Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 15(3): 165-169, Sep. 2017 Regular paper

# Construction of Scientific Impact Evaluation Model Based on Altmetrics

**Jiapei Li[1], Seong Yoon Shin[2], and Hyun Chang Lee[3*], *Member, KIICE***

[1]Department of Library Information Consulting, Hebei Geology University, Shijiazhuang 050031, China
[2]School of Computer Information & Communication Engineering, Kunsan National University, Gunsan 54150, Korea
[3]Department of Digital Contents Engineering, Wonkwang University, Iksan 54538, Korea

## Abstract

Altmetrics is an emergent research area whereby social media is applied as a source of metrics to evaluate scientific impact. Recently, the interest in altmetrics has been growing. Traditional scientific impact evaluation indictors are based on the number of publications, citation counts and peer reviews of a researcher. As research publications were increasingly placed online, usage metrics as well as webometrics appeared. This paper explores the potential benefits of altmetrics and the deep relationship between each metrics. Firstly, we found a weak-to-medium correlation among the 11 altmetrics and visualized such correlation. Secondly, we conducted principal component analysis and exploratory factor analysis on altmetrics of social media, divided the 11 altmetrics into four feature sets, confirming the dispersion and relative concentration of altmetrics groups and developed the altmetrics evaluation model. We can use this model to evaluate the scientific impact of articles on social media.

**Index Terms**: Altmetrics, Correlation analysis, Evaluation model, Principle component analysis

## I. INTRODUCTION

In 2010, Priem et al. [1] first proposed “altmetrics” in “Altmetrics: A Manifesto” as a new source of metrics for measuring scientific impact. Altmetrics are metrics and qualitative data that are complementary to traditional, citation-based metrics. As more and more publications and other research output are used online, “use metrics” (clicks and downloads) and webometrics have come into being [2]. They can include (but are not limited to) peer reviews on Faculty of 1000, citations on Wikipedia and in public policy documents, discussions on research blogs, mainstream media coverage, bookmarks on reference managers like Mendeley, and mentions on social networks such as Twitter [3].

Scientists are introducing all sorts of altmetrics to assess the scientific impact of articles, to make up for the limits of citations such as shortage of diversification and a long time lag, and to strive to reflect the scientific impact of scientific literature objectively. Some scholars have conducted statistical analysis on the relationship between altmetrics on social reference managers like Mendeley, and some have looked into the role of purely social networking websites like Twitter in academic communication, but no calculation model is developed yet to measure the scientific impact of social networking media. In this paper, we collected the value of altmetrics of the top 100 most-discussed journal articles on the website Altmetrics.com in 2016, implemented

**Received** 07 August 2017, **Revised** 14 August 2017, **Accepted** 19 September 2017
***Corresponding Author** Hyun Chang Lee (E-mail:hclglory@wku.ac.kr, Tel: +82-63-850-6260)
Department of Digital Contents Engineering, Wonkwang University, 460, Iksan-daero, Iksan 54538, Korea.

Open Access https://doi.org/10.6109/jicce.2017.15.3.165 print ISSN: 2234-8255 online ISSN: 2234-8883

statistical analysis to find out the relationship between multiple altmetrics on mainstream social media, reduced the dimensions of and grouped altmetrics, and eventually formed a social media-based calculation model for the scientific impact of altmetrics.

## II. DATASETS AND METHODS

In this paper, we chose the top 100 most-discussed journal articles in 2016 from Altmetrics.com, downloaded the DOI, link, theme, summary, media coverage, and metrics on blogs, Twitter, Facebook, Google+, Wikipedia, Video and F1000 (December 7, 2016) via the website link (https://figshare.com/collections/Altmetric_Top_100_2016/3590951).

We used the RStudio version 3.3.3, including its statistical environment and the following packages: xlsx, ggplot2, corrplot, psych, tm, dplyr, wordcloud2, etc. Firstly, we applied Spearman correlation analysis to investigate the correlation between metrics of the top 100 articles on Altmetrics.com in 2016 and found a correlation between Google Scholar Citation and all the altmetrics, ranging from weak to strong. Secondly, we conducted principal component analysis (PCA) on the alternative metrics in the datasets, grouped the 11 metrics to four group based on the analysis result. Finally, we calculated the principal component score coefficient matrix and developed an evaluation model for scientific impact assess of the articles.

