Scientometrics (2016) 107:1265–1282
DOI 10.1007/s11192-016-1926-1

Successful fish go with the flow: citation impact prediction based on centrality measures for term–document networks

Peter Klimek[1,2] · **Aleksandar S. Jovanovic**[2,3] · **Rainer Egloff**[4] · **Reto Schneider**[4]

Received: 1 October 2015 / Published online: 12 April 2016

Abstract In this work we address the challenge of how to identify those documents from a given set of texts that are most likely to have substantial impact in the future. To this end we develop a purely content-based methodology in order to rank a given set of documents, for example abstracts of scientific publications, according to their potential to generate impact as measured by the numbers of citations that the articles will receive in the future. We construct a bipartite network consisting of documents that are linked to keywords and terms that they contain. We study recursive centrality measures for such networks that quantify how many different terms a document contains and how these terms are related to each other. From this we derive a novel indicator—document centrality—that is shown to be highly predictive of citation impact in six different case studies. We compare these results to findings from a multivariable regression model and from conventional network-based centrality measures to show that document centrality indeed offers a comparably high performance in identifying those articles that contain a large number of high-impact keywords. Our findings suggest that articles which conform to the mainstream within a given research field tend to receive higher numbers of citations than highly original and innovative articles.

Keywords Citation impact prediction · Term–document matrix · Network theory · Bipartite networks

Electronic supplementary material The online version of this article (doi:10.1007/s11192-016-1926-1) contains supplementary material, which is available to authorized users.

✉ Aleksandar S. Jovanovic
jovanovic@risk-technologies.com

[1] Section for Science of Complex Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria

[2] Steinbeis Advanced Risk Technologies, Willi-Bleicher-Straße 19, 70174 Stuttgart, Germany

[3] EU-VRi, Willi-Bleicher-Straße 19, 70174 Stuttgart, Germany

[4] Swiss Reinsurance Company Ltd, Mythenquai 50/60, 8022 Zurich, Switzerland

Introduction

It was estimated that as of 2012 there are 571 new websites created every minute of the day, 175 million tweets written each day, and 100 terabytes of data uploaded daily to Facebook.[1] The doubling time of scientific publications is estimated to be in the range of 10–15 years (Larsen and von Ins 2010). This rapid growth poses not only a technological challenge, but it brings about the problem of data relevance. In this ever-growing volume of novel scientific findings, how can we identify those that are likely to have the greatest impact in the future? What is really worth our attention? To face the challenges that this new "big-data-age" brings about we are therefore in dire need of novel methods to retrieve those documents and texts that might be of the greatest relevance for a given problem at hand.

In this work we address the question of how to identify those texts from a set of documents that are most likely to have the biggest impact in the future. We frame this problem in the context of scientific publications, where a commonly accepted method for measuring the impact of an article is its number of citations (MacRoberts and MacRoberts 1996). In particular, we develop a purely content-based methodology in order to rank a given set of documents (abstracts of scientific publications) according to their potential to generate impact as measured by the future numbers of citations of the articles. Our aim is to develop a simple and model-free indicator for citation impact prediction that captures not only how many and which terms a document is described by, but also how these terms are related to each other. To this end we consider the abstracts of all publications within a given year for a given topic and construct the term–document matrix for this collection of documents. This term–document matrix can be represented as a bipartite network with two sets of nodes, one set corresponding to documents (abstracts) and the other set of nodes representing terms. A term and a document are connected by a link if the document contains the given term. We propose a recursively defined centrality measure for such bipartite networks that has already been shown to be predictive for economic growth of a country based on its basket of exported products (Hidalgo and Hausmann 2009), as well as for the average number of sales for albums in a given music genre based on their typical instrumentation (Percino et al. 2014). The idea is to recursively define the "centrality" or "importance" of a term by assuming that a term is central in a network sense if it occurs in a large number of documents that contain a large number of other central terms. We confirm with high statistical significance in six independent case studies that a document, compared to other, similar documents published within the same year, is more likely to receive a high number of citations, if its abstract contains a large number of terms characterized by a high centrality in the bipartite term–document network. This is captured by a novel indicator that is derived from these bipartite centrality measures, the *document centrality*, C. We compare these results to findings obtained from traditional centrality measures and a regression model where we fit the citations received by a publication using the presence or absence of a specific term as variables in a regression. Document centrality is highly correlated with results obtained a posteriori from the regressions, which corroborates that the bipartite network structure of the term–document matrix is indeed predictive of the future impact of the documents as measured in the number of citations.

Since it is well known that citation counts vary across different scientific fields (Radicchi and Castellano 2012), we focus on six different case studies to validate the predictive value of document centrality. We use different topics from the field of "Material Science" which can be easily identified by a simple search query. These six topics were

[1] http://wikibon.org/blog/big-data-statistics/, retrieved 07/29/2015.

selected through a stakeholder process during the EU FP7 project iNTegRisk that identified subjects in Material Science that are, both, currently active areas of research and that have the potential to generate new findings with significant economic, societal and environmental repercussions (Jovanovic and Renn 2013). These case studies include topics which showed hugely varying differences in their increase of the number of published articles over time, as well as in their overall numbers of publications. This shows that the proposed indicator, document centrality, works in both cases, namely in the identification of potential high-impact works about novel, emerging topics (e.g. nanotechnology or brain–computer interfaces) or about established topics where there already exists a solidified body of knowledge (e.g. aging of materials).

Related work

Several approaches have been proposed to understand and, if possible, predict the citation impact of scientific publications. There are two types of features that are typically used to predict citation impact, namely content-based and bibliometric or extrinsic features (Fu and Aliferis 2010). Bibliometric features include, for instance, the reputation effect (Stewart 1983; Danell 2011), i.e. that the rate at which authors attract citations increases with the number of citations an author already has. It has also been shown that papers published in journals with high reputations tend to receive more citations than second-tier journals (Van Dalen and Henkens 2001; Callaham et al. 2002; Didegah and Thelwall 2013) and that articles that cite highly-cited or a large number of other works will be more often cited themselves (Bornmann et al. 2012; Vieira and Gomes 2010). Another extrinsic feature is the domain of the published paper, since citation counts vary between different research fields (Radicchi and Castellano 2012). This finding triggered interest in the question of how to normalize citation counts across different disciplines (Garfield 1979; Leydesdorff and Bornmann 2011). It has also been shown that social media activity within the first 3 days of article publication allows to predict citations (Eysenbach 2011). Content-based features that are known to be related to the received citations include the terms that occur in the title, abstract, or keywords of the article (Fu and Aliferis 2010; Yu et al. 2012).

