Scientometrics (2022) 127:2773–2790
https://doi.org/10.1007/s11192-022-04344-3

Citation burst prediction in a bibliometric network

Tehmina Amjad[1] · Nafeesa Shahid[1] · Ali Daud[2,3] · Asma Khatoon[1]

Received: 25 September 2021 / Accepted: 8 March 2022 / Published online: 25 March 2022

Abstract

In the field of computer science, both journal and conference publications are considered valuable. The popularity of an author is mostly determined by the paper's high citations in a short time. Features that can help to attract higher visibility are not yet thoroughly investigated in the literature. This study aims to investigate the impact of the several features on received citations, for articles published in both journals or conferences. The correlation analysis and multiple linear regression models are applied to explore the strength of all related features. The study helps in finding the impact of the individual features on the number of citations both for journals and conferences, and to predict future citations. *AMiner* citation dataset has been used for experimental analysis. The findings of the study show that in the case of journal publications, "author first-year citations" and "author total citation" are the most important features. While, in the case of conference publications, "author total citation" is more effective as compared to other features. In the case of journal publications, the multiple linear regression model shows the coefficient of determination (R^2) is 0.975 and accuracy 0.846. For the conference publications, the R^2 value and accuracy are 0.877 and 0.846, respectively.

Keywords Journal · Conference · Citation burst · Citations analysis · Features · Correlation · Multiple linear regression

Introduction

There exist large repositories of scientific information on the web such as digital libraries and archives, which help us in developing and exploring the bibliometric networks. The major issue is to determine the quality of the scientific literature. However, it has been observed that the quality of the contents or scientific information is directly extracted from the standing of the publication venue. In academic culture, both journal articles and conference papers are valuable. Especially, the Computer Science (CS)

✉ Ali Daud
alimsdb@gmail.com

[1] Department of Computer Science and Software Engineering, IIU, Islamabad, Pakistan

[2] School of Information Engineering, Zhejiang Ocean University, Zhoushan, 316022, China

[3] Department of Computer Science and Artificial Intelligence, University of Jeddah, Jeddah, Saudi Arabia

community perceives conference papers to be as essential as journal articles for sharing research findings (Bar-Ilan, 2010; Franceschet, 2010). Features, such as Thomson's Impact Factors, H-Index, and Y-Factors are designed for the assessment of journals, and the features like longevity, conference size, prestige, and current popularity are typically used for the assessment of conferences. Existing literature has explored the various factors affecting the citations of journal or conference publications by using advanced data mining techniques (Amjad et al., 2020, 2021; Daud et al., 2017, 2019; Lee & Brusilovsky, 2019; Li et al., 2015; Onodera & Yoshikane, 2015; Zhu & Ban, 2018). Although very few relevant studies exist that considered both journal and conference publications (Kim, 2019; Vrettas & Sanderson, 2015). The number of citations received by an academic entity (authors, paper, journal, conference) is a primary feature for impact evaluation of that academic entity. Therefore, the impact of researcher and research articles is usually measured by the citation count.

Mainly this study compares the journal and conference citation rates. It also focuses on early citations (Early citations represent the citations from years 1–5 of publications (Zhu & Ban, 2018)) and observes the general trends of journal and conference papers. Especially, regarding various features that affect the number of citations and also studies the extent to which these features influence the rate of citations.

Different features have different impacts, Garfield's impact factor is best known to measure the citations (Garfield, 1972). Attracting a high number of citations in a short time is a strong indicator of an author becoming an expert or influential author quickly. Therefore, to analyze the importance of conference and journal papers in CS, researchers studied both journal and conference publications and sometimes considered them individually as well (Franceschet, 2010; Kim, 2019; Lee & Brusilovsky, 2019). These studies provided an overview of the authors and authorship features that are extracted from large-scale publications data like DBLP, Google Scholars, and CiteSeer. Most of the existing work about citations ignores a thorough investigation of features that may help in attracting higher visibility.

In this work, the following contributions are made.

- Using the basic features provided by the dataset, extraction of fourteen features for four different dimensions including authors, venues, papers, and sociability. These features are Author Reputation, Author Productivity, Author h-index, Author Impact Factor, Author Total Reference Papers, Affiliation, Co-author Counts, Co-author Citations, Co-author Publications, Venue Citations, Venue Publications, Venue Impact Factor, Age of Paper, and Title Length.
- Analyzing the relationship among received citations and the extracted academic features using Pearson correlation coefficient. This will help in identifying which features can be more helpful in attaining more citations.
- Analyzing whether conferences were able to gain more citations or the Journals.

The rest of the manuscript is organized in such a way that Section 2 provides details of surveyed literature, Section 3 presents the problem definition, Section 4 explains the proposed methodology, Section 5 covers a discussion on results, and Section 6 concludes the study along with some future directions.

Related work

The difference between journal and conference publications has been deliberated considering the authorship level (Kim, 2019). Kim analyzed the data of 517,763 scholars and found that 64.30% of scholars have published their first work in conferences and 25.44% in the journal during the last 57 years. It was observed that a conference is a more prevalent resource of research communication in CS. Chen and Konstan found the difference between journal and conference publications at the article level (Chen & Konstan, 2010). They found that papers in conferences that have a low rate of acceptance (almost 30%) have more impact and attract the same number of citations or more cations in ACM as compared to journal articles, where the impact is assessed by the total citations that are received. On the other hand, papers that are published in low-quality journals received more citations as compared to papers published in low-quality conferences, and the length of the papers influences the citation rate (Vrettas & Sanderson, 2015). Some researchers raised the concept that bibliometric databases such as Web of Science, Scopus, and ACM Digital library do not cover all conference publications that may underestimate the conference impact (Li et al., 2015). CS conference publications values are more than other academic fields but overall journal citation rates are higher than conferences (Vrettas & Sanderson, 2015).