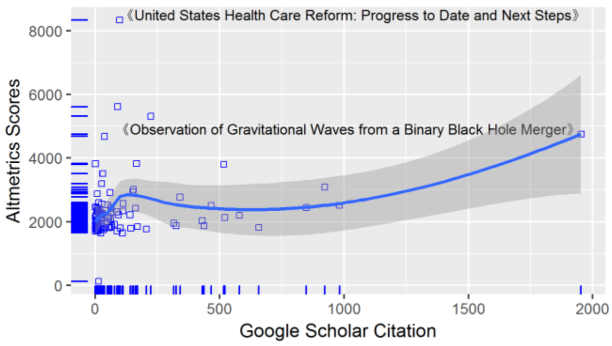


**Fig. 1.** Correlation between Google Scholar citation and Altmetrics Scores.

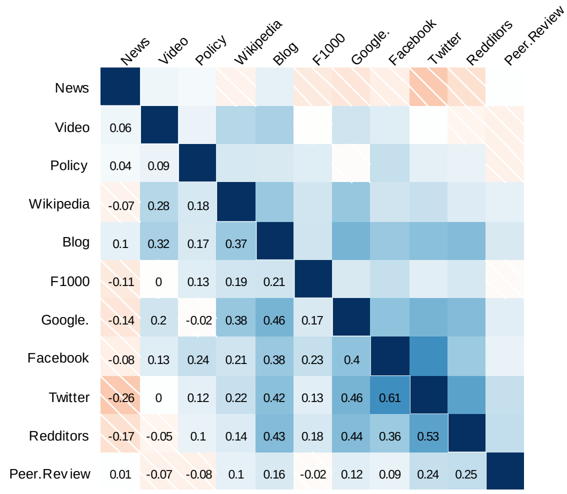


**Fig. 2.** Spearman correlation analysis of 11 altmetrics.

## III. RESEARCH AND ANALYSIS

To reveal the relationship between traditional metrics including citation and alternative metrics, we used the Altmetrics Score provided by Altmetrics.com (which was calculated based on the value of each individual alternative metric) to calculate the coefficient between Google Scholar citation and Altmetrics Score and found their correlation coefficient to be 0.246 (correlation test $p$ = 0.01372 which is less than 0.05 and the correlation coefficient value is acceptable), meaning that the two have weak correlation. We then drew a scatter plot with marginal carpet and locally weighted fitting line, as shown in Fig. 1 from which we can see that, the citations of these top 100 articles in the datasets are mostly within 500 and their Altmetrics Score densely distributed between 2,000 and 4,000. Besides, we drew the locally weighted polynomial fitting line (loess curve) for the statistical analysis and the shadow covers 95% of the fitting area. Meanwhile, two points capture our attention. One is the highest point. It represents the article "United States Health Care Reform Progress in Data and Next Steps", with the Altmetrics Score as high as 8,340 and 98 citations, meaning that it receives much attention from social media. The other is the farthest point. It stands for the article "Observation of Gravitational Waves from a Binary Black Hole Merger", with much social media attention (Altmetrics Score = 4,750) and 1,953 citations. These two points are marked in Fig. 1.

Since samples in the datasets are not in normal distribution, we conducted Spearman correlation analysis on altmetrics and visualized the process [4]. As shown in Fig. 2, variables are News, Video, Policy, Wikipedia, Blog, F1000, Google+, Facebook, Twitter, Redditor, and Peer Review. The correlation between these 11 variables is shown in Fig. 2 (the blue indicates a positive correlation and the red negative correlation). The darker the color, the bigger the absolute value of correlation coefficient is. As far as the correlation coefficient value is concerned, the correlation between these variables ranges from weak to medium [5]. In particular, Blog is positively correlated with the rest 10 variables while News is negatively correlated with most of the other variables.

