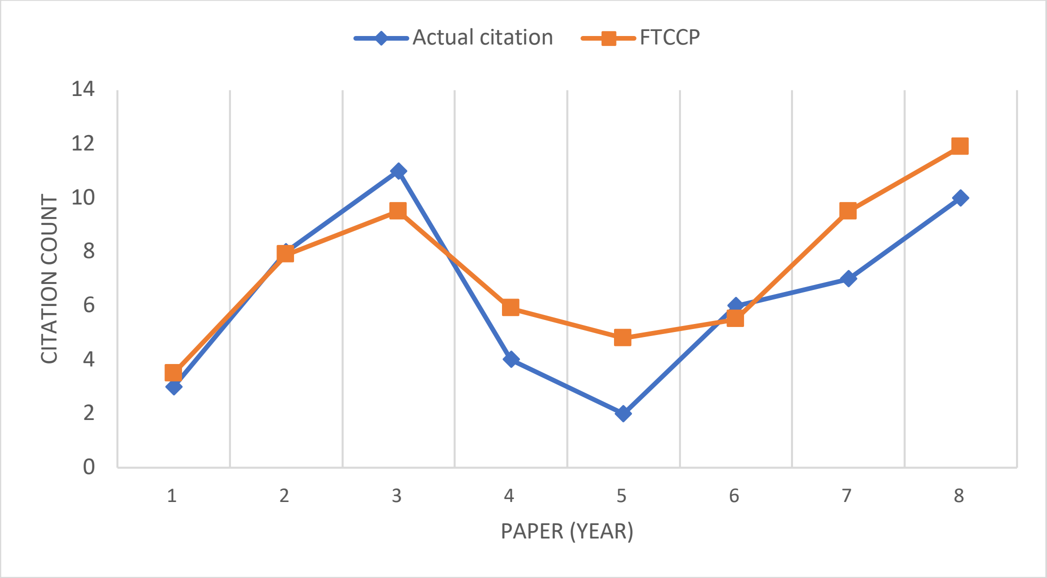


Figure 2. T1-T2-T3-2007-2008-2009 and predict future citations.

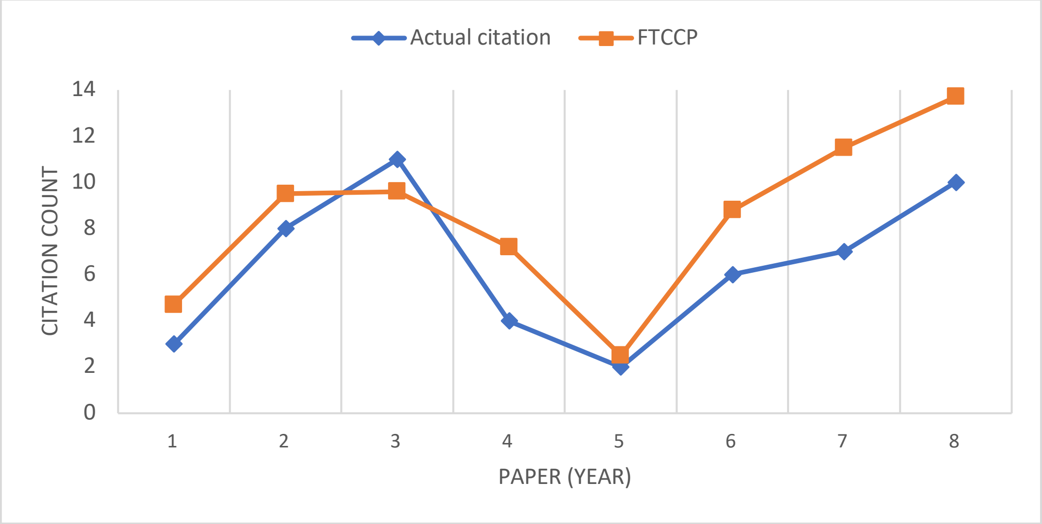


Figure 3. 2007-2008-2009 and predict future citations.