Some of the studies examined the extension of conference publications into journal articles. A study analyzed that in CS almost 25% to 33% of conference publications were later published in journals (Bar-Ilan, 2010). The extension of conference publications into the journal was mostly discussed at the article level. For example, Wainer and Valle (2013), examined the 200 articles of CS and found that 62% in the conference and 55% in journals, authors seemed in extended work and 26% of conference articles extended in journals (Wainer & Valle, 2013). Conversely, Onodera and Yoshikane analyzed the 57 years of CS publications and examined data at the authorship level (Onodera & Yoshikane, 2015). They argued that the title of words and co-authors are not much overlapped in journal and conference publications.

There exist very few relevant studies that considered both journal and conference publications and find out the features which influence more to get citations (Vrettas & Sanderson, 2015). Most of the studies considered journals and very few studies considered conference features. Lee et al. used various conference-related features to understand the impact of the factors on early citations and predict the future citation count of conference papers (Lee & Brusilovsky, 2019). The analysis shows that bookmarks collected within the duration of 4 to 12 months after conference served, for early attention of online readership, it is reliable evidence and also reliable in predicting future citations. Various factors like the type of paper in conference and count of paper presented in the conference also predict better citations in both Scopus and Google scholars. Another study identifies the author-related properties and their predictive power for future citations of the conference paper (Lee, 2020). They studied 21 factors related to the first author and all other authors by considering 28 conferences and found that all author-related factors are high predictors as compared to the first author-related factors, and feature of all authors and first author, degree centrality have highest predictive power for future citations. Another study investigates three types of factors like conference series, individual conference paper, and individual conference (Lee, 2019). They concluded that name, content similarity of papers, international collaboration degree, and age of the conference series both effectively predict the future citations. By using the information of early citations Stern (Stern, 2014) find out high-ranked publications. Yan et al. (Yan et al., 2011, 2012) combined author features, venues features, and content features and used the regression models to predict the function. Bornmann et al. identified that author-related information of the paper could help in predicting the citations. The study argued that it is possible to improve the measurement of citation impact in a short time window by considering factors: number of authors, number of cited references, journal impact factor, and total pages of a paper (Bornmann et al., 2014). This study revealed that citation impact measurement can be improved in the first year after publication. The result was analyzed by using a regression model which showed that by adding the journal impact factor, the number of cited references and total pages increased the value of prediction. Ibáñez et al. focused on different prediction model for Bioinformatic journals and these models used to predict citation count (4 years) of a paper (Ibáñez et al., 2009). Tokens in abstract, section of the journal and 2-weeks post-publication are used as a predictive feature. They used logistic regression and naïve Bayes to define the learning process in nine journal sections with the four years of time horizon. They proved that the appearance of words in the abstract has an impact on the number of citations a paper received. To predict citations factors: weight ratio and abstract ratio both are significant (Sohrabi & Iraj, 2017).

Hence, the previous literature does not provide the aggregate information of how the difference between journals and conferences at the author level and which feature has more impact on citation in conferences are still not identified. Moreover, the analysis is missing in the existing literature, whether the impact of general features is the same in both journal and conference. Therefore, the study aims to complement previous literature by comparing the difference of journal versus conference publication at an author level and to find the features which are important to determine whether a paper gets higher citations to understand them better publication trend in CS.

Problem definition

In a scientific collaboration network, measuring the impact of citations received by an academic publication is an important activity. However, the impact of these features can be different for the conference publications and the journal publications. Thus the impact of these features needs to be examined for conferences as well as journals. Previous studies mostly focused on features that are either specific to a conference or a journal, however, very few researchers have considered the features that are related to both publication venues. It is beneficial for the researchers if they are aware that which features can assist them in gaining citations in a short period. Therefore, this study determines which features can help the authors to gain more citations in a short time span.

Methodology

The proposed methodology is divided into multiple phases. In the first phase, the dataset is extracted and preprocessed. In the second phase, we identified and separated the conference papers and the journal features and extracted the features for them. The extracted features include Author Reputation, Author Productivity, Author h-index, Total References, Affiliation, Co-author Count, Co-author Citations, Co-author Publications, Venue Citations, Venue Publications, Venue Impact Factor, Age of Paper, and Title Length. In the next phase, for each year (2006, 2007, 2008, 2009, and 2010) the correlation between the received citations and each feature is calculated using the Pearson correlation coefficient. Using the data from 2006 to 2009 as training data, multiple

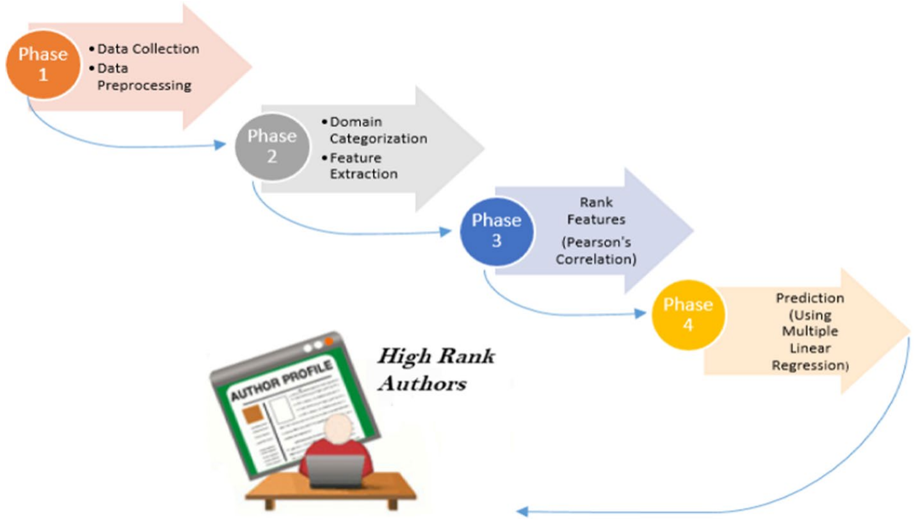


Fig. 1 Abstract representation of the proposed methodology

linear regression model is used is applied to predict total citations in the year 2010 using each feature to identify which feature predicts the future citations with more accuracy. Finally, highly ranked authors with respect to citations are identified as conferences and journals. Figure 1 represents the proposed methodology.

