User Profiling in Altmetrics: the need to enrich altmetric data

By Tint Hla Hla Htoo, Jin-Cheon Na. Nanyang Technological University

1. Introduction

In NISO's altmetrics white paper, data quality and gaming are among the important issues that can undermine the reliability and validity of altmetrics as indicators of research impact (NISO, 2014). In the same paper, developing strategies to build up trust is suggested as one of the potential action items in addressing these issues. In this regard, we observe that altmetric.com, one of the available altmetric tools, has adopted some policies to develop trust. For example, altmetric.com maintains the policy that data curated via altmetric.com must be completely transparent. Thus, altmetric score is calculated with data from auditable sources where the identity of the person is verifiable. To facilitate the interpretation altmetric data, different weightings which reflect the relative reach of data source are also given to different sources when calculating altmetric score (Altmetric.com, 2017).

In this paper, we propose user profiling as an addition measure to improve trust in altmetrics as research impact indicators and to help address data quality concerns such as promotion. We also believe that user profiling is necessary to facilitate the correct interpretation altmetric data. By that, we mean, when calculating altmetric score, different weightings which reflect different types of users may be given to take into consideration promotional and automatic nature of posts by publishers and educational or promotional accounts. Our suggestions were informed by the preliminary findings of a user profiling study which examined six characteristics of users tweeting research papers on Twitter: gender, geographic location, academic, non-academic, individual and organization.

2. Profiling of Twitter users

2.1. Data collection

For our analysis, we used part of the data collected in 2015 for the two previous studies by Htoo & Na (2017) and Na (2015). First, 133 articles published during the period 2008-2013 with the highest number of tweets were selected from top 70 Psychology journals in Social Science Citation Index. For each article, up to 20 tweets in English were then collected. Identical repeated tweets were removed but retweets were included. In the final selected dataset, there were a total of 2,016 tweets for analysis. Additional data, i.e. profile details of each tweeter, were collected from Twitter through API. In the final dataset, name, screen name, description, profile image, time zone, location, statuses count, followers count, and friends count of 1472 tweeters were included.

2.2. Exploratory data analysis

Based on the description in their twitter profile, users were manually categorized into four main types: academic vs non-academic, and individual vs organization. Table 1 in Appendix shows various types of users in each category. Individual users were categorized into male or female categories based on their first name and profile picture. Geographic information was determined based on time zone, instead of location, in their twitter profile.

2.3. Key findings

There are a few key findings. Firstly, we observed that majority of users (86%) were from North America and Europe indicating the possibility that, if, in general, tweets for research articles are mainly in English, Twitter as an alternative metric has Western bias. In terms of gender distribution, male dominant was found with 66% of users being male and only 34% female. With 68% and 32% of tweets respectively, male users tweeted twice more than female users.

Secondly, as shown in Table 2 in Appendix, for tweets identifiable as tweeted by either individuals or organizations, about 77% were tweeted by individuals while 23% by organizations. Among individual users, the number of non-academic users and tweets by non-academic users were twice more than that of academic users and tweets (Table 3). On the other hand, among organization users, the number of tweets by academic organizations were much higher than those by non-academic organizations (Table 4). Overall, if not differentiated between individual and organization users, there was similar number of tweets by academic and non-academic users (Table 4). To sum up, altmetric data was a mix of activities by individuals and organizations. It was, in general, activities by individual users revealed societal impact while activities by organization users revealed more of scholarly impact. Therefore, if altmetric data are to be used as indicators of societal impact, filtering of data is necessary to separate activities by individual and organization users.

Lastly, contrary to popular belief that most tweets are by publishers promoting their own articles, we found that only about 9% of tweets were from organization users such as publishers and educational accounts that use automatic tweeting from RSS feeds from a number of sources. However, it should be noted here that tweets in our dataset were selected from a small number (133) of highly tweeted articles. It is possible that in general, articles normally receive only a few tweets and they are mostly tweeted by publishers and accounts of educational or promotional in nature. So this needs further investigation.