The use of content-based features, e.g. the occurrences of specific terms, to predict citation counts in the scientific literature is a paradigmatic application of several machine learning approaches. These approaches include Naïve Bayesian models where the membership of a document to a specific class is inferred from the frequencies of specific terms within these classes (Gelman et al. 2004; Feng et al. 2011). Another popular approach utilizes support vector machines or networks (Cortes and Vapnik 1995) that map documents into a high-dimensional feature space where a decision surface is constructed that distinguishes different classes of documents (Fu and Aliferis 2010; Kwok 1998; Meyer et al. 2003). The k-nearest neighbor approach achieves a similar task of assigning membership to classes in a non-parametric manner based on the membership of its k nearest neighbors in an abstract feature space (Altman 1992; Jian et al. 2014). In the so-called relational topic model documents are modeled as collections of words and their co-occurrences, which allows to predict the citations of a given paper by examining which other papers share similar topics with the considered publication (Chang and Blei 2009). A similar approach is based on joint latent space models for topics in the texts on one and citations on the other side (Nallapati et al. 2008). For instance, in Latent Dirichlet Allocation models each term or token in a document is associated with a latent variable that is in turn related to one of the underlying topics (Hofmann 2001; Blei et al. 2003; Dietz et al.

2007). In such topic models it is also possible to represent the influence of the communities of co-authors of the authors of a given paper (Liu et al. 2009). Other approaches to construct models for citation impact prediction (using, both, extrinsic and content-based features) include regression (Yu et al. 2014) and mechanistic models (Wang et al. 2013).

The usefulness of network-based measures to quantify the impact of scientific work or of authors has been demonstrated repeatedly using, for instance, coauthorship networks (Newman 2004), article citation networks (Chen et al. 2007), author citation networks (Radicchi et al. 2009), journal citation networks (Bollen et al. 2006), or author cocitation networks (Leydesdorff 2007). Many of these works relied on conventional centrality measures, such as Google's PageRank (Chen et al. 2007) or the betweenness centrality (Leydesdorff 2007). It was soon realized that adaptations of these network measures can lead to better performance in ranking scientific works and productivity. These adaptations that often employ modified weighting schemes for citations, authors, and/or articles include AuthorRank (Liu et al. 2007), Y-Factor (Bollen et al. 2006), CiteRank (Walker et al. 2007), FutureRank (Sayyadi and Getoor 2009), or P-Rank (Yan et al. 2011). Systematic comparisons of traditional bibliometric with network-based indicators suggest that these measures can be classified into two classes that roughly translate into "impact" and "popularity" of the articles (Leydesdorff 2009; Bollen et al. 2009). Note that in this work we do not address coauthorship or related networks, but focus on the structure of the bipartite network that underlies the term–document matrix.

Data and methods

Abstracts

We focus on six different aspects and examples of emerging technologies. For each of these topics we handcrafted a query that was used to extract relevant scientific literature from Web Of Knowledge.[2] For each of the topics we only retrieved articles classified as "Science and Technology" and "Material Science". The "Topic" search function of Web Of Knowledge returns all indexed published articles where the search query or its word stem (lemmatization to identify e.g. plurals or different verb tenses) occurs in the title, abstract, or the author keywords. The six case studies are the following.

- Aging of materials, search query "aging" (which also returns results for "ageing").
- Brain–computer interfaces, search query "brain AND computer AND interface".
- Hydraulic fracturing, search query "fracking".
- Graphene, search query "graphene".
- Liquid natural gas, search query "liquid AND natural AND gas".
- Nanotechnology, search query "nanotech*".

We constructed a corpus of documents by extracting each entry from the Web Of Knowledge scientific literature database that is related to one of these six topics, starting from year 1990. These datasets contain the abstract i, its year of publication t_i, and the number of times the article was cited in the Web Of Knowledge database, tc_i. In the following we will refer to the abstracts as documents $d(i)$, where i labels all abstracts published at time t. We excluded publications from 2010 and later to allow an observation window of at least five years for the publications to gather citations. The analysis is carried

[2] http://apps.webofknowledge.com/.

out separately for each of the six topics. The search retrieved 27,459 results for aging of materials, 2527 for brain–computer interfaces, 2901 for fracking, 3655 for graphene, 3510 for liquid natural gas, and 24,020 for nanotechnology.

Term–document matrix

After removing punctuation, each document is split into individual words and the Porter stemming algorithm is applied to the lower case of each word (Porter 1980). We then apply two filters to identify words that characterize the topics of the abstract. First, each word that is ranked as one of the 5000 most frequent words in the corpus of all New York Times issues (Dodds et al. 2011) is removed. In a second step we remove all words that appear only once in the entire corpus. The first step filters out high frequency words that are not specific for scientific publications, the second step gets rid of highly specialized terms that are not relevant for the vast majority of documents. With the remaining words we construct for each year t the term–document matrix $M(t)$ as

$$M_{wi}(t) = \begin{cases} 1 & \text{if word } w \text{ appears in document } d(i) \text{ in year } t, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The term–document matrix $M(t)$ corresponds to a bipartite network, which is a network with two types of nodes where links always connect nodes of different type. One type corresponds to documents, the other type is given by words that are contained in the given documents. A visualization of such a bipartite network is shown in Fig. 1. In Fig. 1 documents are shown as large, blue nodes whereas words correspond to grey nodes.