Figure 1 depicts the flow of the proposed methodology, however; the pseudocode of the proposed methodology is provided below.

```
1. Selection of the year 2006 to 2010
2. Extract the unique authors in the dataset
3. Find out the number of citations received by each publication of the dataset
4. Calculate citations received in each year
5. Categorized data into Journal and Conference publications
6. Calculate basic features in journal and conference publication individuallyFor each author calculate
 a. Author Reputation
 b. Author Productivity
 c. Author h-index
 d. Total Reference
 e. Affiliations
 f. Co-author Count
 g. Co-author Citations
 h. Co-author Publications
 i. Venue Citations
 j. Venue Publications
 k. Venue Impact Factor
 l. Age of Paper
 m. Title Length
7. Calculate Correlation between citations during the year of 2006 to 2010
8. Features are ranked according to the value of their correlation coefficient.
9. Rank the authors who get more citations in a short time span.
10. Apply Multiple Regression Model to find which feature predicts the future citations with better
   accuracy
 a. 2006 to 2009 data was used for training the model
 b. predict the citations in 2010
```

Identifying early burst

A burst represents the time in which many events are occurred (Zhang & Shasha, 2006). To predict the citation incrementing speed within different periods, the burst calculation is an important step. Calculating the citation increment speed in different periods helps us to track the progress of a researcher. This study considers 5 year times as early citations, considering Δ5 as a burst time, and find out the impact of different features on citations at a particular time. For experimentation, we divided the authors into two categories.

Case 1: Select all authors having a minimum of 1 citation.

Case 2: Select all authors having a minimum of 3 citations and citations are greater than the previous year by a gain factor of 75%.

Dataset description and preprocessing

The dataset used is extracted from AMiner which is an educational research and mining platform and the dataset is author name disambiguated (Tang et al., 2008). This dataset covers 2,092,356 papers from computer science, 8,024,869 citations, and 1,712,433 researchers from the year 1936 to 2014. Dataset consists of journals articles, conference papers, books, and reviews. Each record has a unique id, author name, publication year, publication venue, abstract, affiliation, and references. Many previous studies analyzed *AMiner* data for collaboration mapping, content similarities mapping and data management (Amjad et al., 2015, 2017; Kim, 2019; Li et al., 2015). for experimentation, the publications data ranging from 2006 to 2010 is extracted and a total of 617,740 articles are obtained. During preprocessing, the records which are published other than journals or conferences categories like books, reviews are excluded. The records that have null author names, missing the paper title, and publication year are also omitted. The authors with no citations are also removed from the dataset. The dataset is categorized into two sets, the journal publications, and the conference publications. Table 1 represents the dataset statistics.

Measurement of future scientific impact

For the measurement of the researcher's future scientific impact and citations in journals and conferences, features related to four different factors i.e., authors, venue, paper, and sociability are calculated. Table 2 represents all features that are considered. All features are calculated individually for journal and conference categories. We have also mentioned the existing studies that have used these features in their methodology. Knowledge about these factors helps to estimate early citations (5 years) that a published paper will likely receive.

The first group of factors explains the scientist's performance. Based on prior studies in information science (Lee, 2019; Zhu & Ban, 2018), the citations related factors were calculated within 5 years. In this study, factors are calculated from the year 2006 to 2010. The first group of features consists of six features: author reputation (AR), author productivity (AP), author h-index, author impact (AI), and affiliations. The author's reputation/ total number of citations represent the number of paper which is used as a reference in other work/paper. Danell (2011) proved that the author's reputation or past performance of the researcher has an impact on the citation count (Danell, 2011). Second productivity determines the total number of papers published by an author in the journal or conference.

Table 1 Description of dataset

	Journal	Conference
No. of papers	306,947	310,793
No. of Authors initially	179,489	185,398
No. of Authors after preprocessing	73,826	57,666

Table 2 Features considered in this study

Author-related features	Author reputation (AR) (Zhu & Ban, 2018)
	Author productivity (AP) (Onodera & Yoshikane, 2015)
	Author h-index (Hurley et al., 2013; Yan et al., 2011)
	Author Impact Factor (AIF) (Pan & Fortunato, 2014)
	Author total reference papers (ARF) (Vintzileos & Ananth, 2010)
	Affiliation (Yu et al., 2014)
Sociability features	Co-author counts (CC) (Zhu & Ban, 2018)
	Co-author citations (CAC) (Pan & Fortunato, 2014)
	Co-author Publications (CAP) (Amjad et al., 2018; Daud et al., 2015)
Venue-related features	Venue citations (VC) (Bethard & Jurafsky, 2010)
	Venue publications (VP) (Singh et al., 2017)
	Venue impact factor (VIF) (Bai et al., 2019)
Paper-related features	Age of paper (AP) (Zhu & Ban, 2018)
	Title length (TL) (Lyu & Wolfram, 2018; Rostami et al., 2014)

Several publications by researchers are considered as an important factor for future citations (Onodera & Yoshikane, 2015). The third feature, the AIF is used to measure the impact of a researcher's work. AIF can find the trends of a researcher impact exhibit during their careers. Some measuring metrics for a particular performance area are unable to track the impact variation in careers, AIF fills that gap. It can be measured by the total number of citations a researcher have and normalizing it with the recent publication of a researcher (Pan & Fortunato, 2014). The fourth feature, H-index is used to measure the impact of work and productivity of the published work of a researcher. It is based on the researcher's papers and citations of the papers (Yan et al., 2011). Author h-index was a significant feature to predict citations (Hurley et al., 2013). The fifth feature, reference papers are the source a researcher used in their work or list of the resource's researcher has cited. For publication, it's the most important part because editors use reviewers that are included in the reference list of an author (Vintzileos & Ananth, 2010). The number of references has positive correlation with citations (Yu et al., 2014). The reason is that some of the authors in the reference list have already done work on the same topic. The number of references also distinguishes whether the paper is a survey paper or a regular paper (Li et al., 2015). The institution's reputation indirectly reflects the scientific research ability of an author within the institution (Zhu & Ban, 2018). According to Amara et al.'s (2015) studies, institutional affiliation has a significant impact on citations (Amara et al., 2015). If the reputation of an institute is high then the author's research ability is also high.

