Scientometrics (2019) 119:1575–1595
https://doi.org/10.1007/s11192-019-03052-9

Which can better predict the future success of articles? Bibliometric indices or alternative metrics

Mingyang Wang[1] · Zhenyu Wang[1] · Guangsheng Chen[1]

Received: 26 September 2018 / Published online: 6 April 2019
© Akadémiai Kiadó, Budapest, Hungary 2019

Abstract

In this paper, we made a survey on the prediction capability of bibliometric indices and alternative metrics on the future success of articles by establishing a machine learning framework. Twenty-three bibliometric and alternative indices were collected to establish the feature space for the predication task. In order to eliminate the possible redundancy in feature space, three feature selection techniques of Relief-F, principal component analysis and entropy weighted method were used to rank the features according to their contribution to the original data set. Combining the fractal dimension of the data set, the intrinsic features which can better represent the original feature space were extracted. Three classifiers of Naïve Bayes, KNN and random forest were performed to detect the classification performance of these features. Experimental results show that both bibliometric indices and alternative metrics are beneficial to articles' growth. Early citation features, early Web usage statistics, as well as the reputation of the first author are the most valuable indicators in making an article more influential in the future.

Keywords Highly-cited papers · Bibliometric index · Alternative metrics · Machine learning

Introduction

In the past few decades, the growing popularity of bibliometric indices has led to a thorough study of the citation process. The type of document (Annalingam et al. 2014; Bornmann 2013; Bornmann and Williams 2013; Ingwersen and Larsen 2014), its subject (Antoniou et al. 2015; Bornmann et al. 2012; Dorta-González et al. 2014; van Eck et al. 2013), its publishing venue (Didegah and Thelwall 2013; Falagas et al. 2013; Garner et al. 2014; Jiang et al. 2013; Van Der Pol et al. 2015), authors (Biscaro and Giupponi 2014;

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11192-019-03052-9) contains supplementary material, which is available to authorized users.

✉ Guangsheng Chen
kjc_chen@163.com

[1] College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, People's Republic of China

Collet et al. 2014; Farshad et al. 2013; Pagel and Hudetz 2011) and other characteristics (Tahamtan et al. 2016) all influence its citation impact in, statistically speaking, predictable ways. However, these bibliometric indices from academic publications can only help to measure the impact of research within research itself. Alternative metrics (also known as altmetrics) are regarded as an attractive possibility because they not only enable fast, but might also provide broad impact measurement (Priem and Hemminger 2010; Priem et al. 2012). Altmetrics "focuses on the creation, evaluation and use of scholarly metrics derived from the social web" (Haustein et al. 2014). The question to what extent altmetrics actually permit a broad impact measurement of research is currently an object of scientometric research.

Bibliometricians see some value in altmetrics, especially download metrics (Haustein et al. 2013) and there is already evidence that a range of altmetrics associate with traditional citations counts, with Mendeley (Haustein et al. 2013; Li et al. 2012; Zahedi et al. 2013) and Twitter (Eysenbach 2011; Thelwall et al. 2013) seeming to be the most promising sources. And the weak positive correlations between social media mentions and future citations (Peoples et al. 2016; Ringelhan et al. 2015) suggest that online activity may anticipate or drive the traditional measure of scholarly 'impact'. Online activity also promotes engagement with academic research, scholarly or otherwise, increasing article views and PDF downloads of *PLoS ONE* articles (de Winter 2015; Wang et al. 2014). Thus, altmetrics, and the online activity they represent, have the potential to complement, pre-empt and boost future citation rates, and are increasingly used by institutions and funders to measure the attention garnered by the research they support (Ravenscroft et al. 2017).

The use of altmetrics in information retrieval and research evaluation brings the question: Whether the data Altmetric collects is a leading indicator of later success? Do social media mentions predict or correlate with subsequent citation rates for a given article? The absence of such a detection, however, push this paper to contribute to this discussion. We hope that the combination of traditional bibliometric indices and alternative metrics will provide more complete article profiles as it captures more dimensions of scientific practice.

A machine learning framework was established to detect the usefulness of this combination in predicting the later success of articles. Because the highly-cited status for papers is widely accepted as the one indicator of success or higher citation impact, this work is aiming to detect whether this combination of traditional bibliometric indices and alternative metrics are beneficial to predict the future highly-cited papers. And which one could be better for this prediction? Figure 1 shows the sketch of the prediction framework.

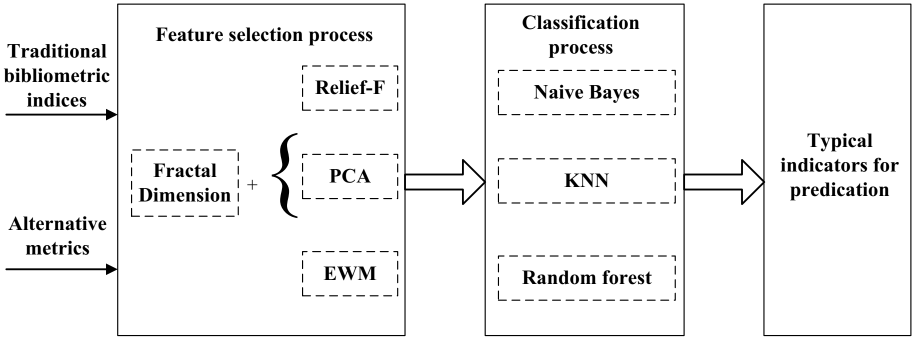

Fig. 1 The sketch for the framework on predicting future highly-cited papers

Both the traditional bibliometric indices and the alternative metrics were combined to establish the feature space for the prediction task. In order to overcome the "dimensionality curse" (Korn et al. 2001; Pagel et al. 2000) which may probably existed in the feature space, a feature selection process was performed to choose the subset of features which maintain the essential characteristics of the data set. Considering that each feature selection technique may be biased to some features due to their initial mechanisms, three different selection techniques of Relief-F, principal component analysis (PCA) and entropy weighted method (EWM) were introduced to rank the features according to the weights calculated by each selection technique. In order to discover how many features are significant to characterize the original dataset, the fractal dimension of the dataset was calculated. Combining the fractal dimension of the data set and the feature-ranking results under each selection technique, the kernel feature subsets were detected for each selection technique, respectively. Finally, three classification methods of Naïve Bayes, KNN and random forest were taken to detect the robust of the three feature subsets. And the typical indicators for predicting future highly-cited papers were hoped to be identified if reasonable classification performances could be got in the classification process.

Related work

Given the important role of citations in measuring the quality of research and researchers, it is reasoning to investigate why some papers achieve more citations than others. Various studies have been conducted to explore the factors influencing citations. Some have attempted to estimate and predict citations of future.

Tahamtan et al. (2016) made a comprehensive review of the factors predicting the frequency of citations. They detected 198 relevant papers and summarized that the three categories of factors–'paper' related factors, 'journal' related factors, and 'author' related factors– are related to the number of citations. Fourteen 'paper' related factors were widely discussed to investigate their influences on predicting paper's future citation counts, e.g. the quality of paper (Buela-Casal and Zych 2010; Patterson and Harris 2009; Stremersch et al. 2007), characteristics of fields/subfield of a discipline and study subject/topics (Glänzel and Schubert 2003; Glänzel et al. 2014; Dorta-González et al. 2014; Gonzalez-Alcaide et al. 2016; Wang et al. 2015a), the characteristics of references (Antoniou et al. 2015; Biscaro and Giupponi 2014; Chen 2012; Didegah and Thelwall 2013; Onodera and Yoshikane 2015; Yu and Yu 2014), the length of paper (Falagas et al. 2013; Stremersch et al. 2015; van Wesel et al. 2014), the early citation and speed of citation (Garner et al. 2014; Glänzel et al. 2012; Hilmer and Lusk 2009a, b), and the accessibility and visibility of papers (Ebrahim et al. 2014; Rees et al. 2012; Yue and Wilson 2004), et al. Four 'journal' related factors, including the journal impact factor (Haslam and Koval 2010; Jiang et al. 2013; Royle et al. 2013; Van Der Pol et al. 2015), language of journal (Borsuk et al. 2009; Leimu and Koricheva 2005; Lira et al. 2013), scope of journal (Bjarnason and Sigfusdottir 2002; Huang et al. 2012; Vanclay 2013), and the form of publication (Ingwersen et al. 2014; Ibáñez et al. 2013; Ke et al. 2014; Sangwal 2012), were investigated to verify their performances in predicting the number of citations. And ten 'author' related factors were also detected by researchers to test whether these factors are related to the frequency of citations, such as the factors of the number of authors (Amara et al. 2015; Glänzel and Thijs 2004; Puuska et al. 2014; Sin 2011; Vieira and Gomes 2010), author's reputation and previous citations (Bornmann et al. 2012; Frandsen and Nicolaisen 2013; Hurley et al. 2013), the international and national collaboration of authors (Chi and Glänzel 2018; Collet et al. 2014; Glänzel and Heeffer 2014; Nomaler et al. 2013; Onyancha and Maluleka 2011; Wang et al. 2015b), authors' country (Lee et al. 2010; Miettunen and Nieminen 2003; Padial et al. 2010; Willis et al. 2011), and author's productivity (Bosquet and Combes 2013; Stremersch et al. 2015), et al.