Recursive bipartite centrality measures

We will now define recursive centrality measures that capture how central or how peripheral the documents are positioned in the bipartite term–document network. Such measures for bipartite networks have been successfully used to show how the export basket of a country is related to its economic growth (Hidalgo and Hausmann 2009), or how instrumentations of a music style are related to its numbers of album sales (Percino et al. 2014). We start by considering each word that is contained in a given document, consider for example the first document $d(1)$ in Fig. 2. From each of the words linked to $d(1)$ we can reach all documents that contain a word that is also contained in $d(1)$ by following its links. If we iterate this procedure two times we reach all the documents that contain a word that is also contained in a document that shares some words with $d(1)$, see the iterative scheme in Fig. 2. The idea is to measure for each document how many paths there are to reach each other document. The higher this number, the more central and ubiquitous are the topics described by the document. If this number is smaller, however, this means that the document contains very specialized terms that are only relevant for a comparably small number of other documents, that is, the document has a high degree of specificity. More formally we recursively define two vectors k and l as

$$\begin{aligned} k_i(n,t) &= \frac{1}{k_i(0,t)} \sum_w M_{wi}(t) l_w(n-1,t), \\ l_w(n,t) &= \frac{1}{l_w(0,t)} \sum_i M_{wi}(t) k_i(n-1,t), \end{aligned} \quad (2)$$

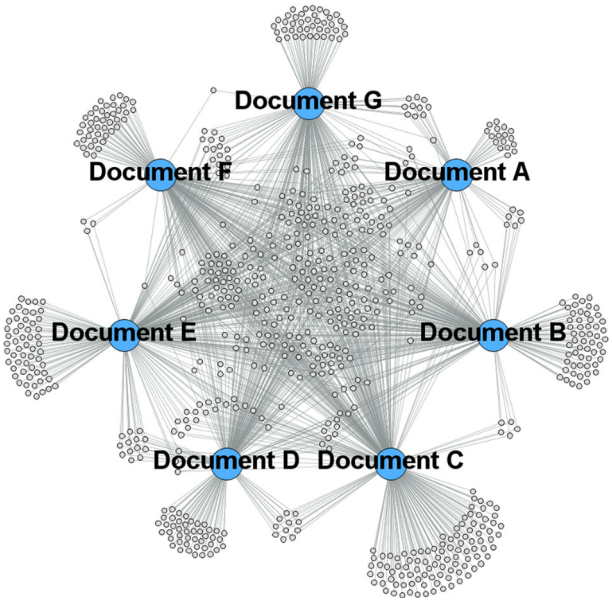

Fig. 1 Visualization of a generic term–document matrix as a bipartite network. One type of nodes corresponds to the documents (*blue nodes*), the other type to words (*grey nodes*). A *link* indicates that a word is contained within the given documents. (Color figure online)

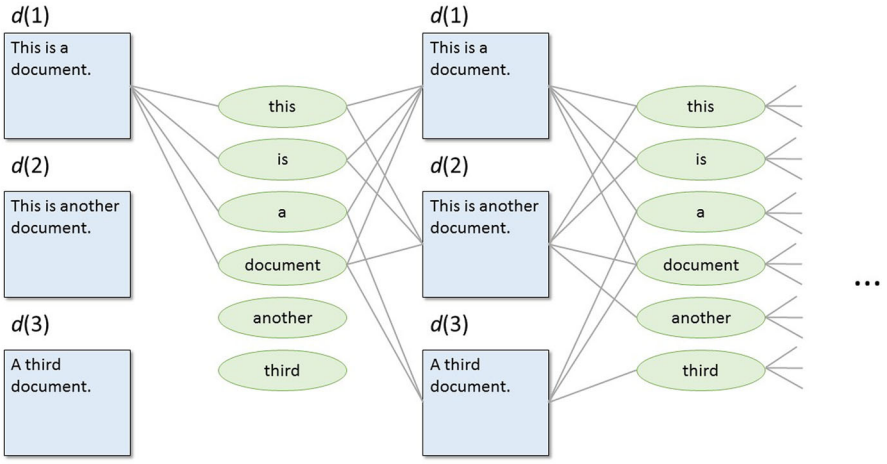

Fig. 2 Recursive method to compute centrality measures on bipartite networks. Starting at a given document, $d(1)$, the first iteration counts the number of documents that contain a term that is also contained in $d(1)$. This procedure is then iterated

with n counting the number of iterations as described in Fig. 2 and the initial conditions $k_i(0,t) = \sum_w M_{wi}(t)$ and $l_w(0,t) = \sum_i M_{wi}(t)$. High values of $k_i(n,t)$ indicate high centrality of the given documents. If the documents are ranked according to their $k_i(n,t)$ values, this ranking typically does not depend on n if n is chosen high enough, see (Hidalgo and Hausmann 2009) for a discussion of convergence properties of the measures in Eq. (2).

In the following we choose to work with $n = 20$ for which we have numerically confirmed that the iteration in Eq. (2) approaches a stationary regime. The values of $k_i(n,t)$ for even and odd values of n lend themselves to different interpretations. $k_i(0,t)$ is a degree centrality that measures the number of out-going links for each document for a given year. High values of $k_i(1,t)$ correspond to documents that contain a large number of keywords that appear in a high number of other documents. Low values of $k_i(1,t)$ indicate that the document contains highly specific terms that are used in few other documents. Values of $k_i(n,t)$ for $n = 2, 4, 6, \ldots$ assign weights to individual keywords based on the number of documents in which the given keyword occurs. The same observations hold for $l_w(n,t)$ with reversed roles for terms and documents.

Defining document centrality

We consider the Z-transforms of the logarithmic recursive bipartite centrality measures,

$$z_i(n,t) = \frac{\log(k_i(n,t)) - \mu[\log(k_i(n,t))]}{\sigma[\log(k_i(n,t))]}, \tag{3}$$

where $\mu[\cdot]$ and $\sigma[\cdot]$ denote mean and standard deviation, respectively. We define the document centrality, $C_i(t)$, through a linear combination of $z_i(0,t)$ and $z_i(20,t)$,

$$C_i(t) = z_i(n = 0, t) + z_i(n = 20, t). \tag{4}$$

A MatLab implementation of an algorithm that computes $C_i(t)$ from a term–document matrix is accompanying this article as electronic supplementary material.

Relation to other centrality measures

We compare the performance of the centrality measures in Eq. (2) to findings from well-known centrality measures, such as eigenvector centrality, Katz prestige, or PageRank. The main rational underlying these measures, which we also adopt for the bipartite recursive centrality measures, is that a node in a network is central if it is connected to nodes that are also central. We consider two different types of unipartite networks that can be derived from the term–document matrix $M(t)$. First, the document–document network $A(t)$ can be obtained as $A(t) = \sum_w M(t)^T M(t)$. The entries $A_{ij}(t)$ are the number of terms that co-occur in documents $d(i)$ and $d(j)$. Note that the bipartite centrality measures for documents in Eq. (2) are also related to the numbers of co-occurring terms between two documents. However, in Eq. (2) these co-occurrences are weighted by the recursively defined centralities of the terms themselves. In brief, unipartite centrality measures for $A(t)$ depend on the raw numbers of co-occurrences of terms, whereas $k_i(n,t)$ is also sensitive to how central the co-occuring terms are. A second unipartite network can be obtained from $M(t)$ by disregarding its bipartite structure. If $M(t)$ contains $D(t)$ different documents and $W(t)$ different terms, the resulting network has $D(t) + W(t)$ nodes. This leads to a network given by the adjacency matrix $B(t)$ as

$$B(t) = \begin{pmatrix} 0_{W(t)\times W(t)} & M(t) \\ M(t)^T & 0_{D(t)\times D(t)} \end{pmatrix}, \tag{5}$$

where $0_{N\times N}$ denotes an N-by-N matrix with all entries being zero and the resulting $B(t)$ being of dimensions $(D(t) + W(t) \times D(t) + W(t))$.

