Measurement of Scholarly Impact: a comparative analysis based on an educational academic journal

Yi Chen, Ameya Karnad, Hui Soo Chae, Gary Natriello
Teachers College Columbia University

Abstract
Identifying influential articles in a corpus is an important and challenging problem in educational research. Scholarly impact has typically been equated with the citation count. However, these methods do not pay enough attention to the context of articles. Continuous improvements in text mining techniques present new opportunities for measuring scholarly impact beyond citation-analysis. But few studies have been done applying these methods in educational research with real world data. In this paper, we introduce an innovative approach to measuring educational scholarly impact based on text mining. About 2000 articles from the Teachers College Record from 1995 to 2016 have been analyzed. We also discuss the potential and the challenges of applying text mining in bibliometrics.

Keywords: Test Mining, Document Influence Model, Negative Binomial Regression, Teachers College Record

Introduction
The growing flood of scholarly literature has exposed the weakness of current citation-based methods for evaluating and filtering articles (Priem & Hemminger, 2010). Today, scholars read 50 percent more papers than they did in the 1970s and consequently are spending less time on average with each one. Moreover, in decisions regarding grant funding and priority in assigning resources, "article overload" besets evaluators and the bibliometrics measurements they employ (Boyack & Jordan, 2011).

Bibliometrics is a systematic method for evaluating research output that can help map changes in the interest of the scientific community over time, and can provide insights into both quantitative and qualitative research trends on a specific topic (Ugolini et al., 2013). Journal impact factor (IF; Garfield, 1999), which is based on citation count, is the most popular way to measure the influence and visibility of a journal, and further help to assess the quality of academic instruments. Despite the popularity of this measurement, it is to some extent slow (Brody, Harnad, & Carr, 2006), narrow (Priem & Hemminger, 2010), secretive as well as irreproducible (Rossner, Van Epps, & Hill, 2008), open to gaming (Falagas & Alexiou, 2008), and based at the journal level rather than the article level (Meho, 2007).
Some investigators have tried to expand citation-based measurements through web-based sources and techniques like social network analysis (Carolan & Natriello, 2005). However, Gerrish and Blei (2010) presented the analysis of scholarly impact directly through the text context with text mining methods. This approach supports a richer picture of an article’s impact. But few research studies have set out a comprehensive and detailed application in education, and discussed the results.

In this study, we use the Document Influence Model (DIM; Gerrish & Blei, 2010) to measure the effect of the influence of articles. We theorize that an influential article will affect how future articles are written. We analyze about 2000 articles appearing over 24 years in the Teachers College Record (TCR), a major outlet in educational research. Using DIM, we identify the most influential article at different periods. Furthermore, we discuss the similarities and differences between text-based and citation-based measurements of article impact.

Literature Review

Databases.
Citation Impact Indicators.
New Methodologies.

Citation Count: strength and weakness. Scholarly impact is one of the strongest currencies in the academy and has traditionally been equated with the number of citations — be it for individuals, articles, departments, universities, journals, or entire fields (Aguinis, Suarez-Gonzalez, Lannelongue, & Joo, 2012). A citation can mean recognition or validation of one’s research by other researchers (D. Wang, Song, & Barabási, 2013). As one of the pioneers of bibliometrics, Eugene Garfield founded the Institute for Scientific Information (ISI), which holds the over-40-year monopoly of citation analysis. Under ISI, the Science Citation Index (SCI; Garfield, 1964) and the Thomson Scientific Journal Impact Factor (JIF; Garfield, 1999) have achieved a dominant position.

Citation analysis is most widely used in ranking academic journals rather than articles. Aggregating articles’ citation count, these databases can be used to estimate a journal’s impact. In addition, citation count influences the ranking of academic institutions in practice. The Academic Ranking of World Universities (ARWU), for instance, relies heavily on the citation-related objective indicators: number of highly cited researchers and number of articles indexed in Science Citation Index (N. C. Liu & Cheng, 2005).

However, it is a sobering fact that 90 percent of papers that have been published in academic journals are never cited and as many as 50 percent of papers are never read by anyone other than their authors, referees and journal editors (Meho, 2007). Even the most widely used ISI’s databases have long been criticized by scholars for the limited number of journal titles. For example, citations from books and most conference proceedings as well as many non-English journal titles are not covered. These limitations in the citation coverage have led to the popularity of web-based sources such as Scopus and Google Scholar as well as new measurements like the 'h-index' (Hirsch, 2010).