However, there are some other factors related to the future frequency of citations which are not classified under the above three categories of factors. For example, the factors represent the knowledge diffusion activities of articles in the scientific environment. Our previous studies showed that the knowledge diffusion activities, represented as the citation distribution of articles in the scientific environment in their early stage after publication, could be good predictors for the article's future citation frequencies (Wang et al. 2012a, b). Such a citation distribution of an article in the scientific environment reflects the scope of the knowledge diffusion of it, indicating the range of its visibility and its contribution on the scientific entities in some extent. And this visibility and contribution has laid an important foundation for the article's citation frequency in future. Thus, in the present study, the citation distribution properties of articles in scientific environment were incorporated in the feature space to detect their capacities on predicting the future success of articles. Furthermore, although large amount of bibliometric factors were detected and regarded as the valuable indicators for predicting articles' future citation frequencies, they only relate to the assessment of scientific impact of articles but lack of reflecting the influence of researches beyond academia.

As scholarly communication migrated to the Web, so did citations. Altmetrics, short for alternative metrics, has been considered as an interesting option to describe Web-based metrics for measuring the societal impact of research (Priem and Hemminger 2010; Piwowar and Priem 2013). Most comments on the benefits of altmetrics relate their potential for measuring the broader impact of research, that is, beyond science (Priem et al. 2012; Bornmann 2014). It is hoped that altmetrics can deliver more transparent descriptions of the interest, usage and reach of scholarly products (Fausto et al. 2012; Taylor 2013) and also more diverse and nuanced forms of impact analyses than traditional biblometrics metrics (Waltman and Costas 2014).

There is evidence that higher altmetric scores associate with higher citation counts for many different indicators, for instance, for articles that were bookmarked on Mendeley (Eysenbach 2011; Herrmannova et al. 2018; Thelwall and Wilson 2016; Thelwall 2018), mentioned in Wikipedia (Didegah et al. 2018; Kousha and Thelwall 2017; Marashi et al. 2013), and tweeted on Twitter (Eysenbach 2011; Thelwall et al. 2013; Ortega 2016; Shu et al. 2018). Thelwall et al. (2013) studied associations between journal citations and different altmetric indicators. They found that six altmetrics sources (tweets, Facebook wall posts, research highlights, blog mentions, mainstream media mentions and forum posts) had significant associations between higher altmetric scores and high numbers of citations, suggesting that multiple different types of altmetrics may be valid and useful. Chi and Glänzel (2017) analyzed the relation between the usage and citation impact and found that the citations and usage counts in Web of Science correlate significantly, especially in the social science. Syamili and Rekha (2017) detected the correlation between the altmetrics scores (saves, views, Mendeley and twitter) with citation and found that all the altmetric scores expect twitter have good correlation with traditional bibliometric citation. Peoples et al. (2016) estimated the relative effects of Twitter activity on Web of Science citation rates and detected that twitter activity was a more important predictor of citation rates than 5-year journal impact factor. They stated that altmetrics and traditional metrics are closely related, but not identical, and suggested that both altmetrics and traditional citation rates can be useful metrics of research impact.

These previous investigations on altmetrics have attempted to compare the various altmetrics metrics with the traditional citation counts. However, they almost always used contemporary rather than future citation counts, leaving it still unknown whether early altmetrics scores can predict later citations for articles. When predicting the future citation status of an article, whether the altmetrics is a useful complement to traditional bibliometric indicators, or is a simple alternative? Or, which one of the two kinds of indicators will provide a better predictive performance?

The present study is mainly aiming to answer these questions. By collecting the early web usage data, and the article's traditional bibliometric indicators, we explored the role of the two kinds of factors in predicting the future citation trend of articles and hoped to give a more clear understanding on their roles in the prediction.

Data set

Our analysis is based on a corpus of 617 scientific articles published in seven journals in Public Library of Science (PLOS) between January 1, 2010 and January 31, 2010. In each journal, the articles collected were ranked according to their total citation counts and then were divided into the highly-cited papers (HCPs), medium-cited papers (MCPs) and Low-cited papers (LCPs) by the following scheme:

(1) *HCPs* Those articles whose accumulated citation ratio reach the 20% of the total citation counts of all articles collected in this journal, were regarded as highly-cited.
(2) *LCPs* Those articles whose accumulated citation ratio located in the last 20% of the total citation counts of all articles collected in this journal, were regarded as low-cited.
(3) *MCPs* Those articles whose accumulated citation ratio located between the ones for HCPs and LCPs in this journal, were regarded as medium-cited.

Table 1 gives the distribution of HCPs, MCPs and LCPs in the seven journals published in PLOS. The purpose of dividing the articles into the three categories is to define the different growth status of articles, where the HCPs is taken for the successful ones. Basing on

Table 1 The distribution of HCPs, MCPs and LCPs in each journal in PLOS

ID.	Journal	Number of HCPs	Number of MCPs	Number of LCPs	Total number
1	PLOS ONE	14	189	253	456
2	PLOS Genetics	3	18	17	38
3	PLOS Pathogens	3	18	16	37
4	PLOS Computational Biology	1	12	18	31
5	PLOS Biology	2	7	11	20
6	PLOS Neglected Tropical Diseases	1	8	9	18
7	PLOS Medicine	1	5	11	17
Total number	–	25	257	335	617

this division, an examination on which indicators are better to predict the future success of articles could be performed.

Indices for predicting future HCPs

Bibliometric indices

In the previous researches, bibliometricians discussed the contribution of characters from authors, journals and articles on the HCPs prediction. Therefore, the indices associated with these aspects were collected to be the members of feature space. At the same time, our previous work showed that the citation distribution of articles in the scientific environment in their early stage after publication also benefits for their later success (Wang et al. 2012a, b). And according to Glänzel (2008), the use of a 3-year citation window is “a good compromise between the fast reception of life science and technology literature and that of the slowly ageing theoretical and mathematical subjects”. Thus, the articles’ citation distribution data in the first 3-year citation window were collected to compose the feature space. Table 2 shows the bibliometrics indices collected from the above consideration.