Scientific impact

We measure the citation impact, $x_i(t)$, of an abstract published in year t as its logarithmic number of citations, re-scaled to lie in the range [0, 1]. This re-scaling is done in each year separately in order to account for the different time spans that the publications have had to gather citations and to eliminate potential age effects or the so-called first-mover advantage, i.e. the effect that the first publications in a field receive a disproportionate amount of citations (Newman 2009). Let tc_i be the raw number of citations of document i, and $tc_{\max}(t)$ the highest number of citations for a document published in year t. The citation impact, $x_i(t)$, is then given by

$$x_i(t) = \frac{\log(tc_i + 1)}{\log(tc_{\max}(t) + 1)}, \tag{6}$$

where we have added one to the number of citations in order to ensure that $x_i(t)$ is also defined for $tc_i = 0$.

Regression model

A linear regression model is built where the citation impact $x_i(t)$ is fitted for each document using the presence or absence of each of the keywords as variables. In order to restrict the number of variables we only include those terms that appear in a sufficient number of documents. That is, we compute the frequency of each term (over all years t) and only use the term as a predictor variable if its frequency scores above the 90th percentile of all frequencies (though we have confirmed that our findings do not depend on the concrete choice of this percentile). As response variable in the regression we use the citation impact, $x_i(t)$, which is then fitted using a linear regression model. The regression model gives the fitted citation impact, $\tilde{x}_i(t)$, which is obtained from a linear model of the type $\tilde{x}_i(t) = \sum_w \alpha_{wi} M_{wi}(t)$, with coefficients α_{wi} that are different from zero whenever the null hypothesis that the true coefficient value is zero can be rejected with a p value of $p < 0.1$. While the generalized diversity measures $k_i(n,t)$ and $l_w(n,t)$ are computed without any information from years after t, the regression model explicitly uses the citation impact and therefore a posteriori knowledge.

Results and discussion

In Fig. 3a we show the number of documents $D(t)$ that has been retrieved for each of the six topics for each year. Each of the topics has an increasing trend in the number of publications by year. For nanotechnology and graphene we find a substantially faster increase in numbers of publications over time when compared to the other topics. We also note that nanotech and aging have in total a much higher number of publications than the remaining topics.

Figure 3b shows results for the averaged keyword diversity $k_i(n = 0, t)_i$, where the average is taken over all documents i published in year t. Figure 3c shows results for the average centrality $k_i(n = 20, t)_i$. The dots in Fig. 3b, c show the median value of the measures and the error bars show the first and third quartile. It is interesting to see that the results for the bipartite centrality measures are to a large extend independent of the year, although the numbers of publications show large changes over the observed time. This shows that the bipartite centrality measures do not depend on the size of the term–

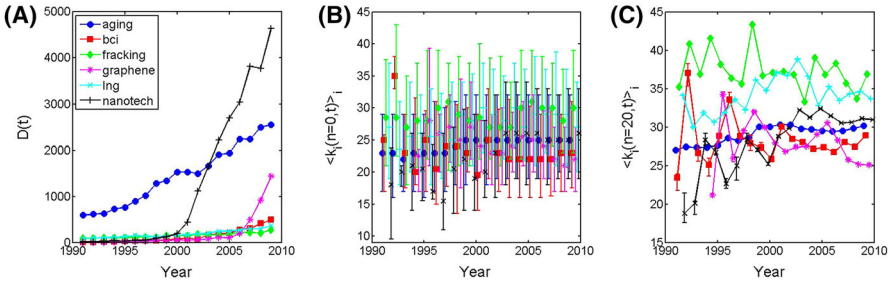

Fig. 3 **a** The number of documents $D(t)$ shows different trends over time for the six topics. There was only little interest in 'nanotech' (*black*) before 2000, but the number of publications quickly ramped up afterwards to almost 5000 publications per year. 'fracking' (*green*) and 'aging' (*blue*) show a steady increase in number of documents over time, with substantially more results for 'aging' than for 'fracking'. **b** The median, shown with first and third quartiles, of the unscaled values for k_i ($n = 0$, t) for each year and topic fluctuates around a value of three. **c** There is a decreasing trend in the median values of the unscaled k_i ($n = 20$, t) values, that is proportional to the increase in number of documents for each topic as shown in **a**. (Color figure online)

document matrix. Some of the different topics, however, show different levels of the centrality measures, compare for instance the results for "fracking" with those for "aging". These differences might suggest that some of the topics typically exhibit a larger or more specialized vocabulary (in the sense of a smaller number of high-frequency keywords) in their publications, which in turn could be related to different degrees of specialization within the different topics.

We form a predictor by taking all documents with values of $z(n, t)$ that fall in a specific range (using forty equidistant bins over the range $[-3.5, 3.5]$), and compare this number to the citation impact $x_i(t)$ within the given set of articles. In the following we refer to these binned entities by dropping the time dependence of the corresponding variables. Note that the term–document matrix M, and therefore the bipartite centrality measures, are computed by using only information from the articles' years of publication. We included a year in this analysis only if we can compute the generalized diversities for at least 200 publications for a given topic and exclude binned data points that correspond to less than three documents to suppress noise in the binning procedure. The values of 200 publications and forty bins have been chosen to minimize noise in the results, however, we confirmed the results do not change qualitatively for a wide range of different choices.