**Table 1.** KMO and Bartlett test

| KMO | Bartlett test *p*-value |
|---|---|
| 0.61 | <2.2e-16 |

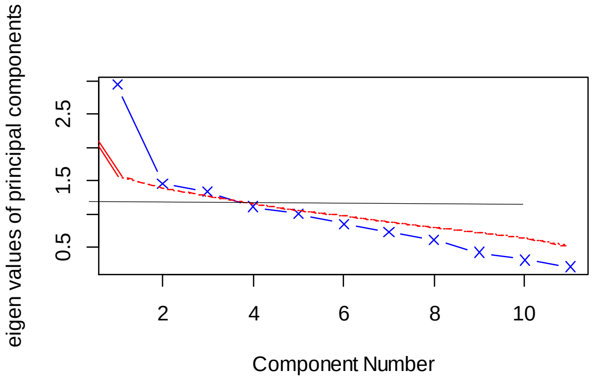


**Fig. 3.** Scree plot with parallel analysis.

**Table 2.** Principal component analysis (step-1)

| | RC1 | RC2 | RC3 | RC4 | h2 | u2 |
|---|---|---|---|---|---|---|
| News | 0.06 | 0.78 | -0.11 | 0.02 | 0.62 | 0.38 |
| Blog | 0.45 | 0.28 | 0.50 | 0.17 | 0.56 | 0.44 |
| Policy | 0.12 | 0.19 | -0.15 | 0.37 | 0.21 | 0.79 |
| Twitter | 0.79 | -0.12 | 0.34 | 0.01 | 0.75 | 0.25 |

**Table 3.** Principal component analysis (step-2)

| | RC1 | RC2 | RC3 | RC4 |
|---|---|---|---|---|
| SS loadings | 2.22 | 1.52 | 2.00 | 1.12 |
| Proportion Var | 0.20 | 0.14 | 0.18 | 0.10 |
| Cumulative Var | 0.20 | 0.52 | 0.38 | 0.62 |
| Proportion explained | 0.32 | 0.23 | 0.29 | 0.16 |
| Cumulative proportion | 0.32 | 0.84 | 0.62 | 1.00 |

PCA is a data-reduction technique that transforms a larger number of correlated variable into a much smaller set of uncorrelated variables called principal components (PC). These PC are linear combinations of the observed variables.

Table 1 shows the result of KMO and Bartlett test. It can be observed that the KMO value is 0.61 and the *p*-value of the Bartlett Test 2.2e-16<0.05, so the datasets is suitable for factor analysis.

In the PCA of the datasets, the first step is to determine the number of principal components. The most common approach is based on the eigenvalue. Each component is associated with an eigenvalue of the correlation matrix. The first PC is an associated with the largest eigenvalue, the second PC with the second largest eigenvalue, and so on. The Kaiser-Harris rule suggests we keep principal components whose eigenvalue is bigger than 1. As shown in Fig. 3, the plot displays the screen test based on the observed eigenvalues (as straight-line segments and x's), the mean eigenvalues derived from 100 data matrices (as dashed lines), and the eigenvalues greater than 1 criteria (as a horizontal line at y=1). In the Fig. 3, a scree plot (the line with x's), eigenvalues greater than 1 criteria (horizontal line), and parallel analysis with 100 simulations (dashed line) suggest retaining four principal component out of the 11 variables.

Table 2 illustrates the result of the four PC extracted via varimax rotation and the loadings (pattern matrix). Form Table 3 we can see that, because of varimax rotation, each principal component is tagged as RC (rotate component) and the component loadings of RC1, RC2, RC3, and RC4 are shown in the table. We can re-analyze the four principal components according to their respective loading for explaining each of the variables, as shown in Table 2. As a matter of experience, we think that when the loading is bigger than 0.5, it means that the variable in question is well explained by the principal component concerned. Thus, as shown in Table 4, we categorize the four principal components into Mass Social Media, Media Coverage, Academic Record and Peer Review. It is noticed that the common factor variance of the component of statistical variable Policy h2=0.21, meaning that the degree of Policy's variance explained is 0.21. For Policy, u2=0.79 (u2=1-h2), which explicates that the ratio of variance that cannot be explained by the principal component is 0.79. It thus can be seen that in the process of principal component analysis, Variable Policy is not well explained by the extracted principal components due to its features. That's why variable Policy doesn't emerge in the four extracted principal components.

RC, principle component using rotation approach, h2 component communalities, the amount of variance in each

**Table 4.** Principal components and their naming

| Principal component | Variables and component loadings | Name |
|---|---|---|
| $RC_1$ | Twitter (0.79), Facebook (0.71), Redditors (0.81) | Mass Social Media |
| RC2 | News (0.78), Video (0.72) | News Media |
| RC3 | Wikipedia (0.90), Google+ (0.86), Blog (0.5)* | Academic Record |
| RC4 | Peer Review (-0.59), F1000 (0.76) | Peer Review |

variable explained by the components; u2, component uniqueness, the amount of variance not accounted for by the components (or 1-h2).