Furthermore, whether citation per se is a reliable measurement for scholarly impact is still an open question (Opthof, 1997; Seglen, 1997; Harter & Nisonger, 1997). One concern is self-citation where scholars deliberately cite the article, journal or the academic institutions they are affiliated with. Ceremonial-citation is also a very common phenomenon in the
academy. Authors sometimes cite an authority in the field without ever having consulted the relevant work itself. Another issue is cronyism, whereby friends or colleagues reciprocally cite each other (Meho, 2007). Citation counts typically ignore the fact that review articles and book chapters are heavily cited. To sum up, traditional citation-analysis does not pay enough attention to the context of the article and so treats the citations’ impacts equally. Even negative or positive citations are counted equally.

**New approach for scholarly impact measurement.** The limitations of citation-related measurements have spurred the study of alternative approaches for scholarly impact measurement. In particular, Google’s method of ranking web pages has inspired many measures that apply social network analysis in citation networks (Lazega, Wasserman, & Faust, 2006). Pinski and Narin (1976) first proposed ranking journals according to their eigenvector centrality, which is a measure of the influence of a node, in a citation network. In addition, citation PageRank (Dellavalle, Schilling, Rodriguez, Van de Sompel, & Bollen, 2007), Scimago Journal Rank (SJR, Chen, Xie, Maslov, & Redner, 2007, and betweenness centrality (Leydesdorff, 2007) have been used to measure the impact of a journal. Aguinis et al. (2012) also used the number of pages as indexed by Google to assess scholarly impact on stakeholders outside and inside the academy.

Similar ideas have also been extended to rank individual articles. For example, Article-Level Metrics (ALMs; Sparc & Tananbaum, 2013) seek to incorporate data from social media mentions along with citations to present a richer picture of how an individual article is being discussed, shared, and used. Efforts to examine various citation metrics include Bollen, Van de Sompel, Hagberg, and Chute (2009) who performed a principal component analysis of the rankings produced by 39 existing and proposed measures of scholarly impact which were calculated based on the citation.

While improvements such as these address deficiencies in citation-analysis, the text content of articles is still ignored. With the rapid development of data mining techniques as well as the massive improvement in computation and storage in the past two decades, scholars today can explore large number of articles in detail. Gerrish and Blei (2010) designed a probabilistic model that captures how past articles exhibit varying influence on future articles: DIM. The basic assumption is that an influential article will affect how future articles are written and that this effect can be detected by examining the way corpus statistics change over time.

**Data Set and Method**

**Document Influence Model.** DIM is a topic modeling approach to measure scholarly impact. A general introduction of topic modeling is beyond the scope of this article. D. M. Blei (2011) provide an introduction to topic modeling from a statistical perspective. L. Liu, Tang, Dong, Yao, and Zhou (2016) provided a review of applications of topic modeling techniques. It is based on a dynamic topic model (DTM, C. Wang, Blei, & Heckerman, 2008) and allows for multiple threads of influence within a corpus. The influence of each article is encoded as a hidden variable and posterior inference reveals the influential article of the collection.

For each publication time period $t$, we associate each document $d_t$ with a vector of influence factors $I_{dt,k}$. These factors represent how much the language used in the document
$d_t$ affects the drift of these topics over a period of time. The larger positive value of an influence factor means that the more influential the document $d_t$ is in the topic $k$.

The full generative model at the time period $t$ can be summarized like this:

1. For topic $k = 1, ..., K$, the word distribution over topic $\beta_{k,t}$ is sampled from a normal distribution.

$$\beta_{k,t}|\beta_{k,t-1}, (w, l, z)_{t-1,1:D} \sim N(\beta_{k,t-1} + \exp(-\beta_{k,t-1} \sum_d l_{d,k} \sum_n w_{d,n} z_{d,n,k}), \sigma^2 I)$$

2. For each document $d_t$
   
   (a) generate all documents at time $t$ using LDA with topics $\beta_t$
   
   $$\theta_{d,t}|\alpha \sim \text{Dirichlet}(\alpha)$$
   $$\phi_{k,t,w} = \frac{\exp(\beta_{k,t,w})}{\sum_w \exp(\beta_{k,t,w})}$$
   $$z_{d,i,t}|\theta_{d,t} \sim \text{Categorical}(\theta_{d,t})$$
   $$w_{d,i,t}|\phi_{k,t,w} \sim \text{Categorical}(\phi_{z_{d,i,t}})$$

   (b) For topic $k = 1, ..., k$, influential weights as:

$$I_{d,k,t} \sim N(0, \sigma^2 I)$$

The length of the time period is a priori, which can be specified based on prior knowledge or techniques including cross-validation or a non-parametric approach. A typical time period used in this model is one year (Gerrish & Blei, 2010). However because the typical educational research article takes a long time to publish and even longer to have an impact in the academy we use two year periods in our analysis.