In Table 1 we show the Pearson correlation coefficient between citation impact and the variables $z_i(n = 0)$ and $z_i(n = 20)$. We see that for most of the topics we find a correlation that is significantly greater than zero, with some exceptions however. The prediction for graphene using $z_i(n = 0)$ does only give results of low significance, similarly the prediction for brain–computer-interfaces using $z_i(n = 20)$ does not produce significant correlations. We therefore employ a combination of these two variables, document centrality C_i. We find significant correlations between citation impact and document centrality for each of the topics, see Table 1. The results are compared to alternative definitions of document centrality where we replace $z_i(n = 20)$ in Eq. (4) by traditional (unipartite) centrality measures, see (Newman 2010). We show results where $z_i(n = 20)$ is replaced by the Z-transforms of the logarithms of the eigenvector centrality, C_i^{EC}, Katz prestige, C_i^{Katz}, or PageRank, C_i^{PR}, of the document–document networks $A(t)$. Similarly, results for the unipartite network $B(t)$ are obtained by computing the centrality measures for the full network and then considering only nodes that correspond to documents. This gives the eigenvector

Table 1 Pearson correlation coefficient between various centrality measures and the received citations of publications for six different emerging technologies

Pearson correlation coefficient ρ	$z_i(0)$	$z_i(20)$	C_i	C_i^{EC}	C_i^{Katz}	C_i^{PR}	D_i^{EC}	D_i^{Katz}	D_i^{PR}
Aging	**0.88*** **	**0.62*** **	**0.97*** **	**0.69** **	**0.94*** **	**0.94*** **	**0.70*** **	**0.82*** **	**0.43**
Brain–computer-interface	**0.94*** **	0.42	**0.78*** **	**0.63***	**0.77*** **	**0.86*** **	**0.57***	**0.90*** **	**0.82*** **
Fracking	**0.89*** **	**0.58***	**0.93*** **	**0.54***	**0.67** **	**0.64** **	**0.65** **	**0.89*** **	**0.71*** **
Graphene	**0.54***	**0.72*** **	**0.65*** **	**0.69*** **	**0.75*** **	**0.64** **	**0.83*** **	**0.47** **	0.30
Liquid natural gas	**0.75*** **	**0.70** **	**0.86*** **	**0.81*** **	**0.78*** **	**0.84*** **	**0.79*** **	**0.92*** **	**0.66*** **
Nanotechnology	**0.94*** **	**0.86*** **	**0.97*** **	**0.72*** **	**0.93*** **	**0.98*** **	**0.81*** **	**0.88*** **	**0.68*** **
Mean	**0.82**	**0.64**	**0.86**	**0.68**	**0.81**	**0.82**	**0.73**	**0.81**	**0.60**

Values that are significantly greater than zero ($p < 0.05$) are shown in bold, we further denote $p < 0.01$ with (*), $p < 0.001$ with (**), and $p < 0.0001$ with (***). For the bipartite centrality measures $z_i(0)$ and $z_i(20)$ we find mixed results, where most of the topics show a significant relation between received citations and centrality measures. By combining these measures, i.e. by using the document centrality $z_i(0) + z_i(20)$ as predictor variable, we find strongly significant relations ($p < 0.0001$) between the numbers of citations and the predictor for each of the studied topics. Results are compared to findings using traditional centrality measures applied to two different unipartite networks, the document–document network $A(t)$ and the adjacency matrix of the term–document matrix $B(t)$. Eigenvector centrality, C_i^{EC} and D_i^{EC}, Katz prestige, C_i^{Katz} and D_i^{Katz}, and PageRank, C_i^{PR} and D_i^{PR}, are shown for $A(t)$ and $B(t)$, respectively. By considering the average correlation coefficient over all case studies, last rows, we confirm that the overall best performs is found for the document centrality, C_i

centrality, D_i^{EC}, Katz prestige, D_i^{Katz}, and PageRank, D_i^{PR}, for documents in the networks $B(t)$. Indeed, the document centrality measure, C_i, shows the best overall performance as measured by the Pearson correlation coefficient with citation impact, averaged over all case studies, see the last row in Table 1. The better performance of document centrality when compared to centralities obtained from the document–document networks $A(t)$ shows that there is indeed some crucial information lost by projecting the term–document matrix onto such a unipartite network, as the entries in $A(t)$ only depend on the raw number of term co-occurrences between two documents. The same cannot be said for the networks $B(t)$. Observe that there is a structural similarity between the computation of the recursive bipartite centrality measures in Eq. (2) and the definition of PageRank for $B(t)$ (assuming that a damping factor of zero is used for the PageRank, see 48), with the main difference being how the contributions of individual nodes are normalized by their degree. In brief, the contributions of a given term to the PageRank of a document are normalized by the term's degree, whereas in Eq. (2) they are normalized by the degree of the document itself.

The results are visualized in Fig. 4. For each topic we show the binned document centrality and the corresponding average citation impact. Error bars denote the standard error over all data points within the bin. Note that Fig. 4 shows that the correlation between document centrality and citation impact extends over the entire range of centrality values (i.e. x-axis) in five out of six case studies. Only in the "brain–computer interface" case study we see that the relation is mostly driven by noise for $C_i > 0$. In all other cases we see that the positive correlation between document centrality and citation impact extends to the top-ranked documents and is therefore not driven by potential spurious correlations between low-ranked documents and small degree.

The five publications with the highest values of document centrality, C_i, are shown in Table 2. Note that the distributions of citation impact in the case studies are skewed

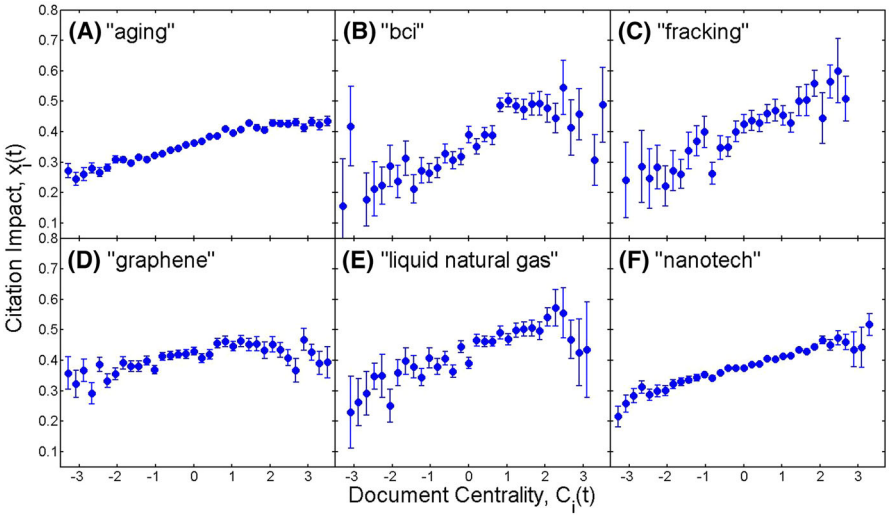

Fig. 4 The abstracts of publications on **a** aging, **b** brain–computer interfaces, **c** fracking, **d** graphene, **e** liquid natural gas, and **f** nanotechnology are grouped according to their document centrality for each year. For each group we compute the average scientific impact as the re-scaled average number of citations. There is a clear trend that higher scientific impact as measured by the number of citations is positively correlated with high generalized keyword diversities, in each of the six independent corpora. The *error bars* show the standard error of the mean in each group