Table 6. Results of scheme 1

Scheme 1 (Features)	R^2
X T1-T2-T3-2007-2008-09, Y 2010	0.748973
X T1-T2-T3-2007-2008-09, Y 2011	0.884231
X T1-T2-T3-2007-2008-09, Y 2012	0.961529
X T1-T2-T3-2007-2008-09, Y 2013	0.955894
X T1-T2-T3-2007-2008-09, Y 2014	0.903412
X T1-T2-T3-2007-2008-09, Y 2015	0.885166
X T1-T2-T3-2007-2008-09, Y 2016	0.862214
X T1-T2-T3-2007-2008-09, Y 2017	0.825412

Table 7. Results of Scheme 2

Scheme 2 (Features)	R^2
X 2007-2008-09, Y 2010	0.703121
X 2007-2008-09, Y 2011	0.863190
X 2007-2008-09, Y 2012	0.920714
X 2007-2008-09, Y 2013	0.906989
X 2007-2008-09, Y 2014	0.898194
X 2007-2008-09, Y 2015	0.866087
X 2007-2008-09, Y 2016	0.844603
X 2007-2008-09, Y 2017	0.797234

In the second experiment corresponding to scheme 2, we restrict the input to Early Citation Count up to the third year after publication, aiming to predict the citation count until the 8th year post-publication. Figure 3 presents the evaluation outcomes of this experiment. As indicated in Table 7, the proposed method FTCCP yields lower R2 values in this scenario when solely using citation counts as input compared to the results obtained in scheme 1 experiments. Consequently, the proposed method demonstrates superior performance when incorporating both FoS trend and Early Citation Count for prediction.

For the third experiment under scheme 3, we utilize the input as top-3 FoS trend and Early Citation Count up to the fifth year after publication, repeating the procedure from the first experiment. In this setup, FoS trend and Early Citation Count history up to the fifth year after publication are employed to predict the citation count up to the 8th year after publication. Figure 4 displays the evaluation outcomes of this experiment. As shown in Table 8, the proposed method FTCCP continues to outperform according to the R2 criterion for the papers within the dataset.

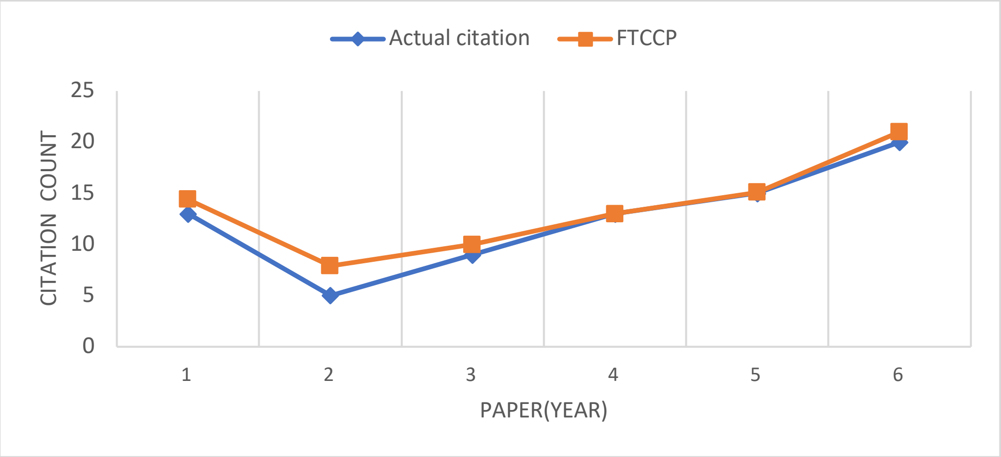


Figure 4. T1-T2-T3-2007-2008-2009-2010-2011 and predict future citations.