The second group of features relates to the research capacity of the collaborators of the target scientists' to determine whether collaborating with successful and experienced collaborators matters for their potential development as a researcher in the initial part of their careers. Therefore, it is important in the disciplines of information science and computer science to investigate the impact of collaborators on future research. Co-author/collaborators show the sociability of an author and also reflect that these particular authors work on similar topics (Zhu & Ban, 2018). The number of co-authors is a highly influential factor that has a positive correlation with citations (Hurley et al., 2013). Hence, a researcher has more co-authors he/she gets more citations because of widely connected authors. Co-author citations are a social feature that reflects the paper's popularity and quality. If a new researcher has few citations and then collaborates with the senior researcher, there are many chances for a new researcher that he/she get more citations in collaborations (Amjad et al., 2018; Daud et al., 2015).

Top venues submit high-quality papers. This submission shows the reputation of a venue. Venue publications refer to how many papers are published in a particular journal or conference. Some venues have higher productivity, and some have low. The number of papers published in journals influenced the citation count (Li et al., 2015; Yu et al., 2014). Venue citations count refers to how often a venue has been cited (Bethard & Jurafsky, 2010). Journal citation or conference citation is the number of citations received in a particular journal or conference. This statistic is common to analyze the impact of the venue. It is an important factor that is positively correlated with citations (Singh et al., 2017). It is a qualitative index to measure the impact of a venue but it cannot access the individual article quality (Bai et al., 2019). Papers that are published in high-impact venues received more citations. Venue impact was identified as an important factor for an article. It shows the average value of citations of published papers in the particular venue.

Besides author and sociability features, additional intuitive features affecting the publication's success are its paper-related features. for citation analysis, the time elapsed from its publication date is very important and it needs to be considered. If the time of publication of a paper is longer then the paper has more readers and it may receive more citations (Lyu & Wolfram, 2018; Zhu & Ban, 2018). The title of a paper usually describes the objective of the study and develops the interest for further reading (Vintzileos & Ananth, 2010). The number of words in a title reflects the title length and it helps to predict the future citation count (Rostami et al., 2014). Paper use increases if the title of the paper is informative and also increases the number of downloads.

Pearson correlation

It is a statistical measure that is used to measure the relationship strength between two variables (Singh et al., 2017). The range of values is − 1 to + 1 and 0. The − 1 value shows a perfect negative relationship between variables, + 1 shows a perfect positive relationship and 0 value shows no relationship between the variables. The Pearson correlation coefficient is calculated as[2]:

$$r = \frac{\sum \left(x - \bar{x}\right)\left(y - \bar{y}\right)}{\sqrt{\sum \left(x - \bar{x}\right)^{2} \sum \left(y - \bar{y}\right)^{2}}} \tag{1}$$

where,$\bar{x}$ and $\bar{y}$ are. Sample means of two arrays.

This study calculated the correlation between yearly citations and factors that are considered. After calculating the correlation, we calculate the mean value of all factor's correlation values.

The prediction model

To predict future citations a statistical technique, multiple line regression (MLR), is applied. For multiple independent variables, it is the most used form of linear regression. MLR is also used to describe the relationship between one dependent variable i.e., Yearly citation, and two or more independent variables i.e., author features. It is used to predict future values and trends. The relationship between a variable Y depending on p variables x_1, x_2… x_p in the following technique:

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \beta_p x_p + \varepsilon \tag{2}$$

where Y is response variable (also dependent variable, output, explained variable), x_1, x_2,… x_p: regress (also predictor, input, explanatory variables, independent variable), ε Random variable representing the error.

In this study, first, the dataset is divided into two categories: journal and conference, and then features are extracted from the dataset. After this, multiple linear regression is applied to predict future citations by using different features. The data from the year 2006 till 2009 is used for training the model and total citations in the year 2010 are predicted with the help of MLR. We predicted the received citations with all features one by one, then by using all features in a group (authors, paper, venue, and sociability), and finally, we predicted the citations by using all 14 features.

Performance evaluation

To validate the efficacy of our proposed method coefficient of determination (R^2) is calculated. We follow Yan et.al (Yan et al., 2011) to use the coefficient of determination (R^2) to be an evaluation metric.

$$\boldsymbol{R}^2 = \frac{\sum_{\boldsymbol{d}\in\boldsymbol{D_T}} \left[\boldsymbol{C}_{\boldsymbol{T_{CCP}}}(\boldsymbol{d}) - \boldsymbol{C_T}\left(\boldsymbol{D_T}\right)\right]^2}{\sum_{\boldsymbol{d}\in\boldsymbol{D_T}} \left[\boldsymbol{C_T}(\boldsymbol{d}) - \boldsymbol{C_T}\left(\boldsymbol{D_T}\right)\right]^2} \tag{3}$$

where $C_{T_{ccp}(d)}$ is the predicted citations for an article d in the test set $\mathrm{D_T}$ and $\mathrm{C_T}$ ($\mathrm{D_T}$) is the mean of the observed citations count for an article d in $\mathrm{D_T}$. The value of R^2 ranges from 0 to 1**.** A higher value indicates better performance.