Table 2 For each case study we rank the publications according to their value of the document centrality and show the top 5 articles

Rank	Aging	Brain–computer-interface	Fracking	Graphene	Liquid natural gas	Nanotech
1	Fang et al., Progress in Natural Science 2009; 19(7): 851–859 [$x_i = 0.14$]	Hawkins, Davis, Pharmacological Reviews 2005; 57(2): 173–185 [$x_i = 0.46$]	Bucher et al., J Petrology 2006; 47(3): 567–93 [$x_i = 0.52$]	Freitag, Merkle, J Comput Theor Nanosci 2008; 5(5): 760–861 [$x_i = 0.35$]	Coopman et al., Forensic Sci Int 2009; 189(11–13): e13–e20. [$x_i = 0.45$]	Lertsapcharoen et al., Indian Heart J 2006; 58(4): 315–20 [$x_i = 0.23$]
2	Yang et al., Progress in Natural Science 2003; 13(11): 861–866 [$x_i = 0.34$]	Grudzinska et al., Neuron 2005; 45(5): 727–739 [$x_i = 0.86$]	Möller et al., Int J Earth Sci 2007; 96(3): 541–566 [$x_i = 0.61$]	Bucknum, Casto, Solid State Sciences 2008; 10(9): 1245–1251 [$x_i = 0.24$]	Frieri et al., Digestion 2006; 73: 58–66 [$x_i = 0.00$]	Schattenburg, J Vac Sci Technol B 2001; 2319 [$x_i = 0.36$]
3	Tepe et al., Invest Radiol 1996; 31(4): 223–239 [$x_i = 0.65$]	Nashmi et al., Biochem Pharmacol 2007; 74(8): 1145–1154 [$x_i = 0.55$]	Gargini et al., Hydrogeology 2006; 125(3): 293–327 [$x_i = 0.35$]	Ramanathan et al., Nature Nanotechnology 2008; 3: 327–331 [$x_i = 0.87$]	MacLean, Lave, PECS 2003; 29: 1–69 [$x_i = 0.39$]	Steinberg et al., Proteomics 2001; 1(7): 841–855. [$x_i = 0.71$]
4	Baker-Cairns, et al., Exp Neurol 1996; 142(1): 36–46 [$x_i = 0.63$]	Kundrát, Naturwissenschaften 2007; 94(6): 499–504 [$x_i = 0.50$]	Dasgupta, Sedimentary Geology 2006; 185(1–2): 59–78 [$x_i = 0.50$]	Lambert et al., J Phys Chem C 2009; 113(46): 19812–19823 [$x_i = 0.67$]	Laherrère, Int J Vehicle Design 2004; 35(1–2): 9–26. [$x_i = 0.83$]	Heller, Neurosurgery 2008; 62(6): 921–940 [$x_i = 0.43$]
5	Abeloa-Formanek et al., J Cataract Refract Surg 2002; 28(1): 50–61. [$x_i = 0.68$]	Rickmane et al., Mol Biol Cell 2006; 17(1): 283–294 [$x_i = 0.55$]	Martin-Izard et al., J of Geochemical Exploration 2009; 100: 51–66 [$x_i = 0.31$]	Petit et al., J Mater Chem 2009; 19: 9176–9185 [$x_i = 0.56$]	Valianou et al., Anal Bioanal Chem 2009; 395(7): 2175–2189. [$x_i = 0.29$]	Moses, Am J Cardiovasc Drugs 2002; 2(3): 163–172 [$x_i = 0.37$]

This illustrates the statistical nature of the relation between document centrality and citation impact. The citation impact is given in rectangular brackets

towards smaller values, with medians ranging from 0.37 to 0.46, 75th quantiles ranging from 0.50 to 0.59, and 95th quantiles range from 0.68 to 0.80. This means that many of the top ranked documents in Table 2 have indeed also highly ranked citation impacts. However, the statistical nature of the relation between citation impact and document centrality is illustrated by some documents with very low citation impact that also occur in these lists.

It is further instructive to compare results of the document centrality with the fitted citation impact, $\tilde{x}_i(t)$. A direct comparison of their average values, together with the standard error of the fitted citation impact for a given value of document centrality, is given in Fig. 5. While still significant, we find substantially lower correlations for the "brain–computer-interface" case study than for the remaining case studies. The second lowest correlation is found for "fracking". It is intriguing to see that these are also the two case studies where the bipartite centrality measures, $z_i(n = 20)$, lead to considerably poorer results as compared to the results, $z_i(n = 0)$, compare Table 1. Note that $z_i(n = 0)$ only measures the number of keywords without containing any information on the centrality of those keywords. From this one might conclude that in these cases the mere appearance of certain concepts, topics, or keywords is of greater importance than with which terms they appear together with. Nevertheless, there is a high statistical significance of the correlation between document centrality and fitted citation impact in each of the case studies. This finding further corroborates that the measure of document centrality helps to identify those articles that contain a large number of keywords that lead to a comparably larger citation impact.