3. Conclusion

User profiling reveals biases in altmetric data as well as different nature of impact generated by different types of users. While there were legitimate tweets by individual users sharing research articles of their interest, there were also promotional tweets by publishers and accounts that use automatic tweeting from RSS feeds from a number of sources. In that sense, the same principle of weighting used for different data sources in calculating altmetric scores provided by altmetric.com may be applied to different user types in calculating altmetric scores.

Based on findings from this study, we would recommend altmetric tool providers to include various user characteristics (e.g., gender, geographic distribution, academic, non-academic, individual, and organization) in altmetric data to enhance correct interpretation of altmetric data and to increase trust in altmetric data.

References

Altmetric.com. (2017). How is the altmetric attention score calculated? Retrieved from https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-attention-score-calculated-

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

Chen, V. H. H., & Wu, Y. (2015). Group identification as a mediator of the effect of players' anonymity on cheating in online games. *Behaviour & Information Technology*, 34(7), 658-667.

Cho, D., Kim, S., & Acquisti, A. (2012). *Empirical analysis of online anonymity and user behaviors: the impact of real name policy*. Paper presented at the 2012 45th Hawaii International Conference on System Sciences.

Christopherson, K. M. (2007). The positive and negative implications of anonymity in Internet social interactions: "on the Internet, nobody knows you're a dog". *Computers in Human Behavior*, 23(6), 3038-3056.

Costas, R., Zahedi, Z., & Wouters, P. (2015). 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*, 66(10), 2003-2019.

Htoo, T. H. H., & Na, J.-C. (2017). Disciplinary differences in altmetrics for social sciences. *Online Information Review*, 41(2), 235-251.

Na, J.-C. (2015). User motivations for tweeting research articles: A content analysis approach. In R. B. Allen, J. Hunter, & M. L. Zeng (Eds.), *Digital Libraries: Providing Quality Information: 17th International Conference on Asia-Pacific Digital Libraries, ICADL 2015, Seoul, Korea, December 9-12, 2015. Proceedings (pp. 197-208).* Cham: Springer International Publishing.

NISO. (2014), *Alternative Metrics Initiative Phase I White Paper.* Retrieved from http://www.niso.org/apps/group_public/download.php/13809/Altmetrics_project_phase1_white_paper.pdf.

Santana, A. D. (2014). Virtuous or vitriolic: the effect of anonymity on civility in online newspaper reader comment boards. *Journalism Practice*, 8(1), 18-33.

Appendix

Table 1. User types

Categories	**Academic**	**Non-Academic**
Organization	Universities Research Institutes Publishers/News Feeds Researcher Networking Groups Research Teams/Projects Libraries	Commercial Organizations Interest and Reference Forums NGOs and NPOs Networking Groups Social Campaign Forums Hospitals and Clinics Product Pages Public Figure Pages Church and Religious Organizations Event Pages Museums

Individual	Researchers Faculty members Postgraduate students Undergraduate students Librarians	Journalists Corporate/private personnel Scientists Healthcare professionals Professionals in health-related industry Others (e.g. Patient/Family member of patient)

Table 2. Individual vs. Organization

	No. of Tweets	**No. of Tweeters**
Individual	1512(77%)	1199 (83%)
Organization	456(23%)	251(17%)
Total	**1968**	**1450**

Table 3. Individual [Academic vs. Non-academic]

	No. of Tweets	**No. of Tweeters**
Academic	577(38%)	400 (33%)
Non-academic	937(62%)	801 (67%)
Total	**1514**	**1201**

Table 4. Organizations [Academic vs. Non-academic]

	No. of Tweets	**No. of Tweeters**
Academic	280(61%)	114(45%)
Non-academic	177(39%)	137(55%)
Total	**457**	**251**

Table 5. Academic vs. Non-academic

	No. of Tweets	**No. of Tweeters**
Academic	857 (43%)	514 (35%)
Non-academic	1,114 (57%)	938 (65%)
Total	**1971**	**1452**