**Negative Binomial Regression.** The citation count in TCR is an over-dispersed count outcome variable. Among 2,700 papers, over 2,500 papers haven’t been cited by another TCR paper ever. In this study, we want to explore the relationship between the citation count and the overall influence of a document. We aggregate the influence value across the topics. The overall influence $O_d$ of a document $d_t$ is calculated as the following formula:

$$o_d = \max(\tilde{l}_{d,t} \circ \tilde{\theta}_{d,t} \circ \sum_d \tilde{\theta}_{d,t+1})$$

Each influence factor is multiplied by its corresponding posterior topic proportion in the current period and marginal posterior topic distribution in the next period. Then, we choose the maximum value of the influence vector as the document overall influence. It indicates how influential a document is based on the text. Taking the overall influence $o_d$ as the explanatory variable and citation count $c_d$ as a response variable, we use the negative binomial regression.
Data

Our data comes from TCR from 1995 to 2016. For dynamic topic modeling, the beginning and end periods have relatively less information than another period during estimation iterations. The model results of the first period (1995-1996) and the last period (2015-2016) would be cut off since they are usually are not stable. But these data are useful in overall model training and parameter estimation. We collected the text context, publication year, and citation-count within TCR for each article. We then conducted the following text data processing operations: word segmentation, punctuation removal, deletion of numbers, transforming the words to lowercase, updating the stop words, removing the stop words, and stemming. After building the corpus, we transformed the data into a document-term matrix and removed the sparse terms. Finally, we derived a $2000 \times 1074$ document-term matrix.

Results

In the application of DIM, the number of topics is usually a modeling choice which we need to specify \textit{a priori}. Including more topics always leads to a better fit at the expense of increased difficulty in interpretation and computation. Following the common practice for specifying the number of topics, we use cross-validation based on the measurement of log likelihood (perplexity is another common choice). We choose eight topics since fitting more than that leads to decreasing improvement gains.

Figure 1 shows the standardized overall influence of each document in a different year. In general, for articles published during the last periods, the overall influence has a relatively small variance and is less likely to have a high value. This phenomenon accords with common sense since it would be difficult for newly published articles to lead to an immediate big impact and huge differences. Recalling the generation process of $\beta_{t+1}$, a
large overall influence means that the marginal topic distribution in the next period would shift towards this document. For every period, several articles have a relatively high overall impact. For more detail, the results from DIM provides the articles’ impact on the eight topics separately.

Table 1 provides more information about the most influential article in each period. In the table, we also provide the citation count based on Google Scholar. There is an obvious positive correlation between the standardized overall influence and Google Scholar citation. To some extent, this proves the reliability of the DIM model in measuring the scholarly impact. As long as the scholarly impact needs to be measured among the articles, we need to include all the documents in the corpus. Unlike Google Scholar, we focus on the TCR articles in our corpus. Thus, it would be more reasonable to use the citation count only within the TCR and see how these measures interact with each other.

<table>
<thead>
<tr>
<th>publish year</th>
<th>paper title</th>
<th>standardized overall influence</th>
<th>Google Scholar Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999-2000</td>
<td>Community Colleges and Contract Training: Content, Origins and Impact</td>
<td>5.21369</td>
<td>93</td>
</tr>
<tr>
<td>2001-2002</td>
<td>&quot;Why Do They Give the Good Classes to Some and Not to Others?&quot; Latino Parent Narratives of Struggle in a College Access Program</td>
<td>5.434379</td>
<td>203</td>
</tr>
<tr>
<td>2007-2008</td>
<td>Refusing to Leave Desegregation Behind: From Graduates of Racially Diverse Schools to the Supreme Court</td>
<td>4.58327</td>
<td>40</td>
</tr>
<tr>
<td>2009-2010</td>
<td>Racial Differences in the Formation of Postsecondary Educational Expectations: A Structural Model</td>
<td>3.531543</td>
<td>41</td>
</tr>
<tr>
<td>2011-2012</td>
<td>A Study of Arizona's Teachers of English Language Learners</td>
<td>2.972088</td>
<td>62</td>
</tr>
<tr>
<td>2013-2014</td>
<td>The Role of Moral and Performance Character Strengths in Predicting Achievement and Conduct Among Urban Middle School Students</td>
<td>1.209532</td>
<td>25</td>
</tr>
</tbody>
</table>
Table 2 summarizes the basic statistical properties of the influence factor for each topic and their correlation with citation count. As we can see, there exists a weak positive correlation between influence factors and citation count. The correlation is smaller than the previous one partly because over 90 percent of TCR articles have not been cited by other TCR articles. However, this result shows the advantage of DIM in handling non-cited articles.

