Some practical difficulties in creating field-normalized altmetrics

Michael Taylor
Digital Science

“There is only one thing in life worse than being talked about, and that is not being talked about.” Oscar Wilde

Abstract: The path to creating field-normalized metrics using altmetric data is not clear. Citation data is sufficiently similar to some altmetrics that citation-based normalization might provide some insights into what metrics can be developed. However there are sufficient differences in the type, structure and distribution of many altmetric data to merit a critical approach before directly replicating the algorithms used in Scopus’s Field Weighted Citation Index (FWCI) and Web of Science’s Normalized Citation Index (NCI). Consideration has to be made on whether the data being normalized is aggregated according to type of data before normalizing for subject area, or at an Altmetric Attention Score level. Alternatively, rather than computing a number, it might be possible to develop mathematical and theoretical metrics that assign documents to categories of behavioural type, that describe their position and performance in the sharing cycle. This presentation will explore some of the practical difficulties in creating field-normalized altmetrics and invite discussion and insights from amongst the workshop participants.

Introduction: Databases such as Scopus and Web of Science calculate field-normalized metrics based on citation count. Both FWCI and NCI present the value as a ratio of citations received to an expected citation value. This latter number is a mean calculated from a set of peer articles, typically published in the same subject area and year. The Relative Citation Ratio (RCR) is a newer metric, developed by the National Institutes of Health, that also presents a ratio of citations received to an expected citation value: however the RCR calculates the expected value by reference to an article’s co-citation network.

An obvious approach would be to apply the same approaches to calculate field-normalized altmetric **numerical values**: dividing some actual altmetric value (e.g. number of tweets) by a mean of a related group of articles, and presenting the ratio. However, the number of zero values associated with many altmetric data is non-trivial, and is often substantially higher than the equivalent number of articles with zero citations. Furthermore, the number of zero values is likely to be more significant than a simple lack of evidence. Both the percentage of documents with zero values and the mean values alter by subject area. Subject areas also vary by type of activity: for example, the number of tweets received by a subject area appears to be independent of the number of Mendeley saves.

However, altmetric data significantly differs from citation data, in a structural sense. When an article receives traditional citations, it can only receive them from other articles: the network is closed and homogeneous. A graph of altmetric data is, in contrast, open and heterogeneous. New forms of documents can be created, made public or otherwise discovered that 'cite' research articles. For example, altmetric providers and researchers are investigating how to include more policy documents and white papers. Furthermore, these citations typically only happen in a single direction: if an article quotes or 'cites' a tweet, it is unlikely that this data would be captured or analysed in any meaningful sense.

When we consider a citation graph, we do so knowing that the graph is a closed, homogenous network. Nodes in the graph can only be credited with citations from other nodes in the graph: an article's citability is a given, once the article has been placed in the graph. (Although altmetrics aren't derived from closed graphs, systems such as Mendeley are much more closely aligned with the closed, citable graph when compared to open social systems.)

Therefore, when considering the meaning of a field normalized metric, we may need to consider that the first stage of activity (of getting a non-zero altmetric value) is, in some way, a measurement of how shareable a document, or set of documents is. Authors, publishers and institutions can all play their part in promoting these attributes to the broader public and disseminating this work.

This may lead us to consider the value of assigning article's performances into categories, and then performing numerical analysis on the contents of those categories.

For example, a study of non-zero values might be useful in understanding different attributes of a subject area, independently of the initial metric. This observation - that an approach might be to divide altmetric data into zeros and non-zeros and report on them separately - might be of particular use to people in the scholarly community. For example, if a journal's articles fail to get any altmetric attention whatsoever (when compared to similar subject areas), it may indicate that their articles are not easy to share, or that the publisher doesn't encourage or support it. Once a set of articles is being shared in public, a different set of analyses might be undertaken, that takes into account the actual level of sharing.

Research is needed to ascertain the potential for normalized, **categorized metrics**: there is some evidence that there are other tiers in the stages of sharing that might be used to develop non-numerical metrics.

Background: The notion that there might be valid tiers, or categories of altmetric activity was first suggested by Mike Taylor (formerly of Elsevier) and Matt Hobby (currently of Elsevier). Their unpublished work suggested that there might be a number (between 3 and 4) of discrete and identifiable categories in altmetric data. It was suggested by Taylor that these might indicate the various stages that an article goes through as it is shared in different networks of individuals.

A future line of work might be to investigate this data with a view to understanding the patterns of network sharing. For example, an article might be initially shared through a small network of colleagues known to each other, then through a wider community of peers without direct connections as a topic cluster, then into a more loosely, ad hoc group, before eventually being spread through a potentially massive network of loosely related users. Taylor and Hobby's work suggested that these data - when examined as distribution curves - varied by discipline. A useful outcome would be to calculate and identify the stages of virality of each article, in the different altmetric data types.

Methodology: A data-set of 9610 articles drawn from the NIH's publication database, which have been tagged into 13 subject areas on an article-by-article basis will be presented. The data - which will be published on Figshare - will include citation and altmetric activity.

Discussion: Various strategies will be presented to create field-normalized values and categories, and the strengths and weakness of each approach are explored and discussions invited. Of particular interest are other approaches to field normalization that might be adapted and explored. The author is open to working with other researchers to explore how the mathematics of categorization might be used in exploring more future-orientated approaches to understanding the application of graph theory and network behaviour, as it relates to the altmetrics of scholarly research.

With contributions from Stacy Konkiel (Altmetric), Aaron Sorensen (Uberresearch).

Acknowledgements to Matt Hobby (Elsevier).