Indices of x_1–x_6 give the characters of authors, in which the *h*-index of authors is used to represent the reputation of them. x_7 is the Impact Factor of journals publishing this article. x_8 is the number of references. Indices of x_9–x_{11} show the basic properties of articles. Indices of x_{12}–x_{13} are the first-cited properties of articles. Initial citations that a paper receives

Table 2 The bibliometric indices

Indices	Detailed information
x_1	Number of authors
x_2	Whether there is international cooperation
x_3	The *h*-index of the first author before publishing the article
x_4	The total citation counts of the first author before publishing the article
x_5	The maximum *h*-index of the corresponding authors before publishing the article
x_6	The maximum total citation counts of the corresponding authors before publishing the article
x_7	The Impact Factor of the journal
x_8	Number of references
x_9	Length of the article
x_{10}	Type of the article
x_{11}	Language of the article
x_{12}	The first-cited-age of the article
x_{13}	The citations obtained in the first-cited-age
x_{14}	Number of citing countries in the first 3-years after publication
x_{15}	Number of types of citing articles in the first 3-years after publication
x_{16}	Number of citing institutions in the first 3-years after publication
x_{17}	Number of citing journals in the first 3-years after publication
x_{18}	Number of citing subjects in the first 3-years after publication
x_{19}	Number of citing languages in the first 3-years after publication
x_{20}	Total citation counts obtained in the first 3-years after publication

are the early feedback of the scientific community about that paper. Van Dalen and Henkens (2005) stated that these two indices could help to determine the rate and the strength of the first citation. The index of x_{12}, the first-cited-age of the article, is calculated by the interval between article's first-cited year and its publication year. For example, if one article P published in year 2000 got its first citations in year 2002, the first-cited-age of P is calculated as $2002 - 2000 + 1 = 3$. Indices x_{14}–x_{20} give the citation distribution characters of articles in the scientific environments in the first 3 years after publication, in other words, the size of the scope of their influences in the scientific environments in their early stage after publication.

All these indices could be obtained through *Web of Science*. By using the "create citation report" in *Web of Science*, the *h*-index of authors could be got. By using the "analyze the indexing results" and the following buttons of "publication year" and "*Web of Science* categories" etc., the indices associated with the first-cited properties and the citation distribution characters in the first 3-year citation window could be obtained.

Alternative metrics

In this paper, all the alternative metrics were collected from PLOS. PLOS inaugurated a program to provide Article-level Metrics (ALMs) on every article across all journals since 2009. ALMs capture the manifold ways in which research is disseminated and can help users determine the value of an article to them and to their scientific community. PLOS ALMs split the alternative metrics into five categories, including: viewed, saved, discussed, recommended, and cited. Table 3 shows the detailed information about these five categories of alternative metrics in PLOS.

More detailed explanation about these five categories of alternative metrics can be found in http://www.lagotto.io/plos/#relativeMetrics. The classification mirrors the whole process of user engagement, from the first look at a paper to its citation in (scholarly) literature and thus relate to the various dimensions of research impact (Neylon and Wu 2009). Because our efforts on discussing the alternative metrics is limited to detect whether there are some metrics besides citation activities potentially benefiting to articles' future citations, the category "cited" is not considered here.

Altmetrics focuses on the exchange and dissemination of article in the social web environment, which can timely reflect the value of article. However, because the Web environment has the characteristics of faster update and more emphasis on new information, Altmetrics has insufficient sustainability for one article. In order to determine a reasonable time for collecting these alternative metrics, we observed the changing characters of these metrics for each of the 617 articles from PLOS ALMs. For each article, its publication date

Table 3 The detailed information about the five categories of alternative metrics in PLOS

PLOS	Detailed information	Indices
Viewed	HTML/PDF (PLOS or PubMed Central), XML (PLOS)	x_{21}
Saved	CiteULike, Mendeley	x_{22}
Discussed	NatureBlogs, ScienceSeeker, ResearchBlogging, PLOS Comments, Wikipedia, Twitter, Facebook	x_{23}
Recommended	F1000Prime	–
Cited	CrossRef, PubMed Central, Web of Science, Scopus	–

is marked as T_0; the date when the metrics firstly chang is marked as T_1; and the date when the metrics will never change is marked as T_2. By recording the changing value of the four categories of alternative metrics for each article (In this paper, the changes in these indices were counted every month), two time intervals were calculated as shown in Table 4.

The time span between the publication date and the first changing date, labeled as ave.$(T_1 - T_0)$ in Table 4, is an expression on first changing speed for the metrics. The time span between the first changing date and the date never changing, labeled as ave.$(T_2 - T_1)$ in Table 4, is used to represent the duration of each metrics. Obviously, the metrics are different in the changing status. Articles would be quickly viewed and discussed than saved according to the shorter time span ave. $(T_1 - T_0)$ for metrics of “Viewed” and “Discussed”. Metrics “Viewed” has the longest time span ave.$(T_2 - T_1)$ than that of “Saved” and “Discussed”, which indicates that articles could be viewed for a relative longer time. It should be mentioned that there is no status for metrics “Recommended”. The reason is that almost all the papers collected are not fortunate enough to be recommended in F1000Prime. Thus any calculation on this metrics is lack of significance. Accordingly, only three alternative metrics of “Viewed”, “Saved” and “Discussed” are left to constitute the feature space, which is labeled as indices x_{21}-x_{23}.

And because this study is based on the monthly unit to record the changes in the value of these indices, it can not accurately locate the specific time point of changes in the value. Furthermore, the analysis on the sustainability of these indices is somewhat rough. It can’t guarantee that some indices would get new attentions and change into a new value as a result after a period of silence. However, this study only wants to determine a point of time when the values of these indices are no longer densely varying, so that we can use the data of these indices before that point to explore the forecasting task.

Based on the results on the duration analysis, all the three metrics tend to be stable in the first 2-years after the publication of articles. We take the date of February 1, 2012, which is just the first 2-year after articles’ publication, to collect the values of “Viewed”, “Saved”, and “Discussed” for each article.

Feature selection process

Feature selection defines the least number of features that could be used to best represent the inter relationships amongst those features (Naraei and Sadeghian 2017). In this paper, three different feature selection techniques of Relief-F, principal component analysis (PCA) and entropy weighted method (EWM) were used to calculate the significance of each feature on the initial data set. However, these techniques can only help to rank the features according to the importance of them to the original data set, while be insufficient in determining the numbers of key features which can best represent the data set. In this paper, the fractal dimension of the data set was introduced to discover how many attributes are significant to characterize the dataset.

Table 4 The changing status of the alternative metrics

Metrics	Viewed (days)	Saved (days)	Discussed (days)	Recommended (days)
Ave. $(T_1 - T_0)$	30	135	30	–
Ave. $(T_2 - T_1)$	698	435	186	–

Fractal dimension of data set

The initial idea for calculating the fractal dimension of one data set aims at detecting the correlations between attributes in a data set, spotting the attributes that can be obtained by some function of others, and defining the reduced data set with only the relevant attributes (Berchtold et al. 1998; Pagel et al. 2000). These considerations lead to the definition of the embedding and intrinsic dimensions, as well as the correlation fractal dimension (Traina et al. 2000).

Definition 1 The embedding dimension E of a data set is the dimension of its address space. In other words, it is the number of attributes in the data set.

Definition 2 The intrinsic dimension D of a data set is the dimension of the spatial object represented by the data set, regardless of the space where it is embedded.

Note that if a data set has all of its features independent of the others, its intrinsic dimension is the embedding dimension ($D=E$). However, whenever there is a correlation between two or more features, the intrinsic dimensionality of the data set is reduced accordingly. Through the intrinsic dimension of a data set it is possible to decide how many attributes are actually required to characterize it.

Definition 3 Correlation Fractal dimension: Given a dataset presenting self-similarity in the range of scales [r_1,r_2], its Correlation Fractal dimension D_2 for this range is measured as:

$$D_2 = \frac{\partial log \sum_i C_{r,i}^2}{\partial \log(r)} = \frac{\partial \log(S(r))}{\partial \log(r)}, r \in \left[r_1, r_2\right] \tag{1}$$

As shown in (Belussi and Faloutsos 1995), the correlation fractal dimension correspond to the intrinsic dimension of the dataset. Thus, in the present work we use D_2 as the intrinsic dimension D, just like Belussi and Faloutsos (1995) have done. The intrinsic dimensionality gives a lower bound of the number of attributes needed to keep the essential characteristics of the dataset. The detailed calculation process on correlation fractal dimension could be found at Belussi and Faloutsos (1995) and Traina et al. (2000).