Results and discussion

Feature analysis

Feature analysis of 14 features is performed and features are ranked in journal and conference categories based on correlation values. The correlation between yearly citations and features is calculated. In the next step, the mean value of all features is calculated. After that, features are ranked in descending order according to their mean correlation values for journals and conferences. Table 3 shows the ranked list of all features for cases 1 and 2 in both categories.

Figure 2 represents the correlation values for case 1 and case 2. Some features are highly correlated, some show medium or no correlation. From all features, author-related features have high correlation values. Author reputation has the highest value for both cases and both categories. Thus, when author-related features values are increased then citations also increase, or if citations of any author increase then values of author-related features also increase. The venue citation feature is showing more high correlation for the case 2 conferences as compared to case 1 conferences. This shows that although the number of citations received by journals was much higher than the number of citations received by the conferences (Table 4), still it is surprising to see the venue citations feature shows much higher correlation for the conferences as compared to the journals. It is also observed that the Age of paper feature is significantly highly correlated for the conferences as compared to journals. While this feature shows no impact on prediction of future citations (Table 8). All other features have positive but not much high relation with citations. CAP, VP, and VIF have approximately 0 relation with citations which means by increasing the values of these features, the value of citations will not change. We also observed that the number of words in the title has a negative correlation which represents that if the title of any paper is too long then it may have a negative correlation impact on the citations. It is also shown that in both cases the trend of conference papers correlation is higher than the trend of journal paper correlational values for same features.

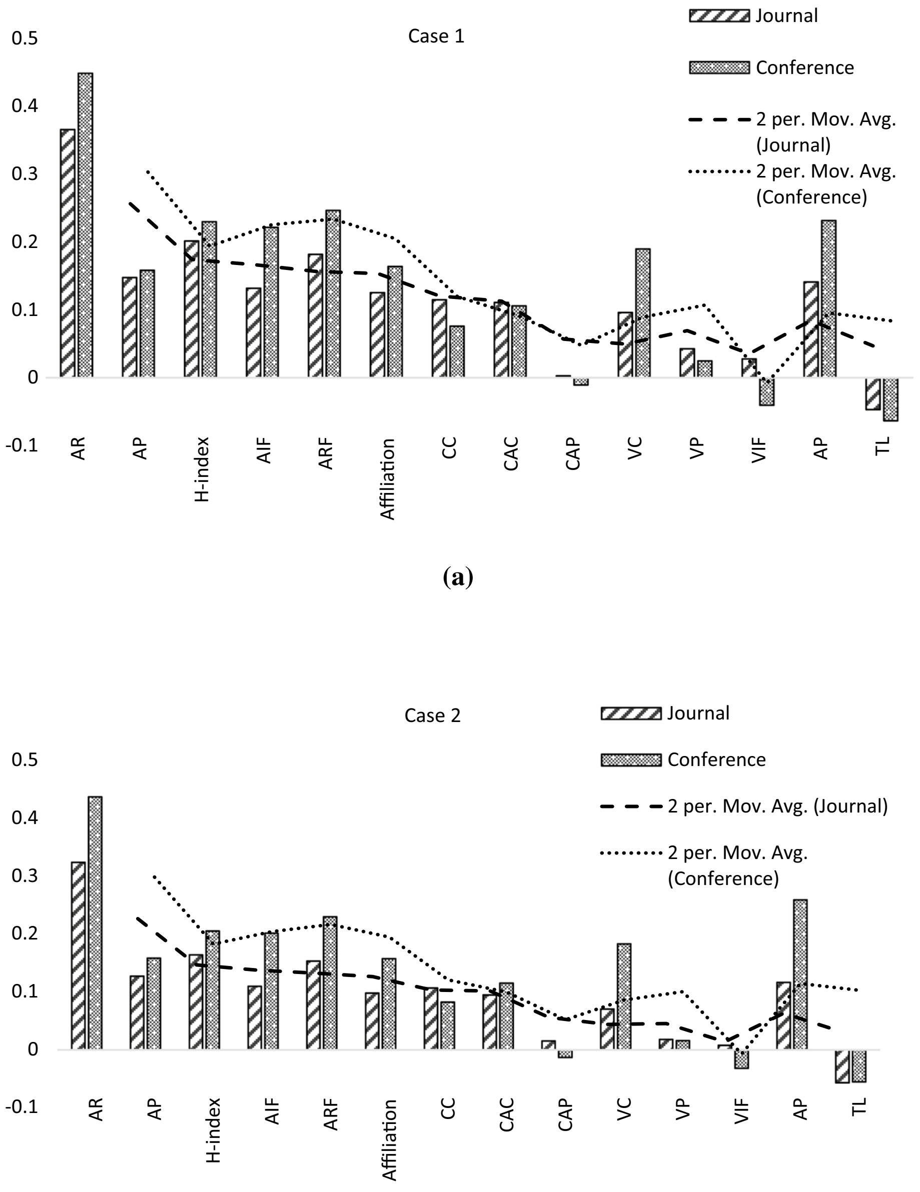
Authors ranking

In this section, we rank the top 15 authors in journals and conferences that received more citations in a short period (within 5 years). Table 4 represents the author rank list for case 1 and case 2.

We can also conclude that in both cases, authors that published their work in journals received more citations, and authors that published their work in the conference received fewer citations.