What are the words that publications with a high number of citations typically contain? To study this question one can exploit the symmetry between terms and documents in the bipartite centrality measures of Eq. (2). This allows us to define the term centrality, C'_w,

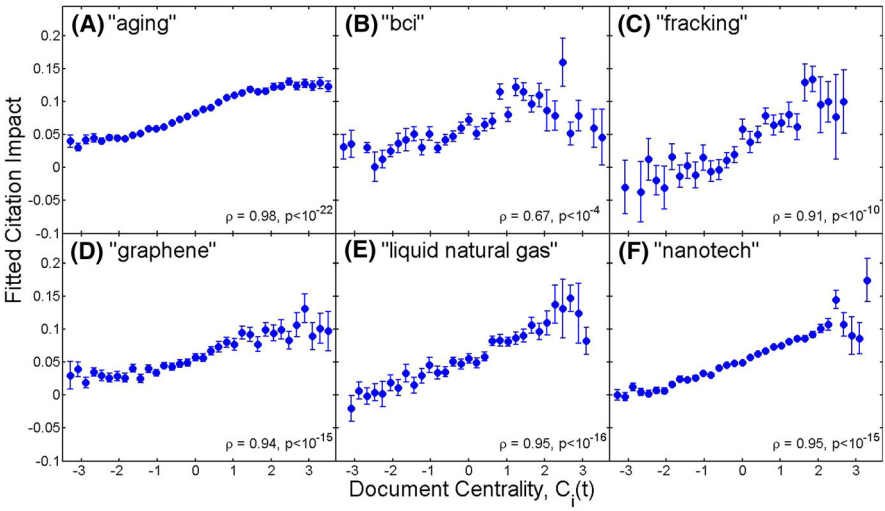

Fig. 5 Comparison of results for document centrality, C_i, and the fitted citation impact for the six case studies, **a** aging, **b** brain–computer interfaces, **c** fracking, **d** graphene, **e** liquid natural gas, and **f** nanotechnology. We also show results for the Pearson correlation coefficient and the p value to reject the null hypothesis that the true coefficient value is zero. There is a highly significant correlation between results of the regression model and document centrality in each of the case studies, which shows that the network based method identifies those articles that contain keywords that in turn correlate with a high citation impact

Table 3 Results for top 5 terms for each case study as ranked by their term-centrality measure, C'_w, (values given in the brackets), i.e. a formulation of C_i where terms and document exchange roles

Rank	Aging	Brain–computer-interface	Fracking	Graphene	Liquid natural gas	Nanotech
1	Temperature (5.1)	Interface (5.0)	Fractures (5.2)	Graphene (3.9)	Liquid (5.0)	Helmus (18)
2	Alloy (4.8)	Brain–computer (3.6)	Hydraulic (5.1)	Carbon (3.8)	Samples 3.6	Nanotechnology (4.3)
3	Mechanism (4.5)	Function (2.9)	Fluids (3.3)	Nanotubes (3.2)	Phase 3.4	Nanoparticles (3.3)
4	Phase (4.4)	Classifier (2.8)	Simulation (3.0)	Graphite (2.5)	Concentration 3.4	Molecular (3.1)
5	Precipitation (4.3)	Accuracy (2.7)	Permeability (2.9)	Layers (2.5)	Temperature (3.3)	Characterization (3.0)

that can be obtained analogously to the document centrality, C_i, by exchanging the roles of terms and documents in Eqs. (2–4). Results for C'_w for the top 5 terms in the six different case studies are given in Table 3. Terms that are closely related to the topic itself feature prominently in these lists, e.g. nanotechnology and nanoparticles for "nanotech*". The other terms hint at several related issues that lie at the core of the particular topics, such as the controversy surrounding the permeability of rock with respect to fluids used in hydraulic fracturing. The results also contain more specific terms, such as the name of author Michael N. Helmus who wrote a series of thesis articles for Nature Nanotechnology on the commercialization of nanotechnology, with his name appearing in the abstract in each case.

Conclusions

To summarize, in this work we have introduced a novel and exclusively content-based method for ranking documents according to estimates of their potential impact. The method is based on a bipartite network representation of the term–document matrix for a given corpus of documents. We compute bipartite centrality measures that adapt the conventional recursive centrality measures on networks (such as Katz prestige, eigenvector centrality, or Google's PageRank) to the bipartite situation. That is, a document is regarded as "central" if it contains a large number of terms that commonly appear together with other terms, which in turn appear in a large number of other documents. The recursive nature of this measure can be expressed as the fact that a document is "central" if it is linked by central terms to other documents that are also regarded as central. This is a crucial difference to conventional centrality measures that typically do not depend on which kind of terms overlap between two documents. We constructed a measure, document centrality, which combined two indicators: The first indicator measures the number of different terms in a given document, i.e. it can be interpreted as the degree (number of links) of a document in the bipartite network of the term–document matrix. While this indicator is in a sense "blind" to which terms co-appear in the document, this is not true for the second indicator that can be interpreted as a recursive centrality measure for bipartite networks. We demonstrated the ability of document centrality to predict the citation impact of scientific publications in six case studies from the field of material science. These case studies covered, both, fields that show rapidly increasing attention and interest (e.g. "graphene" or "fracking") and also field that show a substantially slower growth in the number of publications (e.g. aging). In each of these case studies the values of document centrality showed a strongly significant correlation with the citation impact of the publications. Note that the term–document matrices, and thereby the document centralities, were computed using only knowledge from the year of publication. That is, we can exclude the possibility that there is any cross-talk between the computation of the indicators and the citation impact that we want to predict using them. We found that while both indicators, $z_i(n = 0)$ and $z_i(n = 20)$, are often significant predictive indicators for citation impact themselves, the best performance is found for a combination of them. These findings were further substantiated by a comparison of the values for document centrality with results from traditional, unipartite centrality measures and from a linear regression model where we explicitly fit citation impact of documents using the presence or absence of terms as variables (hence the regression model has no predictive value whatsoever and is prone to overfitting the data). However, we find a strongly significant correlation between results of

this regression model and document centrality, which shows that the network-based indicators correctly identify those articles that tend to contain terms associated with large citation impact.

Limitations of our current approach include that it is based on a bag-of-word representation of text and does not take phrases, lemmatization, or n-grams of words into account. It will be interesting to explore to which extend the performance of the document centrality measure can be further improved by using such more refined text representations. In the current work we have only used the terms that appear in the abstract of a publication, but not the full text of the article. While the abstract is certainly an extremely relevant description of the full text of the article, it remains to be seen to which extend our findings would apply to the full articles. Another interesting extension of our work could lie in an ontological annotation of the terms or phrases to study potential differences due to semantic or syntactic information (consider, for example, the extremely high term centrality found for a person in the nanotechnology case study). This also includes potential differences due to works that do not adhere to standards of the particular field, as it has recently been shown that the style of writing impacts the success of published articles too (Moohebat et al. 2015).