We measured the importance of each variable in the process of evaluation after specifying the number of PC to be four. First, Table 5 gives the PC$scores coefficient matrix. Accordingly, we derived the formulas to calculate the score of each PC Then, we used the proportion of variance explained to the accumulative proportion of variance explained ratio as the weight of each principal component and eventually developed the impact evaluation model. According to Table 5, we can calculate the score of RC1, RC2, RC3, and RC4, respectively. The formula is as follows:

$$\begin{aligned} RC_1 = {} & 0.01N + 0.12B + 0.07P \\ & + 0.35T + 0.24PR + 0.32FB \\ & - 0.15W - 0.05G + 0.43R \\ & - 0.09V + 0.07F \end{aligned} , \tag{1}$$

$$\begin{aligned} RC_2 = {} & 0.53N + 0.14B + 0.12P \\ & - 0.14T - 0.09PR + 0.22FB \\ & - 0.09W + 0.01G - 0.08R \\ & + 0.48V - 0.02F \end{aligned} , \tag{2}$$

$$\begin{aligned} RC_3 = {} & -0.11N + 0.19B - 0.12P \\ & + 0.05T - 0.13PR - 0.07FB \\ & + 0.15W + 0.45G - 0.15R \\ & + 0.07V - 0.01F \end{aligned} , \tag{3}$$

$$\begin{aligned} RC_4 = {} & 0.00N + 0.14B + 0.32P \\ & + 0.00T - 0.53PR + 0.09FB \\ & - 0.06W - 0.05G - 0.05R \\ & - 0.05V + 0.69F \end{aligned} . \tag{4}$$

According to Table 3, the Proportion Var of the four principal components is 0.20, 0.14, 0.18, and 0.10, ranked in the descending order. We used the result and Cumulative Var to calculate the portion of variance explained by each principal component, which was 0.32, 0.23, 0.29 and 0.16, respectively. Thus we developed the following evaluation model:

$$\begin{aligned} RC = {} & 0.32RC_1 + 0.23RC_2 \\ & + 0.29RC_3 + 0.16RC_4 \end{aligned} . \tag{5}$$

To uncover the latent structure in the set of variables, we conducted exploratory factor analysis (EFA) on the datasets. The goal of EFA is to explain the correlations among a set of observed variables by uncovering a smaller set of more fundamental unobserved variables underlying the data. These hypothetical, unobserved variables are called factors (each factor is assumed to explain the variance shared among two or more observed variables, so technically, they're called common factors). In fact, though most researchers don't think there's a big gap between PCA and EFA (also known as the principal axis factor method), Widman argues that the principal axis factor method has more accurate factor loadings than PCA for it uses the square of the multiple correlation coefficient as the initial estimation of the common variance, and repeats the process again and again until it obtains a definite value of common factor. Hence Widman recommends the principal axis factor method over PCA.

EFA takes similar steps with PCA. First, we decided the number of common factors to be extracted. For EFA, the eigenvalue according to the Kaiser-Harris rule should be bigger than 0 (not 1), so we decided to extract four common factors from this datasets. Secondly, we used the axis iteration method to extract these common factors. Thirdly, to explain the meaning of the exported loading matrix better, we used oblique rotation to extract common factors. Finally, drew the result was shown in Fig. 4. The figure shows that

**Table 5.** Principal component score (PC score) coefficient matrix

| | RC1 | RC2 | RC3 | RC4 |
|---|---|---|---|---|
| News Story (N) | 0.01 | 0.53 | -0.11 | 0.00 |
| Blog (B) | 0.12 | 0.14 | 0.19 | 0.14 |
| Policy (P) | 0.07 | 0.12 | -0.12 | 0.32 |
| Tweet (T) | 0.35 | -0.14 | 0.05 | 0.00 |
| Peer Review (PR) | 0.24 | -0.09 | -0.13 | -0.53 |
| Facebook (FB) | 0.32 | 0.22 | -0.07 | 0.09 |
| Wikipedia (W) | -0.15 | -0.09 | 0.51 | -0.06 |
| Google+ (G) | -0.05 | 0.01 | 0.45 | -0.05 |
| Redditors (R) | 0.43 | -0.08 | -0.15 | -0.05 |
| Video (V) | -0.09 | 0.48 | 0.07 | -0.03 |
| F1000 (F) | 0.07 | -0.22 | -0.01 | 0.69 |