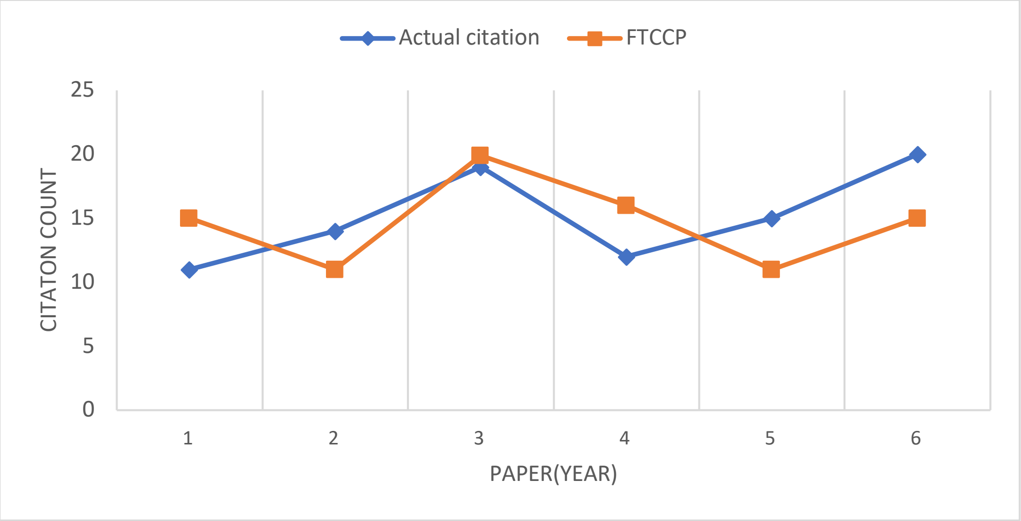


Figure 5. 2007-2008-2009-2010-2011 and predict future citations.

Table 8. Results of scheme 3.

Scheme 3 (Features)	R²
X T1-T2-T3-2007-2008-2009-2010-2011, Y 2012	0.752313
X T1-T2-T3-2007-2008-2009-2010-2011, Y 2013	0.901120
X T1-T2-T3-2007-2008-2009-2010-2011, Y 2014	0.979803
X T1-T2-T3-2007-2008-2009-2010-2011, Y 2015	0.972214
X T1-T2-T3-2007-2008-2009-2010-2011, Y 2016	0.921014
X T1-T2-T3-2007-2008-2009-2010-2011, Y 2017	0.904024

Table 9. Results of scheme 4

Scheme 4 (Features)	R^2
X 2007-2008-2009-2010-2011, Y 2012	0.721101
X 2007-2008-2009-2010-2011, Y 2013	0.880120
X 2007-2008-2009-2010-2011, Y 2014	0.952102
X 2007-2008-2009-2010-2011, Y 2015	0.933465
X 2007-2008-2009-2010-2011, Y 2016	0.904241
X 2007-2008-2009-2010-2011, Y 2017	0.872417

In the fourth experiment corresponding to scheme 4, we confine the input to Early Citation Count up to the fifth year after publication, and we replicate the procedure from the second experiment. In this setup, Early Citation Count up to the fifth year after publication is employed to predict the citation count up to the 8th year. Figure 5 illustrates the results of this experiment. Table 9 presents the outcomes of this experiment based on R2 scores. It is noteworthy that the prediction results obtained by FTCCP are more accurate when utilizing both FoS trend and Early Citation Count, compared to using citation count alone, as evidenced by the R2 scores.

Table 10. Results of proposed scheme and comparison with literature.