Feature selection technique of Relief-F

The basic idea of Relief-F is to draw instances at random, compute their nearest neighbors, and adjust a feature weighting vector to give larger weight to features that discriminate the instances from neighbors of different classes (Kononenko 1994). Specially, for a random selected sample x_i, it finds the k nearest neighbors $\{h_j\}$in the same class C with x_i as well as the k nearest neighbors$\{m_j\}$ in the each of the other classes $\{S\}$ besides C, respectively. And then the relief-F tries to find a good estimate of the following probability to assign as the weight for feature A:

$$
\begin{aligned}
W(A) = W(A) - \sum_{j=1}^{k} \operatorname{diff}(A, x_i, h_j)/(\mathrm{mk}) \\
+ \sum_{S \neq C} \left[\frac{p(C)}{1 - p(S)} \sum_{j=1}^{k} \operatorname{diff}(A, x_i, m_j \right] /(\mathrm{mk})
\end{aligned}
\tag{2}
$$

$$
\operatorname{diff}(A, x_i, x_j) = \begin{cases} \frac{\left| x_i[A] - x_j[A] \right|}{\max(A) - \min(A)} & A \text{ is continuous} \\ 0 & A \text{ is discrete and } x_i[A] = x_j[A] \\ 1 & A \text{ is discrete and } x_i[A] \neq x_j[A] \end{cases}
$$

where $p(C)$ is the probability of class C, and $p(S)$ is the probability of the other class $\{S\}$ besides C.

By calculating the weight for every feature in the original data set by Eq. (2), we can get the features ranked by their weights.

Feature selection technique of PCA

The principal component analysis technique is a statistical analysis approach to map multiple characteristic parameters to a few comprehensive features. These PCA-based comprehensive features are not related to each other and can represent original features effectively (Abdi et al. 2013). Based on these comprehensive features, the absolute weight to the original features could be calculated.

For a given feature vector set $x = \{x_1, x_2, \dots, x_m\}$, $x_i \in R^n$ which consists of m features vectors (m samples), each with n-dimensional, the algorithm to extract sensitive features and calculate the weight for original features is taken as follows (Xu et al. 2008):

(1) Calculate the average value:

$$
\boldsymbol{\mu} = \frac{1}{m} \sum_{i=1}^{m} x_i \tag{3}
$$

(2) Compute the covariance matrix C of eigenvectors:

$$
\mathrm{C} = \frac{1}{m} \sum_{i=1}^{m} \left(x_i - \boldsymbol{\mu} \right) \left(x_i - \boldsymbol{\mu} \right)^{\boldsymbol{T}} \tag{4}
$$

(3) Compute the eigenvalues λ_i and eigenvectors $\boldsymbol{v_i}(i = 1, 2, \dots, n)$ of C:

$$
\mathrm{C}\boldsymbol{v_i} = \lambda_i \boldsymbol{v_i} \tag{5}
$$

(4) Arrange the eigenvalues in descending order $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, composite the first k eigenvalues $\Delta = \left(\lambda_1, \lambda_2, \dots, \lambda_k \right)$ and corresponding eigenvectors $\boldsymbol{W} = \left[\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w_k} \right]$. Thus the cumulative contribution rate is defined as:

$$
R_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{6}
$$

R_k indicates the percentage of the total variance by the first k principal components (PCs). The first principal component is oriented in the direction of the largest variance. The following components are furthermore oriented in the direction of the decreasingly ordered further variances. In this paper, the most k representative PCs were selected by $R_k \geq 0.8$.

(5) The linear combination can be expressed for the kth dimension of the projected feature vector as follows:

$$PC_k = \sum_j w_j^k x_j \tag{7}$$

The weighting reflects the contribution of the original features to the linear combination, and thus is related to the original variance of the data samples.

(6) Calculate the absolute weight for the original features. By using the similar process proposed by Xu et al. (2008), the absolute weight $w_j^{'}$ of features x_j to the original data set is calculated by the selected components.

$$w_j^{'} = \frac{\sum_{i=1}^{k} w_{j\times}^{i} \lambda_i}{\sum_{i=1}^{k} \lambda_i} \tag{8}$$

By this method, a large $w_j^{'}$ represents a large contribution of the feature x_j to the original data set. Following the same computational procedure, the weights of all features were computed. Taking the order of weights from large to small, we can get all the features ranked.

Feature selection technique of entropy weight method (EWM)

The entropy weight method (EWM) determines weights through quantifying the disorder extent of a particular system (Huang et al. 2015). Because the weighting factors are purely dependent on the value of indices rather than human subjective assessment, EWM was recognized as an objective method for weight calculation. For a given feature vector set $\mathbf{x} = \left\{x_1, x_2, \ldots, x_m\right\}$, $x_i \in R^n$ which consists of m samples, each with n-dimensional, the main steps for weighting the features with the entropy weight method are as follows:

Step 1 Normalization of the features. In order to ensure the uniformity of indices' units or value range, the normalization of all features is performed as:

$$\begin{gathered} y_{ij} = \frac{x_{ij} - \min_j \left(x_{ij}\right)}{\max_j \left(x_{ij}\right) - \min_j \left(x_{ij}\right)} \left(\max_{\text{new}} - \min_{\text{new}}\right) + \min_{\text{new}} \\ i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, n \end{gathered} \tag{9}$$

where $\left[\min_{\text{new}}, \max_{\text{new}}\right]$ is the new value range for all the features, which is usually set as $\left[\min_{\text{new}}, \max_{\text{new}}\right] = [0, 1]$.

Step 2 Calculation of weighting coefficients. The information entropy of each feature is calculated by:

$$E_j = -(\ln n)^{-1} \sum_{i=1}^{n} p_{ij} \ln p_{ij} \tag{10}$$

where E_j is the information entropy of each feature, p_{ij} can be calculated by $p_{ij} = {y_{ij}}\big/{\sum_{i=1}^{n} y_{ij}}$.

Based on the value of information entropy E_j, the weighting factor of each feature is calculated by:

$$\omega_j = \frac{1 - E_j}{\sum_{j=1}^{n}(1 - E_j)} = \frac{1 - E_j}{n - \sum_{j=1}^{3} E_j} \tag{11}$$

where $\sum_{j=1}^{n} \omega_j = 1\text{and}0 \le \omega_j \le 1.1 - E_j$ indicates the inconsistency degree of each sample under the jth feature from the theory of information entropy. Then, the feature that can create a larger inconsistency degree among samples, in other words, which has a larger capacity to discriminate samples, would have a larger weighting coefficient.

Also, basing on the weights to each feature, it can easily get the features ranked.

Then, by combining the fractal dimension of the data set, which determine the numbers of key attributes well characterizing the data set, and the ranked features in each of the feature selection techniques, the key feature subset of the data set under each feature selection technique could be achieved.

The classification process to verify the robustness of the feature subsets

To verify the robustness of the feature subsets selected under the three feature selection techniques, three machine learning techniques, Naïve-Bayes classifier, K-nearest-neighbor classifier (KNN), and random forest classifier based on decision trees were employed on the obtained features.

Naïve-Bayes classifier (Langley et al. 1992) predicts the probability that a given sample belongs to a particular class. Given a sample ***X***, the classifier will predict that ***X*** belongs to the class having the highest a posteriori probability, conditioned on ***X***.

K-nearest Neighbor (KNN) classifier simply retains the entire training set during learning and assigns to each query a class represented by the majority label of its k-nearest neighbors in the training set. In the present study, we used the distance-weighted KNN proposed by Dudani (1976) to perform the classification process.

Random forest combines several randomized decision trees and aggregates their predictions by averaging (Breiman 2001). In this study, the random forest classifier consists of seven trees, with each is grown with the classification and regression tree (CART) algorithm (Breiman et al. 1984). To classify a new dataset, each case of the datasets is passed down to each of the seven trees. The forest chooses a class having the most out of seven votes to be as the final class label of the case.

The detailed calculation process for these three classifiers were discussed in the "Supplementary Material".