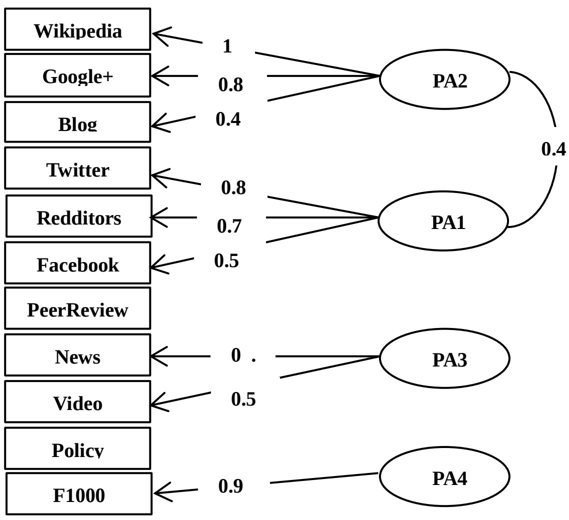


**Fig. 4.** EFA of the oblique four-factor solution.

three out of the four common factors corroborate the principal components of PCA, namely, Academic Records (Wikipedia, Google+, Blog), Mass Social Media (Twitter, Facebook, Redditor) and News Media (News, Video). In EFA, factor 4 (PA4) includes only F1000 and excludes Peer Review. What's more, the variable Policy is not explained by any common factor. It thus can be seen that except for "Peer Review", the EFA result and the PCA result are fitted. In other words, the reduction of dimensions of the 11 variables is true and acceptable.

## IV. CONCLUSIONS AND OUTLOOKS

We found mainly a weak-to-medium correlation between altmetrics using correlation analysis. But they are also concentrated to some degree for they can be divided into four categories: Academic Records, Mass Social Media, News Media, and Peer Review, based on the PCA results. The evaluation model can be used to evaluate the scholar impact of an article on the social media.

Altmetrics give us a unique social perspective to analyze the impact of academic research findings and trace academic communication among readers. Social media platforms contain a lot of comment texts about scientific articles [6]. We should better analyze them through statistical analysis [7], sentiment analysis, text classification and clustering, and machine learning to obtain implicit, unknown useful information from them, and thus better support scientific research and discovery.

## REFERENCES

[1] J. Priem, T. Taraaborelli, P. Groth, and Neylon, "Altmetrics: a manifesto," 2010 [Internet], Available: http://altmetrics.org/manifesto/.

[2] L. Bornmann, "Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics," *Journal of Informetrics*, vol. 8, no. 4, pp. 895-903, 2014.

[3] S. Haustein and V. Lariviere, "A multidimensional analysis of Aslib proceedings: using everything but the impact factor," *Aslib Journal of Information Management*, vol. 66, no. 4, pp. 358-380, 2014.

[4] X. Li, M. Thelwall, and D. Giustini, "Validating online reference managers for scientific impact measurement," *Scientometrics*, vol. 91, no. 2, pp. 461-471, 2011.

[5] R. Costas, Z. Zahedi, and P. Wouters, "Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective," *Journal of the Association for Information Science and Technology*, vol. 66, no. 10, pp. 2003-2019, 2015.

[6] M. Erdt, A. Nagarajan, S. C. J. Sin, and Y. L. Theng, "Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media," *Scientometrics*, vol. 109, no. 2, pp. 1117-1166, 2016.

[7] M. Taylor, "Exploring the boundaries: how altmetrics can expand our vision of scientific communication and social impact," *Information Standards Quarterly*, vol. 25, no. 2, pp. 27-32, 2013.

**Jiapei Li**
received her M.S. degree from information department in Tianjin normal university in China. From 2008 to the present, she has been an assistant professor in the Library of Hebei geology university in China. Her research interests include data science and text mining.

**Seong Yoon Shin**
received his M.S. and Ph.D. degrees from the Dept. of Computer Information Engineering of Kunsan National University, Gunsan, Korea, in 1997 and 2003, respectively. From 2006 to the present, he has been a professor in the same department. His research interests include image processing, computer vision, and virtual reality.

**Hyun Chang Lee**
Dean of Digital Content Engineering Department at Wonkwang University in Korea. He obtained his Ph.D. and M.S. from the University of Hongik, Seoul in Korea, in 2001 and 1996, respectively. Since 2008, he has been a professor in Wonkwang University. His major interests include business intelligence, IoT, AR&VR and data science fields.