Table 3 Ranking factors according to correlation values

Case 1		Case 2	
Journal	Conference	Journal	Conference
Author total citations	Author total citations	Author total citations	Author total citations
Author impact	Author H-Index	Author impact	Author H-Index
Author H-Index	Author impact	Author H-Index	Age of paper
Author total Ref. Papers	Author Total Ref. Papers	Author total publications	Author Impact
Co-Author counts	Co-author citations	Age of Paper	Author Total Ref. Papers
Author total publications	Author total publications	Author Total Ref. Papers	Venue impact
Venue impact	Co-author publications	Co-author Citations	Co-author citations
Co-author citations	Venue impact	Co-author publications	Author total publications
Co-author publications	Co-author counts	Co-author counts	Co-author count
Venue citations	Venue citations	Venue impact	Co-author Publications
Venue publications	Affiliations	Venue citations	Venue citations
Affiliation	Venue publications	Venue publications	Affiliations
Age of paper	Age of paper	Affiliations	Venue Publications
No. of words in title	No. of words in title	No. of Words in Title	No. of Words in Title



(a)

(b)

Fig. 2 **a** Correlation between features for case 1. **b** Correlation between features for case 2

Predicting citations and impact of features on citations in journal and conference

In this section, we predict the future citations of authors and find out the effect of different features that are used to predict these citations. To predict the future citations, we used multiple linear regression model to analyze and to find the various feature's impact. The data from 2006 to 2009 is used to train the model and the citations received in 2010 are predicted. We added all attributes one by one and then observed the effect of all features on citations individually. To determine the effect of these features on the citations, R^2 is calculated.

The value of R^2 ranges from 0 to 1, the features that generate higher R^2 in prediction have a high impact on received citations and features that generate low R^2 values have less impact on received citations. For the purpose of analysis, we divided the features with respect to their impact on received citations into four categories including high impact, medium impact, low impact, and negative impact. The features that have $R^2 > 0.5$ are termed as high impact, features that have a R^2 between 0.2 and 0.5 are medium impact, features with R^2 between 0.2 and 0.1 are termed as low impact features and finally, the features with $R^2 = 0$ are no impact features.

Table 5 represents the list of features that have a high impact on citations in the journal and conference category. In the journal category author total citations and first-year citations have a higher impact on citations, which shows that to get early citations these two factors are most important. By using these two features get R^2 0.969 for case 1 and 0.925 for case 2. Still, in the conference category, only author total citations have a higher impact and by using this, we get R^2 0.822 for case 1 and 0.784 for case 2.

In case 1, author h-index and author impact had a medium effect on citations in the journal category, and by using these two features, we get R^2 0.449. In the conference category first-year citations, author total reference papers, author h-index, and author impact have a medium effect, and we get R^2 68.10% by using all these features. In case 2 and both categories author impact had a medium effect on citations, and we get R^2 0.461 and 0.477 (Table 6).

Table 7 represents the list of features that have a low impact on citations. In journal features: author total publications, author complete reference papers. Co-author count, co-author citations, co-author publications, venue citations, venue impact, and the number of words in the title have low impact. We get R^2 0.0580 and in the conference get R^2 0.2840 by using the features: author publications, co-author count, co-author citations, co-author publications, venue citations, venue impact, venue publications, and the number of words in the title. We get R^2 0.034 for journals and 0.101 for the conference.

Some factors have 0 impacts on citations and these features venue publications and age of paper in a journal and conference age of paper has 0 impacts. These factors are shown in Table 8.

Afterwards, we used all the features at once for the prediction of citations. In case 1, by considering all features we get R^2, 0.976 in the journal, and 0.843 in the conference category. The value of R^2 shows that by using various factors we accurately predict citations 97.60% in the journal and 84.30% in the conference. In case 2, by considering all features we get R^2, 0.976 in the journal, and 0.846 in the conference category. The value of R^2 shows that by using various factors we accurately predict citations 94.10% in the journal and 84.60% in the conference.

Discussion

This study identifies the features that are important for getting high citations in journals and conferences. Talking about a high number of citations, it was observed from the results of the study that journal papers received more citations as compared to conference papers.

Table 4 Authors rank list

Rank No.	Case 1				Case 2			
	Journal		Conference		Journal		Conference	
	Citations	Author name	Citations	Author name	Citations	Author name	Citations	Author name
1	776	P. Gupta	335	Eugene Agichtein	776	P. Gupta	335	Eugene Agichtein
2	742	V. Tarokh	310	Jure Leskovec	742	V. Tarokh	310	Jure Leskovec
3	630	Christopher M. Bishop	286	Mihir Bellare	630	Christopher M. Bishop	286	Mihir Bellare
4	621	S. M. Alamouti	269	Sachin Katti	621	S. M. Alamouti	268	Ashwin Machanavajjhala
5	609	G. Bianchi	268	Ashwin Machanavajjhala	609	G. Bianchi	262	Herbert Bay
6	564	R. Ahlswede	262	Herbert Bay	564	R. Ahlswede	257	Cynthia Dwork
7	501	S. Haykin	257	Cynthia Dwork	501	S. Haykin	228	Dan Boneh
8	497	D. L. Donoho	244	Svetlana Lazebnik	497	D. L. Donoho	218	Sören Auer
9	423	G. Caire	228	Dan Boneh	423	G. Caire	199	Nicholas Nethercote
10	407	T. J. Richardson	218	Sören Auer	407	Dacheng Tao	198	Martin Gebser

Table 5 High impact features

Case 1		Case 2	
Journal	Conference	Journal	Conference
Author total citations	Author total Citations	Author total citations	Author total citations
Author first year citations	–	Author first year Citations	–

Regarding the relationship of features with citations, we analyzed that some features are highly correlated, some are medium or no correlation. From all features, author-related features have a high correlation with received citations. Other features have a positive correlation, but it is not significantly high. Features like CAP, VP, and VIF have approximately 0 correlation. We also observed that the number of words in the title has a negative correlation. We observed that in both categories (journal and conference) the impact of features was not similar as all features have a different impact on citations. In the journal category author total citations and first-year citations have a higher impact on citations. The result is similar to the study (Silva et al., Aug. 2020) but in the conference category, only author total citations have a higher impact on citations. The author h-index and author impact had a medium effect on citations in the journal category. In the conference category first-year