Table 2  
Statistics Summary of Influence Factors and overall influence

<table>
<thead>
<tr>
<th>Measure</th>
<th>mean</th>
<th>Sd</th>
<th>Correlation with citation count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>0.002203</td>
<td>0.0046332</td>
<td>0.010729595</td>
</tr>
<tr>
<td>Topic 2</td>
<td>0.005599</td>
<td>0.0133193</td>
<td>0.059727508</td>
</tr>
<tr>
<td>Topic 3</td>
<td>0.002380</td>
<td>0.0047990</td>
<td>0.044717551</td>
</tr>
<tr>
<td>Topic 4</td>
<td>0.003848</td>
<td>0.0075350</td>
<td>0.004884340</td>
</tr>
<tr>
<td>Topic 5</td>
<td>0.002462</td>
<td>0.0052886</td>
<td>0.008275504</td>
</tr>
<tr>
<td>Topic 6</td>
<td>0.003215</td>
<td>0.0066733</td>
<td>0.033400151</td>
</tr>
<tr>
<td>Topic 7</td>
<td>0.013783</td>
<td>0.0232415</td>
<td>0.058190470</td>
</tr>
<tr>
<td>Topic 8</td>
<td>0.002793</td>
<td>0.0057384</td>
<td>0.021701854</td>
</tr>
<tr>
<td>Overall Influence</td>
<td>0.005087</td>
<td>0.0009423</td>
<td>0.044448355</td>
</tr>
</tbody>
</table>

To handle this unbalanced data, it would be better to use the negative binomial regression to explore the relationship between the overall influence and the citation count. As we can see from Table 3, the parameters in the regression are significant which means the text-based overall influence indeed correlated with the citation count. We can conclude that for a one unit increase in the overall influence, the logs of the expected citation count is expected to change by 0.125.

Table 3  
Negative Binomial Regression Result

| Variable          | Estimate | Std. Error | z Value | Pr(>|z|) |
|-------------------|----------|------------|---------|---------|
| Intercept         | -0.67046 | 0.05077    | -13.206 | <2e-16***|
| Overall Influence | 0.12591  | 0.05345    | 2.356   | 0.0185*  |

The results from both correlation analysis and negative binomial regression show a positive relationship between overall influence and citation count. This indicates that an articles’ text cannot fully explain its citation count. As we discussed before, some other factors may also have some influence on the citation behavior, including ceremonial-citation, self-citation, and cronyism. To sum up, DIM provides a reasonable and somewhat richer picture of the scholarly impact than citation count alone.
Discussion

This study is an example of how we can use text mining methods in the measurement of scholarly impact in education. These techniques empower the researchers in education to analyze text data and explore the stream of hidden patterns. However, more research should be done to address the remaining challenges in the application of text mining in educational research. And more advanced models should be designed for different research problems.

The biggest challenge in the application of DIM lies in the demanding computational requirements. The unknown parameters in the DIM model would increase exponentially if an increasing number of articles are included in the corpus. This is also true when we need to build more complicated models or a bigger corpus. As a consequence, this method may be useful to identify the research trends in a specific field or for some specific journals, but it is still not practical as a citation-based measurement for the whole academy.

To alleviate this problem, several parameters estimation technologies have been designed. For example, the stochastic variational inference has been designed for analyzing very large data sets (D. M. Blei, Kucukelbir, & McAuliffe, 2017). On the one hand, it is an open question as to how much information from an article should be included in the corpus for identifying the articles’ latent characteristics. Instead of incorporating the whole context, we could input only the abstract and keywords to reduce the scale of the dataset. Additionally, we can also remove more sparse terms in the document term matrix and update the stop words specifically for education research.

On the other hand, text mining is still an important amendment and supplement for citation-based measurements. This technique fills in the important details of scholarly impact beyond the citation such as: (1) how much influence an article has on different topics in subsequent time periods; (2) which is the most important one among all the citations for an article; and (3) is the article talking about similar topics as its citations? All of these questions can be answered through topic modeling approaches such as DIM.

In conclusion, topic modeling is a very flexible model building framework that allows scholars to incorporate more information in their analyses. Moreover, topic modeling will permit us to include currently used citation or social network information in the model. One approach to accomplish this is to use a hierarchical topic model (D. Blei, Griffiths, Jordan, & Tenenbaum, 2004). It is also possible to include article downloads and view times within the DIM framework by combining supervised Latent Dirichlet Allocation (Li, Ouyang, & Zhou, 2015) with DIM. With the flexibility to create more robust interpretations of scholarly impact, we anticipate greater use of topic modeling as research in this area moves forward.

References


Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ (Clinical research ed.)*.