Experimental results and discussion

According to the analysis of fractal dimension in Sect. 5.1, the slope of (log(r),log(S(r)) in Eq. (1) was calculated with a value of 7.98, which is the lower bound of the number of attributes needed to keep the essential characteristics of the dataset. That is to say, in all

Table 5 Key feature subsets under the three feature selection techniques

Feature selection techniques	Feature subsets
Relief-F	$\{\mathbf{x_{12}}, \mathbf{x_{18}}, \mathbf{x_{14}}, \mathbf{x_{17}}, \mathrm{x}_{16}, \mathbf{x_{22}}, \mathrm{x}_{21}, \mathrm{x}_{13}\}$
PCA	$\{\mathbf{x_{17}}, \mathbf{x_{18}}, \mathbf{x_{22}}, \mathbf{x_{12}}, \mathrm{x}_{3}, \mathbf{x_{14}}, \mathrm{x}_{20}, \mathrm{x}_{21}\}$
EWM	$\{\mathbf{x_{18}}, \mathbf{x_{12}}, \mathrm{x}_{16}, \mathbf{x_{17}}, \mathrm{x}_{20}, \mathbf{x_{14}}, \mathbf{x_{22}}, \mathrm{x}_{23}\}$

Bold values indicate that these five features are the core characteristics of the initial data set

the twenty-three indices collected for the prediction task, there are at least eight ones are most significant to represent the original data set. According to this result, the first eight features were selected to be the most important representing features for the original data set according to the ranked results in each feature selection technique. Table 5 shows the results for the selected feature subset under each feature selection technique.

Not all the selected indices exist in Table 5. In fact, only eleven indices appear in the three feature subsets. This means that these eleven indices will have larger predictive capacities than those not appearing in Table 5.

Among the eleven features in Table 5, the indices $\{x_{12}, x_{14}, x_{17}, x_{18}, x_{22}\}$ exist in all the three feature subsets, which indicates that these five features are the core characteristics of the initial data set. That is to say, these five features are those playing dominated roles in determining which articles could grow up into highly-cited ones. Index x_{12} shows the first-cited age of articles, which represents the speed with which the results of a research is disseminated in the scientific community. Indices $\{x_{14}, x_{17}, x_{18}\}$ show the scope of knowledge diffusion for one article in the scientific community. Index x_{22} represents the saved times of the article in various social medias in the first 2-year time window after publication. It suggests that the alternative metrics, especially the saved activity, also has kernel influence on articles' future success. Besides these five corn characteristics, the h-index of the first author before publishing this article (x_3), the citations got in the first-cited age (x_{13}), the number of citing institutions in the first 3 years after publication (x_{16}), the total citations got in the first 3 years (x_{20}), the viewed times (x_{21}) and the discussed times (x_{23}) are all helpful for the future growth of articles.

Then, the three classifiers of Naïve-Bayes, K-nearest-neighbor (KNN), and random forests were operated to test the performance of these feature subsets on predicting future HCPs. Table 6 shows the final classification performance of each feature subset under each of the three classifiers.

Obviously, all the feature subsets have got considerable classification performance under each of the classifiers. Even the worst accuracy has reached to 0.882. And the feature

Table 6 The classification performance of the feature subsets

Classification performance	Relief-F	PCA	EWM	Ave.
Naïve Bayes	0.941	0.944	0.915	0.933
KNN	***0.882***	0.929	0.907	0.906
Random Forest	0.943	$\underline{\textbf{0.947}}$	0.886	0.925
Ave.	0.922	0.940	0.903	–

The feature subset selected by PCA has gained the largest precision with 0.947 trained by the randomforests subset selected by PCA has gained the largest precision with 0.947 trained by the random forests. In fact, the feature subset selected by PCA has got the best classification performance regardless of the classifiers. The last row in Table 6 gives the average classification accuracy for each of the feature subsets under different classifiers; and the last column is for the average classification accuracy for each of the classifiers under different feature subsets. All the average classification accuracies are above 0.9 no matter to classifiers or to feature subsets. The results indicate that the feature subsets extracted by the three feature selection techniques are stable and valuable to classify and predict the future HCPs, although there's little differences for the values of accuracy. Based on this considerable classification results, this study combines all the features appearing in the three feature subsets as the typical indicators for the future HCPs.

Table 7 shows the final typical indicators for the future HCPs. There are eleven features are verified to be the typical indicators for future HCPs, with eight from traditional bibliometric indices and three from alternative metrics. This indicates that both the bibliometric indices and the alternative metrics do benefit to the future success of articles. And the altmetrics does offer new ways to measure the impact of publications which may complement rather than replace traditional indicators for research evaluation and prediction. Table 7 also shows the frequency of occurrence of each index. Features $\{x_{12}, x_{14}, x_{17}, x_{18}, x_{22}\}$ exist in all the three feature subsets, showing their leading role in predicting future HCPs. Indices $\{x_{16}, x_{20}, x_{21}\}$ are also existed as the informative features with two times' occurrence. Indices $\{x_3, x_{13}, x_{23}\}$ present for one time in the feature subsets. Although these eleven features are diverse in the occurrence and in the predictive capacities, we believe that the perfect classification performance is the result of the combined effects of these indices. Thus, all these eleven indices in Table 7 are presented as the typical indicators for the future HCPs. Here, we give a detailed analysis on these features.

(1) The first-cited performance of articles: The two first-cited indices $\{x_{12}, x_{13}\}$for articles are both selected as the valid predictors, where x_{12} shows the citation rate and x_{13} is the citation strength in the first-cited year. Van Dalen and Henkens (2005) stated that the status of uncitedness of a paper becomes a stigma and the longer a paper is uncited, the lower its quality and the less inclined researchers will be to cite it. This stigma for uncitedness indicates the important role of a paper's first-cited performance on its later citation life. Our results show that a quicker accept speed and a higher accept strength for one article in the scientific community are important for its future success. This results is consistent with the conclusions by some previous works (Adams 2005; Bornmann and Daniel 2010; Chakraborty et al. 2014; Garner et al. 2014; Guerrero-Bote and Moya-Anegón 2014; Hilmer and Lusk 2009a, b). These works stated that based on the speed with which the results of a research is disseminated in the scientific community and is being cited, future citations can be predicted.

(2) The early-stage citation diffusion performance of articles: Five indices $\{x_{14}, x_{17}, x_{18}, x_{16}, x_{20}\}$associated with articles' citation diffusion performance in their early stage, 3 years after publication in this study, are also considerable typical indicators for future HCPs. These five indices show the scope of the influence of articles in their citing environment. If one article has got citations from more countries, institutions, subjects and journals, etc., it means that the knowledge carried by the article has been diffused into a more diverse fields accompany with the occurrence of citation activities. This wider visibility of the article impact, in turn, bring more opportunities for articles to

Table 7 Typical indices for the future HCPs

Indices	Detailed information	Frequency of occurrence	Source
x_{12}	The first-cited age of this article	3	Bibliometrics index
x_{14}	The number of countries citing this paper in the first 3 years after publication	3	Bibliometrics index
x_{17}	The number of journals citing this paper in the first 3 years after publication	3	Bibliometrics index
x_{18}	The number of subjects citing this paper in the first 3 years after publication	3	Bibliometrics index
x_{22}	The saved times in the first 2 years after publication	3	Alternative metric
x_{16}	The number of institutions citing this paper in the first 3 years after publication	2	Bibliometrics index
x_{20}	The total citations to this article in the first 3 years after publication	2	Bibliometrics index
x_{21}	The viewed times in the first 2 years after publication	2	Alternative metric
x_3	The *h*-index of the first author before publishing this articles	1	Bibliometrics index
x_{13}	The citations to this article in the first-cited year	1	Bibliometrics index
x_{23}	The discussed times in the first 2 years after publication	1	Alternative metric

gain new citations. Aksnes (2003) stated that the increasing visibility of paper would lead to further citations to it. Our experimental results confirm this viewpoint.

(3) The early-stage web usage statistics of articles: All the three indices$\{x_{21}, x_{22}, x_{23}\}$ related to articles' web usage are extracted as typical predictors. Numbers of works have presented to verify the positive relationship between the number of citations and various alternative metrics by statistical techniques (McCabe and Snyder 2015; Yuan and Hua 2011; Eysenbach 2011; Neylon and Wu 2009; Thelwall et al. 2013; Li et al. 2012; Haustein et al. 2013; Zahedi et al. 2013). However, fewer have focused on the predictive performance of these metrics on future citation numbers. Our study made a preliminary attempt on this issue and showed that the earlier web usage metrics can also be valuable predictors of later citation flourish.