Table 6 Medium impact features

Case 1		Case 2	
Journal	Conference	Journal	Conference
Author h-index	Author first year citations	Author impact	Author impact
Author impact	Author h-index	–	–
–	Author total reference paper	–	–
–	Author Impact	–	–

Table 7 Low impact features

Case 1		Case 2	
Journal	Conference	Journal	Conference
Author total publications	Author total publications	Author total publications	First year Citations
Author total reference papers	Co-author's publications	Author total reference papers	Author Total Publications
Co-author's publications	Co-author's citations	Co-author's Publications	Venue Citations
Co-author's citations	Co-author count	Venue Impact	Venue Impact
Co-author count	Venue citations	Author Total Publications	First year Citations
Venue citations	Venue impact	–	–
Venue impact	No. of words in title	–	–
No. of words in title	Age of paper	–	–

citations, author total reference papers, author h-index, and author impact have a medium effect. Some factors have 0 impacts on citations and these features venue publications and age of paper in a journal and in conference age of paper has 0 impacts. Overall author-related features have more correlation values as well as more impact on citations.

The baseline method used Deep Neural Networks, Support Vector Machines, and Multiple Linear Regression. In this study we have only applied the multiple linear regression because its performance is the best as per results of X. P. Zhu et al. (Zhu & Ban, 2018). The proposed method incorporated the features for journal and conference publications and also studied the impact of each factor on received citations while the baseline method only considered the journal features. Table 9 shows that previous work achieved 88.87% accurate prediction for journal publications using MLR. The proposed model performs better when compared to the baseline by achieving 97.06% accurate prediction in journal publication and 84.3% accurate prediction in conference publications.

Conclusion and future work

This study identifies the features that are more helpful for the researchers to gain more citations in a short time. from the findings of the study, it was observed that computer science researchers publish more articles in conferences as compared to the journals but journals articles were receiving more citations as compared to the conference publications. An interesting finding was observed that relationship between various features and the author's annual citations in both categories (journal and conference) are different. The citations based features like author h-index, author impact show a positive correlation with future citations. Factors, like title length, have a negative correlation. For prediction, we applied multiple linear regression models in both categories and studied the impact of the individual features on citations. When considering authors that have at least 1 citation, in the journal category impact of 'total author citations' and 'first-year author citations' is high with R ^{2} 0.969 and by using all considered features the R^2 is 0.975. In the conference category, only 'author total citations' has more impact, and we get R^2 0.772 and when all factors are combined we get R^2 0.843. it is observed that to get citations in a short time most essential features in the journal category is 'first-year citations'. Other factors like author

Table 8 Zero impact factors

Case 1		Case 2	
Journal	Conference	Journal	Conference
Affiliations	Affiliations	Author h-index	Author publications
Venue Publications	Venue Publications	Affiliations	Affiliations
Age of Paper	–	Co-author citations	Author h-index
–	–	Co-author publications	Co-author citations
–	–	Co-author count	Co-author publications
–	–	Venue citations	Co-author Count
–	–	Venue publications	Venue publications
–	–	Age of paper	Age of paper
–	–	Title length	Title length

Table 9 comparison with baseline method

	Proposed Method	Baseline Method
Dataset	*AMiner*	*AMiner*
Features	Author Reputation	Paper Citation quality
	Author Productivity	Paper Published time
	Author h-index:	Paper Novelty
	Author Impact Factor	Paper Popularity
	Author Total Reference Papers	Paper Diversity Author citations
	Affiliation:	Author h-index
	Co-author counts	Author ability
	Co-author citations	Author versatility
	Co-author Publications	Venue citations
	Venue Citations	Venue productivity
	Venue Publications	Betweenness centrality
	Venue Impact Factor	Page rank
	Age of Paper	Community
	Title Length	Authority
R^2 (MLR)	0.976 for journal	Experimentation for journals only
	0.843 for conference	For SVM, MRL and DNN 88.87% accurate prediction

h-index, author reference paper, venue impact, author impact are also positively corelated. Some factors like venue publications and title length are not important for getting citations fast. We obtained R^2 0.941 in the journal and 0.846 in the conference category when we predicted the citations in the early burst. Papers that are published in journals achieved more citations as compared to conference publications. Overall author-related factors have more impact.

This study can be improved by using more features like author-related features, content, expertise, and reinforcement. Also expanding the time and finding out the impact of the features in different stages like middle and late, and comparing more categories like books, notes, proceeding papers. The proposed approach can be applied to other entities such as papers/articles, collaborators, and venues. For example, paper authors, paper co-authors, paper venue and title, and so on could be considered in the case of a paper as entity. The technique could be further enhanced after considering various factors related to collaborators like considering the impact of the collaborates and their bibliometric features and impact fator of publications venues and their topic of publications.

References

Amara, N., Landry, R., & Halilem, N. (2015). What can university administrators do to increase the publication and citation scores of their faculty members? *Scientometrics, 103*(2), 489–530.

Amjad, T., Daud, A., Che, D., & Akram, A. (2015). MuICE: Mutual influence and citation exclusivity author rank. *Information Processing & Management, 52*(3), Art. no. 3.

Amjad, T., et al. (2017). Standing on the shoulders of giants. *Journal of Informetrics, 11*(1), Art. no. 1.

Amjad, T., Daud, A., Khan, S., Abbasi, R. A., & Imran, F. (2018). Prediction of Rising Stars from Pakistani Research Communities. In *2018 14th International Conference on Emerging Technologies (ICET)*, 2018, pp. 1–6.

Amjad, T., Sabir, M., Shamim, A., Amjad, M., & Daud, A. (2021). Investigating the citation advantage of author-pays charges model in computer science research: a case study of Elsevier and Springer. *Library Hi Tech*.