Features		R^2
Early Citation Count	Δt = 3	0.920714
	Δt = 5	0.962102
Early Citation Count + FoS Trend	Δt = 3-T1-T2-T3	0.961529
	Δt = 5-T1-T2-T3	0.979803
Paper + Author + Network [9]	Δt = 3 (using MLR)	0.5536
	Δt = 5	0.5505
	Δt = 3 (using SVM[1])	0.6287
	Δt = 5	0.6254

[1] SVM: Support Vector Machine

Features		R^2
EarlyCitationCount	Δt = 3	0.920714
	Δt = 5	0.962102
EarlyCitationCount + FoS Trend	Δt = 3-T1-T2-T3	0.961529
	Δt = 5-T1-T2-T3	0.979803
Paper+Author+Network [9]	Δt = 3 (using MLR)	0.5536
Paper+Author+Network [9]	Δt = 5 (using MLR)	0.5505
Paper+Author+Network [9]	Δt = 3 (using SVM)	0.6287
XinPing et al.,2018 [9]	Δt = 5 (using SVM)	0.6254

The results indicate that the inclusion of the FoS trend feature significantly enhances the accuracy of predicting future citation counts compared to relying solely on citation count. Additionally, the findings highlight that utilizing FoS trend and Early Citation Count up to the fifth year after publication leads to higher R2 scores compared to using data up to the third year. Incorporating both FoS trend and Early Citation Count as input features provides the model with more information, resulting in improved performance in predicting future citation counts.

In comparison with a previous study [9], which achieved R2 scores of 0.5536 for 3-years and 0.5505 for 5-years using paper, author, and network features for prediction via MLR, our approach yielded higher R2 scores of 0.961529 for 3-years and 0.979803 for 5-years when utilizing FoS trend and early citation count. This demonstrates the effectiveness of our method in leveraging FoS trend and Early Citation Count for improved prediction accuracy.

5. Conclusion and Recommendations

This study presents a pioneering approach named FoS Trend-based Citation Count Prediction (FTCCP), designed to forecast the citation count of scientific articles using their Field of Study (FoS) trend and Early Citation Count. Fundamentally, our proposed model utilizes the FoS trend and Early Citation Count of a paper within the initial years following its publication (specifically 1-3 years and 1-5 years) to anticipate its citation count over a longer-term duration. Importantly, our analysis does not include other features such as author details, venue, or journal.

While some literature explores more extensive sets of features for citation prediction methods, we intentionally restrict the inputs to simply the FoS trend and Early Citation Count. This decision is aimed at keeping the research problem accessible and applicable across diverse research domains. Our study focuses on two main categories of features for FTCCP: FoS trend and early citation count, employing Multiple Linear Regression for analysis.

Experimental evaluations conducted on the Microsoft Academic Graph dataset yield promising results, indicating that FTCCP achieves accuracy when utilizing FoS trend and early citation history compared to relying solely on citation history from publication, as evidenced by R2 scores.

5.1 Future Work

Moreover, the proposed features demonstrate superior effectiveness compared to conventional ones. For future investigations, we plan to explore the application of the proposed method for predicting the future h-index of authors, as well as enhancing the overall performance of our prediction methodology.