(4) The prestige of authors: The *h* index of the first author before publishing this article, labelled as index $\{\mathrm{x}_3\}$ is also an valuable predictor to show the importance of author's prestige. A larger *h*-index indicates that the author has gained considerable research capabilities or reputations in science. Hurley et al. (2013) presented that *h*-index of the author group influences citation frequency. Schilling and Green (2011) stated that prior experience and publishing success may act as a signaling and legitimization device that serves to increase the likelihood of others reading and citing the work. Researcher also showed that the number of citations to an author's precious papers can be considered as a good predictor for citations to further papers (Tang et al. 2014; Walters 2006; Yu et al. 2014). Our study show that comparing with the authors' previous citations, *h*-index of the first author would be a more predictive indicator.

Conclusions

Is it possible to identify the importance of an article earlier in the read-cite cycle? This paper is aiming to answer the above question by identifying the early-stage predictors from traditional bibliometric indices and alternative metrics. By establishing a feature space with twenty-three indices, a manifold characters from authors, journals, articles, early citation distribution performance, as well as the early web usage statistics were tested to find the key predictors for articles' future success. Combining with the fractal dimension of the data set, three feature selection techniques of Relief-F, principal component analysis (PCA) and entropy weighted method (EWM) were performed to extract the key feature subsets which can better represent the original data set. Then three kinds of classifiers, Naïve Bayes, KNN and random forest, were taken to verify the classification performance of the feature subsets. Experimental results on articles published in the seven journals in PLOS showed that both traditional bibliometric indices and alternative metrics are valuable predictors for future HCPs. These predictors are mainly from four aspects: the first-cited performance of articles, the early-stage citation diffusion characters of articles, the early-stage web usage statistics of articles, and the prestige of authors. It indicates that the data altmetric collects is also a useful leading indicator for later success. The web-based access to the research literature does offer a potential measure for the future impact of articles. And the combination of traditional and alternative metrics for research prediction could provide more complete article profiles as it captures more dimensions of scientific practice.

The limitation of the present analysis is that it focused on the seven journals published in PLOS only, where all the articles collected are life sciences and medical literatures. The results on this limited corpus would not be universal to articles in other fields. And the current study is based solely on the analysis of alternative indicators provided by PLOS, and there may be other alternative indicators on other sites, but this is not covered in this study. However, PLOS ALM is an important altmetrics application platform with high authority and availability. The results from PLOS could also provide certain reference on the following evaluation or prediction-related research.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 71473034), the financial assistance from Postdoctoral Scientific Research Developmental Fund of Heilongjiang Province (Grant No. LBH-Q16003), and the national undergraduate training programs for innovation (Grant No. 201510225167).

References

Abdi, H., Williams, L. J., & Valentin, D. (2013). Multiple factor analysis: Principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics, 5*(2), 149–179.

Adams, J. (2005). Early citation counts correlation with accumulated impact. *Scientometrics, 63*(3), 567–581.

Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation, 12*(3), 159–170.

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.

Annalingam, A., Damayanthi, H., Jayawardena, R., & Ranasinghe, P. (2014). Determinants of the citation rate of medical research publications from a developing country. *SpringerPlus, 3*(1), 1–6.

Antoniou, G. A., Antoniou, S. A., Georgakarakos, E. I., Sfyroeras, G. S., & Georgiadis, G. S. (2015). Bibliometric analysis of factors predicting increased citations in the vascular and endovascular literature. *Annals of Vascular Surgery, 29*(2), 286–292.

Belussi, A., & Faloutsos, C. (1995). Estimating the selectivity of spatial queries using the 'correlation' fractal dimension. In *Proceedings of the 21th international conference on very large data bases* (pp. 299–310).

Berchtold, S., Böhm, C., & Kriegel, H.-P. (1998). The pyramid-tree: Breaking the curse of dimensionality. In *Proceedings of the ACM SIGMOD international conference on management of data* (pp. 142–153).

Biscaro, C., & Giupponi, C. (2014). Co-authorship and bibliographic coupling network effects on citations. *PLoS ONE, 9*(6), e99502.

Bjarnason, T., & Sigfusdottir, I. D. (2002). Nordic impact: Article productivity and citation patterns in sixteen Nordic Sociology departments. *Acta Sociologica, 45*(4), 253–267.

Bornmann, L. (2013). The problem of citation impact assessments for recent publication years in institutional evaluations. *Journal of Informetrics, 7*(3), 722–729.

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics, 8*(4), 895–903.

Bornmann, L., & Daniel, H. D. (2010). Citation speed as a measure to predict the attention an article receives: An investigation of the validity of editorial decisions at Angewandte Chemie International Edition. *Journal of Informetrics, 4*(1), 83–88.

Bornmann, L., Schier, H., Marx, W., & Daniel, H.-D. (2012). What factors determine citation counts of publications in chemistry besides their quality? *Journal of Informetrics, 6*(1), 11–18.

Bornmann, L., & Williams, R. (2013). How to calculate the practical significance of citation impact differences? An empirical example from evaluative institutional bibliometrics using adjusted predictions and marginal effects. *Journal of Informetrics, 7*(2), 562–574.

Borsuk, R. M., Budden, A. E., Leimu, R., Aarssen, L. W., & Lortie, C. J. (2009). The influence of author gender, national language and number of authors on citation rate in ecology. *Open Ecology Journal, 2*(1), 25–28.

Bosquet, C., & Combes, P. P. (2013). Are academics who publish more also more cited? Individual determinants of publication and citation records. *Scientometrics, 97*(3), 831–857.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. New York: Wadsworth.

Buela-Casal, G., & Zych, I. (2010). Analysis of the relationship between the number of citations and the quality evaluated by experts in psychology journals. *Psicothema, 22*(2), 270–276.

Chakraborty, T., Kumar, S., Goyal, p. Ganguly, N. & Mukherjee, A. (2014). Towards a stratified learning approach to predict future citation counts. In *Proceedings of the ACM/IEEE joint conference on digital libraries*.

Chen, C. M. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology, 63*(3), 431–449.

Chi, P. S., & Glänzel, W. (2017). An empirical investigation of the associations among usage, scientific collaboration and citation impact. *Scientometrics, 112*(1), 403–412.

Chi, P. S., & Glänzel, W. (2018). Comparison of citation and usage indicators in research assessment in scientific disciplines and journals. *Scientometrics, 116*(1), 537–554.

Collet, F., Robertson, D. A., & Lup, D. (2014). When does brokerage matter? Citation impact of research teams in an emerging academic field. *Strategic Organization, 12*(3), 157–179.

Dalen, Van, & Henkens, H. P. K. (2005). Signals in science-On the importance of signaling in gaining attention in science. *Scientometrics, 64*(2), 209–233.

de Winter, J. (2015). The relationship between tweets, citations, and article views for PLOS ONE articles. *Scientometrics, 102*(2), 1773–1779.

Didegah, F., Bowman, T. D., & Holmberg, K. (2018). On the difference between citations and altmetrics: An investigation of factors driving altmetrics versus citations for Finnish articles. *Journal of the Association for Information Science and Technology, 69*(6), 832–843.

Didegah, F., & Thelwall, M. (2013). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology, 64*(5), 1055–1064.

Dorta-González, P., Dorta-González, M. I., Santos-Peñate, D. R., & Suárez-Vega, R. (2014). Journal topic citation potential and between-field comparisons: The topic normalized impact factor. *Journal of Informetrics, 8*(2), 406–418.

Dudani, S. A. (1976). The distance-weighted k-nearest neighbor rule. *IEEE Transactions on System Man and Cybernetics, 6*(4), 325–327.

Ebrahim, N. A., Salehi, H., Embi, M. A., Tanha, F. H., Gholizadeh, H., & Motahar, S. M. (2014). Visibility and citation impact. *International Education Studies, 7*(4), 120–125.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research, 13*(4), e123.

Falagas, M. E., Zarkali, A., Karageorgopoulos, D. E., Bardakas, V., & Mavros, M. N. (2013). The impact of article length on the number of future citations: A bibliometric analysis of general medicine journals. *PLoS ONE, 8*(2), e49476.