Amjad, T., Rehmat, Y., Daud, A., & Abbasi, R. A. (2020). Scientific impact of an author and role of self-citations. *Scientometrics, 122*(2), 915–932.

Bai, X., Zhang, F., & Lee, I. (2019). Predicting the citations of scholarly paper. *Journal of Informetrics, 13*(1), 407–418.

Bar-Ilan, J. (2010). Web of Science with the Conference Proceedings Citation Indexes: The case of computer science. *Scientometrics, 83*(3), 809–824.

Bethard, S., & Jurafsky, D. (2010). Who should I cite: learning literature search models from citation behavior. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp. 609–618.

Bornmann, L., Leydesdorff, L., & Wang, J. (2014). How to improve the prediction based on citation impact percentiles for years shortly after the publication date? *Journal of Informetrics, 8*(1), 175–180.

Chen, J., & Konstan, J. A. (2010). Conference paper selectivity and impact. *Communications of the ACM, 53*(6), 79–83.

Danell, R. (2011). Can the quality of scientific work be predicted using information on the author's track record? *Journal of the American Society for Information Science and Technology, 62*(1), 50–60.

Daud, A., Ahmad, M., Malik, M. S. I., & Che, D. (2015). Using machine learning techniques for rising star prediction in co-author network. *Scientometrics, 102*(2), Art. no. 2.

Daud, A., et al. (2017). Who will cite you back? Reciprocal link prediction in citation networks. *Library Hi Tech, 35*(4), Art. no. 4.

Daud, A., Amjad, T., Siddiqui, M. A., Aljohani, N. R., Abbasi, R. A., & Aslam, M. A. (2019). Correlational analysis of topic specificity and citations count of publication venues. *Library Hi Tech.*

Franceschet, M. (2010). The role of conference publications in CS. *Communications of the ACM, 53*(12), 129–132.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178*(4060), 471–479.

Hurley, L. A., Ogier, A. L., & Torvik, V. I. (2013). Deconstructing the collaborative impact: Article and author characteristics that influence citation count. *Proceedings of the American Society for Information Science and Technology, 50*(1), 1–10.

Ibáñez, A., Larrañaga, P., & Bielza, C. (2009). Predicting citation count of Bioinformatics papers within four years of publication. *Bioinformatics, 25*(24), 3303–3309.

Kim, J. (2019). Author-based analysis of conference versus journal publication in computer science. *Journal of the Association for Information Science and Technology, 70*(1), 71–82.

Lee, D. (2020). Author-related factors predicting citation counts of conference papers: focusing on computer and information science. *The Electronic Library*.

Lee, D. H. (2019). Predictive power of conference-related factors on citation rates of conference papers. *Scientometrics, 118*(1), 281–304.

Lee, D. H., & Brusilovsky, P. (2019). The first impression of conference papers: Does it matter in predicting future citations? *Journal of the Association for Information Science and Technology, 70*(1), 83–95.

Li, C.-T., Lin, Y.-J., Yan, R., & Yeh, M.-Y. (2015). Trend-based citation count prediction for research articles. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 659–671.

Lyu, P., & Wolfram, D. (2018). Do longer articles gather more citations? Article length and scholarly impact among top biomedical journals. *Proceedings of the Association for Information Science and Technology, 55*(1), 319–326.

Onodera, N., & Yoshikane, F. (2015). Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology, 66*(4), 739–764.

Pan, R. K., & Fortunato, S. (2014). Author Impact Factor: Tracking the dynamics of individual scientific impact. *Scientific Reports, 4*, 4880.

Rostami, F., Mohammadpoorasl, A., & Hajizadeh, M. (2014). The effect of characteristics of title on citation rates of articles. *Scientometrics, 98*(3), 2007–2010.

Silva, F. N., et al. (2020). Recency predicts bursts in the evolution of author citations. *Quantitative Science Studies, 1*(3), 1298–1308. https://doi.org/10.1162/qss_a_00070

Singh, M., Jaiswal, A., Shree, P., Pal, A., Mukherjee, A., & Goyal, P. (2017). Understanding the impact of early citers on long-term scientific impact. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 1–10.

Sohrabi, B., & Iraj, H. (2017). The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts. *Scientometrics, 110*(1), 243–251.

Stern, D. I. (2014). High-ranked social science journal articles can be identified from early citation information. *PLoS ONE, 9*(11), e112520.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: extraction and mining of academic social networks, pp. 990–998. Retrieved August 14, 2014 from http://dl.acm.org/citation.cfm?id=1402008.

Vintzileos, A. M., & Ananth, C. V. (2010). How to write and publish an original research article. *American Journal of Obstetrics and Gynecology, 202*(4), 344-e1.

Vrettas, G., & Sanderson, M. (2015). Conferences versus journals in computer science. *Journal of the Association for Information Science and Technology, 66*(12), 2674–2684.

Wainer, J., & Valle, E. (2013). What happens to computer science research after it is published? Tracking CS research lines. *Journal of the American Society for Information Science and Technology, 64*(6), 1104–1111.

Yan, R., Tang, J., Liu, X., Shan, D., & Li, X. (2011). Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1247–1252. Retrieved May 09, 2017, from http://dl.acm.org/citation.cfm?id=2063757

Yan, R., Huang, C., Tang, J., Zhang, Y., & Li, X. (2012). To better stand on the shoulder of giants. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pp. 51–60. Retrieved 24 November, 2016, from http://dl.acm.org/citation.cfm?id=2232831.

Yu, T., Yu, G., Li, P.-Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics, 101*(2), 1233–1252.

Zhang, X., & Shasha, D. (2006). Better burst detection. In *22nd International Conference on Data Engineering (ICDE'06)*, pp. 146–146.

Zhu, X. P., & Ban, Z. (2018). Citation count prediction based on academic network features. In *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*, pp. 534–541.