It is worth stressing that the methodology of this work can be applied to a wide range of text corpora, not necessarily scientific publications. One of the main motivations of this work is indeed the often encountered problem of identifying the most relevant item from a large set of unstructured documents. The case studies using scientific publications therefore have the appeal of giving a (more or less) objective way to measure impact of documents, namely in terms of received citations. The document centrality measure might therefore have applications as a quantitative tool to aid the steering of research funding and to qualify the potential of research proposals by, e.g., funding agencies or the host institutions. Alternatively one might recognize that low values of the bipartite centrality measures can be interpreted as indicators for high originality of a work (that is, articles with low values exhibit substantial deviations in their usage of terms from the mainstream represented by the other, published research papers in this field). In this sense papers are highly original if they offer a different perspective or novel context to an already established topic. Our findings that $z_i(n = 20)$ has a positive correlation with citation impact has then a clear interpretation: Highly original research papers tend to be punished in terms of their received numbers of citations.

Acknowledgments PK acknowledges financial support from the European Commission, EU FP7 Project MULTIPLEX, No. 317532. We thank the anonymous referees for providing extremely helpful comments and suggestions.

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*(3), 175–185.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bollen, J., Rodriguez, M. A., & Van De Sompel, H. (2006). Journal status. *Scientometrics, 69*(3), 669–687.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 30 scientific impact measures. *PLoS One, 4*(6), e6022.

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.

Callaham, M., Wears, R. L., & Weber, E. (2002). Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *Journal of the American Medical Association, 287*(21), 2847–2850.

Chang, J., & Blei, D. M. (2009). Relational topic models for document networks. In *Proceedings of the 12th international conference on artificial intelligence and statistics (AISTATS)* (Vol. 5, pp. 81–88).

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics, 1*(1), 8–15.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

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.

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.

Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on machine learning* (pp. 233–240).

Dodds, P. S., Harris, K., Kloumann, I., Bliss, C., & Danforth, C. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE, 6*(12), e26752.

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.

Feng, G., Guo, J., Jing, B.-Y., & Hao, L. (2011). A Bayesian feature selection paradigm for text classification. *Information Processing and Management, 48*(2), 283–302.

Fu, L. D., & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics, 85*, 257–270.

Garfield, E. (1979). *Citation indexing: Its theory and application in science, technology, and humanities*. Ney York: Wiley.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Hidalgo, C. A., & Hausmann, R. (2009). Building blocks of economic complexity. *Proceedings of the National Academy of Sciences, 106*(26), 10570–10575.

Hofmann, T. (2001). Unsupervised learning by probabilistic semantic analysis. *Machine Learning, 42*, 177–196.

Jian, L., Cai, Z., Wang, D., & Zhang, H. (2014). Bayesian citation-KNN with distance weighting. *International Journal of Machine Learning and Cybernetics, 5*(2), 193–199.

Jovanovic, A. S., & Renn, O. (2013). Search for the 'European way' of taming the risks of new technologies: The EU research project iNTeg-Risk. *Journal of Risk Research, 16*(3–4), 271–274.

Kwok, J.T.-Y. (1998). Automated text categorization using support vector machine. In *Proceedings of the international conference on neural information processing (ICONIP)* (pp. 347–351).

Larsen, P. O., & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics, 84*(3), 575–603.

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Techology, 58*(9), 1303–1319.

Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the American Society for Information Science and Technology, 60*(7), 1327–1336.

Leydesdorff, L., & Bornmann, L. (2011). Integrated impact indicators (I3) compared with impact factors (Ifs): An alternative design with policy implications. *Journal of the American Society for Information Science and Technology, 62*(7), 1370–1381.

Liu, Y., Niculescu-Mizil, A., & Gryc, W. (2009). Topic-link LDA: Joint models of topic and author community. In *Proceedings of the 26th annual international conference on machine learning (ICML)* (pp. 665–72).

Liu, L. G., Xuan, Z. G., Dang, Z. Y., Guo, Q., & Wang, Z. T. (2007). Weighted network properties of Chines nature science basic research. *Physica A, 377*(1), 302–314.

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics, 36*(3), 435–444.

Meyer, D., Leisch, F., & Hornik, K. (2003). The support vector machine under test. *Neurocomputing, 55*(1), 169–186.

Moohebat, M., Raj, R. G., Kareem, S. B. A., & Thorleuchter, D. (2015). Identifying ISI-indexed articles by their lexical usage: A text analysis approach. *Journal of the Association for Information Science and Technology, 66*(3), 501–511.

Nallapati, R., Ahmed, A., Xing, E., & Cohen, W. W. (2008). Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD)* (pp. 542–550).

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences, 101*(1), 5200–5205.
Newman, M. E. J. (2009). The first-mover advantage in scientific publication. *Europhysics Letters, 86*(6), 68001.
Newman, M. E. J. (2010). *Networks: An introduction*. Oxford: Oxford University Press.
Percino, G., Klimek, P., & Thurner, S. (2014). Instrumentational complexity of music genres and why simplicity sells. *PLoS ONE, 9*, e115255.
Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.
Radicchi, F., & Castellano, C. (2012). Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. *Journal of Informetrics, 6*(1), 121–130.
Radicchi, F., Fortunato, S., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, 056103.
Sayyadi, H., & Getoor, L. (2009). FutureRank: Ranking scientific articles by predicting their future PageRank. In *The 9th SIAM international conference on data mining*.
Stewart, J. A. (1983). Achievement and ascriptive processes in the recognition of scientific articles. *Social Forces, 62*(1), 166–189.
Van Dalen, H. P., & Henkens, K. (2001). What makes a scientific article influential? *The case of demographers. Scientometrics, 50*(3), 455–482.
Vieira, E. S., & Gomes, J. A. N. F. (2010). Citation to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics, 4*(1), 1–13.
Walker, D., Xie, H., Yan, K. K., & Maslov, S. (2007). Ranking scientific publications using a simple model of network traffic. *Journal of Statistical Mechanics*, P06010. doi:10.1088/1742-5468/2007/06/P06010.
Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science, 342*(6154), 127–132.
Yan, E., Ding, Y., & Sugimoto, C. R. (2011). P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology, 62*(3), 467–477.
Yu, X., Gu, Q., Zhou, M., & Han, J. (2012). Citation prediction in heterogeneous bibliographic networks. In *SDM* (Vol. 12, pp. 1119–1130).
Yu, T., Yu, G., Li, P.-Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics, 101*, 1233–1252.