Farshad, M., Sidler, C., & Gerber, C. (2013). Association of scientific and nonscientific factors to citation rates of articles of renowned orthopedic journals. *European Orthopedics and Traumatology, 4*(3), 125–130.

Fausto, S., Machado, F. A., Bento, L. F. J., Iamarino, A., Nahas, T. R., & Munger, D. S. (2012). Research blogging: indexing and registering the change in science 2.0. *PLoS One, 7*(12), e50109.

Frandsen, T. F., & Nicolaisen, J. (2013). The ripple effect: Citation chain reactions of a nobel prize. *Journal of the American Society for Information Science and Technology, 64*(3), 437–447.

Garner, J., Porter, A. L., & Newman, N. C. (2014). Distance and velocity measures: Using citations to determine breadth and speed of research impact. *Scientometrics, 100*(3), 687–703.

Glänzel, W. (2008). Seven myths in bibliometrics. About facts and fiction in quantitative science studies. In *4th International conference on webometrics, informetrics and scientometrics & 9th COLLNET meeting*, Berlin, Germany.

Glänzel, W., & Heeffer, S. (2014). Cross-national preferences and similarities in downloads and citations of scientific articles: a pilot study. In E. Noyons (Ed.), *Proceedings of the STI conference 2014*, Leiden (pp. 207–215).

Glänzel, W., Rousseau, R., & Zhang, L. (2012). A visual representation of relative first-citation times. *Journal of the American Society for Information Science and Technology, 63*(7), 1420–1425.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics, 56*(3), 357–367.

Glänzel, W., & Thijs, B. (2004). Does co-authorship inflate the share of self-citations? *Scientometrics, 61*(3), 395–404.

Glänzel, W., Thijs, B., & Debackere, K. (2014). The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparision and institutional research assessment. *Scientometrics, 101*(2), 939–952.

Gonzalez-Alcaide, G., Calafat, A., Becona, E., Thijs, B., & Glänzel, W. (2016). Co-citation analysis of articles published in substance abuse journals: Intellectual structure and research fields (2001-2012). *Journal of Studies on Alcohol and Drugs, 77*(5), 710–722.

Guerrero-Bote, V. P., & Moya-Anegón, F. (2014). Relationship between downloads and citations at journal and paper levels, and the influence of language. *Scientometrics, 101*(2), 1043–1065.

Haslam, N., & Koval, P. (2010). Predicting long-term citation impact of articles in social and personality psychology. *Psychological Reports, 106*(3), 891–900.

Haustein, S., Larivière, V., Thelwall, M., Amyot, D., & Peters, I. (2014). Tweets vs. Mendeley readers: How do these two social media metrics differ? *IT-Information Technology, 56*(5), 207–215.

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Hadas, S., & Terliesner, J. (2013). Coverage and adoption of altmetrics sources in the bibliometric community. In *Proceeding of 14th international society of scientometrics and informatics conference* (pp. 468–483).

Herrmannova, D., Patton, R. M., Knoth, P., & Stahl, C. G. (2018). Do citations and readership identify seminal publications? *Scientometrics, 115*(1), 239–262.

Hilmer, C. E., & Lusk, J. L. (2009a). Determinants of citations to the agricultural and applied economics association journals. *Review of Agricultural Economics, 31*(4), 677–694.

Hilmer, C. E., & Lusk, J. L. (2009b). Determinants of citations to the agricultural and applied economics association journals. *Reviews of Agricultural Economics, 31*(4), 677–694.

Huang, H., Andrews, J., & Tang, J. (2012). Citation characterization and impact normalization in bioinformatics journals. *Journal of the American Society for Information Science and Technology, 63*(3), 490–497.

Huang, S., Chang, J., Leng, G., & Huang, Q. (2015). Integrated index for drought assessment based on variable fuzzy set theory: A case study in the Yellow River basin. *Journal of Hydrology, 527,* 608–618.

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 ASIST Annual Meeting, 50*(1), 1–10.

Ibáñez, A., Bielza, C., & Larrañaga, P. (2013). Relationship among research collaboration, number of documents and number of citations: A case study in Spanish computer science production in 2000-2009. *Scientometrics, 95*(2), 689–716.

Ingwersen, P., & Larsen, B. (2014). Influence of a performance indicator on Danish research production and citation impact 2000-12. *Scientometrics, 101*(2), 1325–1344.

Ingwersen, P., Larsen, B., Garcia-Zorita, J. C., Serrano-Lopez, A. E., & Sanz-Casado, E. (2014). Influence of proceedings papers on citation impact in seven sub-fields of sustainable energy research 2005-2011. *Scientometrics, 101*(2), 1273–1292.

Ke, S. W., Lin, W. C., Tsai, C. F., & Hu, Y. H. (2014). Citation impact analysis of research papers that appear in oral and poster sessions: A case study of three computer science conference. *Online Information Review, 38*(6), 738–745.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. In *European conference on machine learning* (pp. 171–182).

Korn, F., Pagel, B.-U., & Faloutsos, C. (2001). On the ‘dimensionality curse’ and the ‘self-similarity blessing’. *IEEE TKDE, 13,* 96–111.

Kousha, K., & Thelwall, M. (2017). Are Wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology, 68*(3), 762–779.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifers. In *Proceedings of the 10th national conference on artificial intelligence* (pp. 223–228).

Lee, S. Y., Lee, S., & Jun, S. H. (2010). Author and article characteristics, journal quality and citation in economic research. *Applied Economics Letters, 17*(17), 1697–1701.

Leimu, R., & Koricheva, J. (2005). What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution, 20*(1), 28–32.

Li, X., Thelwall, M., & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. *Scientometrics, 91*(2), 461–471.

Lira, R. P. C., Vieira, R. M. C., Goncalves, F. A., Ferreira, M. C. A., Maziero, D., & Arieta, C. E. L. (2013). Influence of English language in the number of citations of articles published in Brazilian journals of Ophthalmology. *Arquivos Brasileiros de Oftalmologia, 76*(1), 26–28.

Marashi, S. A., Hosseini-Nami, S., Alishah, K., Hadi, M., Karimi, A., Hosseinian, S., et al. (2013). Impact of wikipeida on citation trends. *Excli Journal, 12,* 15–19.

McCabe, M. J., & Snyder, C. M. (2015). Does online availability increase citations? Theory and evidence from a panel of economics and business journals. *Review of Economics and Statistics, 97*(1), 144–165.

Miettunen, J., & Nieminen, P. (2003). The effect of statistical methods and study reporting characteristics on the number of citations: A study of four general psychiatric journals. *Scientometrics, 57*(3), 377–388.

Naraei, P. & Sadeghian, A. (2017). A PCA based feature reduction in intracranial hypertension analysis. In *IEEE international conference on 30th Canadian conference on electrical and computer engineering* (pp. 1–6).

Neylon, C., & Wu, S. (2009). Article-level metrics and the evolution of scientific impact. *PLoS Biology, 7*(11), e1000242.

Nomaler, T., Frenken, K., & Heimeriks, G. (2013). Do more distant collaborations have more citation impact? *Journal of Informetrics, 7*(4), 966–971.

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.

Onyancha, O. B., & Maluleka, J. R. (2011). Knowledge production through collaborative research in sub-Saharan Africa: How much do countries contribute to each other's knowledge output and citation impact? *Scientometrics, 87*(2), 315–336.

Ortega, J. (2016). To be or not to be on Twitter, and its relationship with the tweeting and citation of research papers. *Scientometrics, 109*(2), 1353–1364.

Padial, A. A., Nabout, J. C., Siqueira, T., Bini, L. M., & Diniz-Filho, J. A. F. (2010). Weak evidence for determinants of citation frequency in ecological articles. *Scientometrics, 85*(1), 1–12.

Pagel, P. S., & Hudetz, J. A. (2011). Scholarly productivity of United States academic cardiothoracic anesthesiologists: Influence of fellowship accreditation and transesophageal echocardiographic credentials on h-index and other citation bibliometrics. *Journal of Cardiothoracic and Vascular Anesthesia, 25*(5), 761–765.

