Challenges and Opportunities for detecting and measuring diffusion of scientific impact across heterogeneous `altmetric` sources

Brian Davis[1], Ioana Hulpuş[1], Mike Taylor[2], Conor Hayes[1]

[1] Insight Centre for Data Analytics,
National University of Ireland, Galway,
Ireland
{first.last}@insight0-centre.org
[2]
{mike.taylor}@elsevier.com
Elsevier Labs, Oxford,
United Kingdom

1 Introduction

Alternative Metrics (also known as altmetrics) - alternative measures of scholarly impact has achieved increasing recognition among scientists and researchers of scientometrics, the field dedicated to measuring science, technology and innovation. Currently our lab is working on developing an innovative platform for incorporating diverse sources of scientometric data; both traditional (e.g. bibliometric) and new (e.g. social media) in order to capture a comprehensive view of scientific practice and discourse as well as the diffusion of scientific impact across the Web and social media - a type of altmetrics *analytics dashboard*.

We argue for an approach based on ingesting and integrating vast amounts of social media and traditional mainstream web content into a large knowledge graph for analysis. Examples of data sources include such as mainstream news, blogs, microblogs, data from official government agencies i.e. white papers, legislation and policies and of course traditional bibliometric sources. This involves challenges in multiple fields such as text mining and graph mining and analytics. Text mining techniques must be researched to identify mentions of scientists and organisations (entity linking), but also to detect most informative scientific statements in both scientific publications and text produced by the lay person. Furthermore, tracking of these claims over vast amounts of web content is a challenge that needs to be addressed.

The second class of challenges we identify are in the field of network analysis. While the text analysis allows us to link texts to scientists, and to create networks of scientific claims, research must be done in order to understand how these complex networks can be most effectively and efficiently used for the purpose of altmetrics. The current context of social media, mainstream news, the ease of access to government documents, as well as to massive organised collections of scientific publications brings the opportunity to study the complex links and connections between people (journalists, politicians, scientists) at the same time as the links between the content they produce (hyperlinks, paraphrases). This is a great opportunity for understanding scientific impact and scientific finding diffusion, but it comes with great research challenges and questions, some of which we present in the following section.

2. Challenges in Altmetrics Research

What information was diffused? While chasing direct references (through `doi` or hyperlinks) and mentions (i.e. a scientist's name) can reveal parts of the impact of scientific contributions, most of the scientific output is not referred directly. Does this mean that altmetrics should only give credit to scientists who benefit from such direct exposure? Text-mining techniques must be researched that are able to automatically extract statements and scientific claims from heterogeneous altmetrics sources i.e. news, blogs or government policies, and trace them back to the originator scientist or scientific community. Such efforts would allow linking between statements in online media and the scientist who may have inspired them. This might be crucial in order to maximise the coverage of altmetrics over the scientific community. Significant progress made in domains like text-summarisation made it already possible to extract such informative statements from texts [3]. Also, progress in meme-tracking makes it already possible to track such statements [4].

How did the scientific information diffuse? Did the diffusion occur via mainstream news to the blogosphere or microblogs? Who are the key players, stakeholders or gatekeepers of scientific diffusion? While some answers to this are already formed by researchers [4], there is no work to focus especially on scientific content. We think such a research direction is of utmost importance in order to understand what it takes for a scientific finding to become viral, and in what measure such phenomena are driven by scientific quality, or whether they their merit is a consequence of other external factors.

How to quantify the impact of individual scientists / publications / organisations? The ultimate goal of altmetrics is to rank publications, scientists and organisations. Such rankings imply the existence of scores that can be computed. In this regard, we aim to engage the altmetrics community into finding potential definitions for the scientific impact. Intuitively, and in light of the traditional ways of computing importance on the Web (i.e PageRank), one can argue that the mere count of direct / indirect mentions is not enough, but a certain reputation of the mentioning entity must be part of the equation. For instance, a mention / quotation of a paper in the homepage of a scientist might bear less weight than such a mention in a worldwide news article. In general, which are the generally accepted intuitions about scientific impact? After defining such intuitions, we plan to propose scientific measures that are able to capture and quantify them. However, a recurrent problem in altmetrics is the problem of evaluation, addressed in the next question

How can altmetrics be scientifically evaluated? As with everything that bears a strong innovative aspect, we are facing a lack of ground-truth knowledge. What is an acceptable and scientifically sound way of evaluating altmetrics? The most available quality data concerning scientific impact is that of traditional bibliometrics. How much correlation with bibliometrics can we expect? High correlation might mean that an altmetric does not bring new knowledge, while low correlation might be judged as a failure to capture the impact. What are the alternative ways of evaluating altmetrics? Questionnaires to the scientific community? Peer judgement? We consider it of utmost importance to the altmetric community that it produces such a ground-truth data set, otherwise research in the domain cannot be empirically evaluated and will suffer greatly. It is well known among scientists that there is little drive towards research in domains in which ground-truth is not available.

3. Conclusion

In summary we have described and outlined research challenges and opportunities for detecting and measuring diffusion of scientific impact across heterogeneous altmetric sources on the Web. We argue that the research questions described above are important and warrant presentation and discussion within the altmetics community and we would welcome the opportunity to do so .

References

1. Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A.: The Stanford Digital Library Metadata Architecture. Int. J. Digit. Libr. 1 (1997) 108–121
2. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)
3. GUPTA, V., LEHAL, G.. A Survey of Text Summarization Extractive Techniques. Journal of Emerging Technologies in Web Intelligence, North America, 2, aug. 2010.
4. Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09). ACM, New York, NY, USA, 497-506.