References

1. Xia F, Wang W, Bekele TM, Liu H. Big scholarly data: A survey. IEEE Transactions on Big Data. 2017 Jan 9;3(1):18-35.
2. Fiala D, Tutoky G. PageRank-based prediction of award-winning researchers and the impact of citations. Journal of Informetrics. 2017 Nov 1;11(4):1044-68.
3. Zhang J, Ning Z, Bai X, Kong X, Zhou J, Xia F. Exploring time factors in measuring the scientific impact of scholars. Scientometrics. 2017 Sep;112:1301-21.
4. Xia F, Su X, Wang W, Zhang C, Ning Z, Lee I. Bibliographic analysis of nature based on Twitter and Facebook altmetrics data. PloS one. 2016 Dec 1;11(12):e0165997.
5. Li J, Ye FY. Distinguishing sleeping beauties in science. Scientometrics. 2016 Aug;108:821-8.
6. Adankon MM, Cheriet M. Genetic algorithm–based training for semi-supervised SVM. Neural Computing and Applications. 2010 Nov;19:1197-206.
7. Cao X, Chen Y, Liu KR. A data analytic approach to quantifying scientific impact. Journal of Informetrics. 2016 May 1;10(2):471-84.
8. Sandulescu V, Chiru M. Predicting the future relevance of research institutions-The winning solution of the KDD Cup 2016. arXiv preprint arXiv:1609.02728. 2016 Sep 9.
9. Zhu XP, Ban Z. Citation count prediction based on academic network features. In2018 IEEE 32nd international conference on advanced information networking and applications (AINA) 2018 May 16 (pp. 534-541). IEEE.
10. Sinha A, Shen Z, Song Y, Ma H, Eide D, Hsu BJ, Wang K. An overview of microsoft academic service (mas) and applications. InProceedings of the 24th international conference on world wide web 2015 May 18 (pp. 243-246).
11. Garfield E. Impact factors, and why they won't go away. Nature. 2001 May 31;411(6837):522.
12. Fersht A. The most influential journals: Impact Factor and Eigenfactor. Proceedings of the National Academy of Sciences. 2009 Apr 28;106(17):6883-4.
13. Dimitrov JD, Kaveri SV, Bayry J. Metrics: journal's impact factor skewed by a single paper. Nature. 2010 Jul 8;466(7303):179.
14. Hirsch JE. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. Scientometrics. 2010 Dec;85(3):741-54.
15. Stallings J, Vance E, Yang J, Vannier MW, Liang J, Pang L, Dai L, Ye I, Wang G. Determining scientific impact using a collaboration index. Proceedings of the National Academy of Sciences. 2013 Jun 11;110(24):9680-5.
16. Bethard S, Jurafsky D. Who should I cite: learning literature search models from citation behavior. InProceedings of the 19th ACM international conference on Information and knowledge management 2010 Oct 26 (pp. 609-618).
17. Brody T, Harnad S, Carr L. Earlier web usage statistics as predictors of later citation impact. Journal of the American Society for Information Science and Technology. 2006 Jun;57(8):1060-72.
18. Lokker C, McKibbon KA, McKinlay RJ, Wilczynski NL, Haynes RB. Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. Bmj. 2008 Mar 20;336(7645):655-7.
19. Fu LD, Aliferis C. Models for predicting and explaining citation count of biomedical articles. InAMIA Annual symposium proceedings 2008 (Vol. 2008, p. 222). American Medical Informatics Association.
20. Ibáñez A, Larrañaga P, Bielza C. Predicting citation count of Bioinformatics papers within four years of publication. Bioinformatics. 2009 Dec 15;25(24):3303-9.
21. Callaham M, Wears RL, Weber E. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. Jama. 2002 Jun 5;287(21):2847-50.
22. Kulkarni AV, Busse JW, Shams I. Characteristics associated with citation rate of the medical literature. PloS one. 2007 May 2;2(5):e403.
23. Pobiedina N, Ichise R. Citation count prediction as a link prediction problem. Applied Intelligence. 2016 Mar;44:252-68.
24. Yan R, Huang C, Tang J, Zhang Y, Li X. To better stand on the shoulder of giants. InProceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries 2012 Jun 10 (pp. 51-60).
25. Livne A, Adar E, Teevan J, Dumais S. Predicting citation counts using text and graph mining. InProc. the iConference 2013 workshop on computational scientometrics: Theory and applications 2013 Feb 12 (pp. 16-31).
26. Abrishami A, Aliakbary S. Predicting citation counts based on deep neural network learning techniques. Journal of Informetrics. 2019 May 1;13(2):485-99.
27. Tranmer M, Elliot M. Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR). 2008;5(5):1-5.
28. Li CT, Lin YJ, Yan R, Yeh MY. Trend-based citation count prediction for research articles. InAdvances in Knowledge Discovery and Data Mining: 19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22, 2015, Proceedings, Part I 19 2015 (pp. 659-671).
29. Steel RG, Torrie JH. Principles and procedures of statistics. Principles and procedures of statistics. 1960.