Pagel, B.-U., Korn, F. & Faloutsos, C. (2000). Deflating the dimensionality curse using multiple fractal dimensions. In *16th ICDE* (pp. 589–598).

Patterson, M. S., & Harris, S. (2009). The relationship between reviewers' quality-scores and number of citations for papers published in the journal physics in medicine and biology from 2003-2005. *Scientometrics, 80*(2), 343–349.

Peoples, B., Midway, S., Sackett, D., Lynch, A., & Cooney, P. B. (2016). Twitter predicts citation rates of ecological research. *PLoS ONE, 11,* e0166570.

Piwowar, H., & Priem, J. (2013). The power of altmetrics on a CV. *Bulletin of the Association for Information Science and Technology, 39*(4), 10–13.

Priem, J., & Hemminger, B. M. (2010). Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. *First Monday*, 15(7). Retrieved from https://journals.uic.edu/ojs/index.php/fm/article/view/2874/2570.

Priem, J., Parra, C., Piwowar, H., Groth, P., & Waagmeester, A. (2012). Uncovering impacts: a case study in using altmetrics tools. In *Second international conference on the future of scholarly communication and scientific publishing*. Heraklion, Greece. http://jasonpriem.org/self-archived/altmetrics/sepublica/cameraready.pdf. Accessed 19 Mar 2013.

Puuska, H. M., Muhonen, R., & Leino, Y. (2014). International and domestic co-publishing and their citation impact in different disciplines. *Scientometrics, 98*(2), 823–839.

Ravenscroft, J., Liakata, M., Clare, A., & Duma, D. (2017). Measuring scientific impact beyond academia: an assessment of existing impact metrics and proposed improvements. *PLoS ONE, 12,* e0173152.

Rees, T., Ayling-Rouse, K., & Smith, S. (2012). Accesses versus citations: Why you need to measure both to assess publication impact. *Current Medical Research and Opinion, 28,* S9–S10.

Ringelhan, S., Wollersheim, J., & Welpe, I. (2015). I like, I cite? Do facebook likes predict the impact of scientific work? *PLoS ONE, 10,* e0134389.

Royle, P., Kandala, N. B., Barnard, K., & Waugh, N. (2013). Bibliometrics of systematic reviews: Analysis of citation rates and journal impact factors. *Systematic reviews, 2,* 74.

Sangwal, K. (2012). On the relationship between citations of publication output and Hirsch index h of authors: Conceptualization of tapered Hirsch index h T, circular citation area radius R and citation acceleration a. *Scientometrics, 93*(3), 987–1004.

Schilling, M. A., & Green, E. (2011). Recombinant search and breakthrough idea generation: An analysis of high impact papers in the social sciences. *Research Policy, 40*(10), 1321–1331.

Shu, F., Lou, W., & Haustein, S. (2018). Can Twitter increase the visibility of Chinese publications? *Scientometrics, 116*(1), 505–519.

Sin, S. C. J. (2011). International coauthorship and citation impact: A bibliometric study of six LIS journals, 1980-2008. *Journal of the American Society for Information Science and Technology, 62*(9), 1770–1783.

Stremersch, S., Camacho, N., Vanneste, S., & Verniers, I. (2015). Unraveling scientific impact: Citation types in marketing journals. *International Journal of Research in Marketing, 32*(1), 64–77.

Stremersch, S., Verniers, I., & Verhoef, P. C. (2007). The quest for citations: Drivers of article impact. *Journal of Marketing, 71*(3), 171–193.

Syamili, C., & Rekha, R. V. (2017). Do altmetric correlate with citation? A study based on PLOS ONE journal. *Journal of Scientometrics and Information Management, 11*(1), 103–117.

Tahamtan, I., Afshar, A. S., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics, 107*(3), 1195–1225.

Tang, X., Wang, L., & Kishore, R. (2014). Why do is scholars cite other scholars? An empirical analysis of the direct and moderating effects of cooperation and competition among is scholars on individual citation behavior C3. In *35th International conference on information systems (ICIS 2014).*

Taylor, M. (2013). Exploring the boundaries: how altmetrics can expand our vision of scholarly communication and social impact. *Information Standards Quarterly, 25*(2), 27–32.

Thelwall, M. (2018). Early Mendeley readers correlate with later citation counts. *Scientometrics, 115*(3), 1231–1240.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. (2013). Do altmetrics work? Twitter and ten other candidates. *PLoS ONE, 8*(5), e64841.

Thelwall, M., & Wilson, P. (2016). Mendeley readership altmetrics for medical articles: An analysis of 45 fields. *Journal of the Association for Information Science and Technology, 67*(8), 1962–1972.

Traina, C., Traina, A., Wu, L., & Faloutsos, C. (2000). Fast feature selection using fractal dimension. In *Proceeding 15th Brazilian symposium on database (SBBD)* (pp. 158–171).

Van Der Pol, C. B., McInnes, M. D. F., Petrcich, W., Tunis, A. S., & Hanna, R. (2015). Is quality and completeness of reporting of systematic reviews and meta-analyses published in high impact radiology journals associated with citation rates? *PLoS ONE, 10*(3), e011892.

van Eck, N. J., Waltman, L., van Raan, A. F. J., Klautz, R. J. M., & Peul, W. C. (2013). Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PLoS ONE, 8*(4), e62395.

Van Wesel, M., Wyatt, S., & ten Haaf, J. (2014). What a difference a colon makes: How superficial factors influence subsequent citation. *Scientometrics, 98*(3), 1601–1615.

Vanclay, J. K. (2013). Factors affecting citation rates in environmental science. *Journal of Informetrics, 7*(2), 265–271.

Vieira, E. S., & Gomes, J. A. N. F. (2010). Citations to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics, 4*(1), 1–13.

Walters, G. D. (2006). Predicting subsequent citations to article published in twelve crime-psychology journals: Author impact versus journal impact. *Scientometrics, 69*(3), 499–510.

Waltman, L., & Costas, R. (2014). F1000 recommendations as a potential new data source for research evaluation: A comparison with citations. *Journal of the Association for Information Science and Technology, 65*(3), 433–445.

Wang, X., Liu, C., Fang, Z., & Mao, W. (2014). From attention to citation, what and how does altmetrics work? http://arxiv.org/abs/1409.4269

Wang, J., Thijs, B., & Glänzel, W. (2015a). Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PLoS ONE, 10*(5), e0127298.

Wang, L., Thijs, B., & Glänzel, W. (2015b). Characteristics of international collaboration in sport sciences publications and its influence on citation impact. *Scientometrics, 105*(2), 843–862.

Wang, M., Yu, G., An, S., & Yu, D. (2012a). Discovery of factors influencing citation impact based on a soft fuzzy rough set model. *Scientometrics, 93*(3), 635–644.

Wang, M., Yu, G., Xu, J., He, H., Yu, D., & An, S. (2012b). Development a case-based classifier for predicting highly cited papers. *Journal of Informetrics, 6*(4), 586–599.

Willis, D. L., Bahler, C. D., Neuberger, M. M., & Dahm, P. (2011). Predictors of citations in the urological literature. *BJU International, 107*(12), 1876–1880.

Xu, J. L., Xu, B. W., Zhang, W. F., & Cui, Z. F. (2008). Principal component analysis based feature selection for clustering. In *2008 international conference on machine learning and cybernetics* (Vol. 1, pp. 460–465).

Yu, T., & Yu, G. (2014). Features of scientific papers and the relationships with their citation impact. *Malaysian Journal of Library and Information Science, 19*(1), 37–50.

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.

Yuan, S. B., & Hua, W. N. (2011). Scholarly impact measurements of LIS open access journals: Based on citations and links. *Electronic Library, 29*(5), 682–697.

Yue, W. P., & Wilson, C. S. (2004). Measuring the citation impact of research journals in clinical neurology: A structural equation modelling analysis. *Scientometrics, 60*(3), 317–332.

Zahedi, Z., Costas, R. & Wouters, P. (2013). How well developed are Altmerics? Cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. In *14th International society of scientometrics and informatics conference* (pp. 876–884).