Journal of Informetrics 12 (2018) 461–480

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty

Henry Small*

SciTech Strategies, Inc., 105 Rolling Road, Bala Cynwyd, PA 19004, USA

ARTICLE INFO

Article history:
Received 14 February 2018
Received in revised form 21 March 2018
Accepted 22 March 2018

Keywords:
Biomedicine
Highly cited papers
Citation contexts
Method papers
Uncertainty
Hedging
Machine learning
Corpus linguistics
Logistic regression
Empiricism
Lowry's method

ABSTRACT

The top 1000 biomedical papers by number of citations are classified by method, type of method and non-methods by examination of citation contexts. Supervised machine learning is applied to the context data for a training sample of papers which is then used to classify the full list, revealing that words indicating utility are most important for the classification of methods. Further word analysis is carried out using corpus linguistics to uncover context words that characterize non-methods. Hedging words are found to play an important role for non-methods, and several are selected for further analysis with logistic regression. Other variables in the regression are a consensus variable based on the similarity of contexts for a paper and another variable based on whether citations come from "methods" sections of citing papers. Accuracy of predictions from logistic regression is comparable to machine learning. The results are interpreted in terms of the perceived certainty or uncertainty of the underlying knowledge, that is, methods and their outputs have higher certainty, and non-methods higher uncertainty. Evidence is found that hedging is inversely related to citation frequency. Implications of this work for the study of the development of science and the role of methods and tools in biomedical research are discussed.

1. Introduction

For many years scientists and bibliometricians have been puzzled by lists of the most cited papers in science. Why do these lists not conform to our expectation that key discoveries in science such as the theory of relativity, the genetic code, or quantum mechanics should appear near the top of the citation count rankings? Instead, we find that methodologies dominate. Even Eugene Garfield, as he was creating the first citation index in 1961, was somewhat dismayed that a paper by Oliver H. Lowry on protein determination was so heavily cited (Wouters, 1999, 72), so much so that for a moment he had doubts about the usefulness of his index. With stacks of printouts of citations to Lowry on the floor of his office, he wrote to Joshua Lederberg, the Nobel laureate who encouraged him to undertake the project: "I have a sort of panic about this sample and wonder whether this can be useful to anyone." (Wouters, 1999) Lederberg, however, told him not to worry, that the paper was the most frequently quoted paper in biochemistry because it had become the standard method for the protein determination. The Lowry paper turned out to only be the tip of the iceberg and many other highly cited method papers would be highlighted in subsequent years.

* Corresponding author.
E-mail address: hsmall@mapofscience.com

https://doi.org/10.1016/j.joi.2018.03.007
1751-1577/© 2018 Elsevier Ltd. All rights reserved.

Why are our expectations so far off the mark? Are we working under the false assumption that citations are a pure reflection of what is important in science, and that discoveries must carry the greatest importance? Garfield's eventual explanation was that breakthrough papers such as the Watson-Crick discovery of the DNA double-helix can be quickly superseded and replaced by improved formulations, or obliterated by becoming standard usage (Garfield, 1977). Of course, it has been found that some discoveries do achieve high rates of citation in a relatively short period of time, and appear on highly cited lists. Using data from a recent study, it was estimated that at least seven percent of papers in the top 1000 papers ranked by total citations were discoveries (Small, Tseng & Patek, 2017). An alternative hypothesis is that papers containing the methods and tools that scientists use to arrive at their findings should be expected to be the most heavily cited. This might be termed the utilitarian hypothesis, and begs the question, what makes Lowry's paper so useful and compelling to scientists? Is the frequent use of some methods simply the reflection of scientists wanting to obtain credible data to support their hypotheses?

The goal of this paper is to study the phenomenon of highly cited papers from the standpoint of what authors say when they cite them within the so-called citation contexts or citing passages. We will analyze citation contexts for linguistic markers that are associated with methods, and explore the hypothesis that high citation rates are associated with the certainty of the knowledge that is generated. Citation contexts for non-method papers such as discoveries will also be examined for possible linguistic cues that differentiate them from method papers and reveal their role in the knowledge system.

2. Background

The important role of methods in the advancement of biomedical knowledge has often been commented on. For example, Olby in his history of the double helix describes the crucial role that methods played in the elucidation of the structure of DNA (Olby, 1974, 435). Generally, methods and tools in science are seen as providing relatively firm points of reference against which theories can be tested or constructed. For example, Pierre Duhen asserted "Agreement with experiment is the sole criterion of truth for a physical theory." (Duhem, 1962, 21). And John Ziman commented ". . experimental evidence is public knowledge, *par excellence*, with the power of carrying complete conviction." (Ziman, 1968, 32). A successful theory can be seen as consisting of a mix of assumptions and empirical findings which fit together like the pieces of a puzzle. In Kuhn's theory, the paradigm provides a framework of high certainty for experimental and theoretical findings (Kuhn, 1970). In times of crisis, however, when experiment disagrees with theory, the weak link in the chain of reasoning must be found. According to Duhem, this is often more a matter of intuition than of logic (Duhem, 1962, 216). In the history of science, it is most often the theoretical constructs that must give way rather than the experimental findings obtained from the application of methods and tools. Thus, the development of science is critically dependent on the perceived uncertainty of theoretical constructs and the relative certainty of experimental methods.

Garfield was the first to draw attention to the prevalence of method papers in highly cited lists. Over the years, he published numerous essays in *Current Contents* that presented lists of most cited papers from various time periods and journal subsets that highlighted the prominence of methods (Garfield, 1977; Garfield, 1990; Garfield, 1991). In addition, the Citation Classics Commentary series published in *Current Contents*, where authors discussed their highly cited papers, often featured method papers. Lowry, himself, provided material for one such commentary in which he stated that his method, though widely cited, was not a great scientific accomplishment, but merely a more reliable version of earlier methods (Lowry, 1977). Garfield commented, "Is any reasonable person going to claim that the intellectual achievement represented by Einstein's Unified Field Theory is less significant than a convenient method of protein determination simply because Einstein is cited less frequently?" (Garfield, 1973). He goes on to suggest that perhaps it has to do with the relative number of investigators doing protein determination versus field theory.

Method papers also emerged as an issue in early clustering experiments with co-citation (Small & Griffith, 1974). It was found that very highly cited method papers had to be removed or normalized prior to clustering to break up large macro-cluster or giant components that joined together the various specialty clusters. Method papers were like diffuse clouds hovering over the specialties. This work illustrated the trans-specialty and sometimes trans-disciplinary nature of methods.

Studies that attempt to classify the reasons papers are cited usually come up with substantial numbers of citations that fall into a "methods or tools" category (Bornmann & Daniel, 2008), and many of the citer motivation classification schemes have explicit categories for the citation of methods. However, these studies usually are focused on samples of citing papers and do not look at the nature of the cited work. More recent studies that attempt to automate the recognition of citer motivation use a combination of the linguistic analysis of the citation context and location within the IMRaD structure of the scientific paper, but the focus is on the citing instance and not the cited work (Bertin, Atanassova, Sugimoto, 2016; Teufel, Siddharthan, & Tidhar, 2006). Recent studies of the distribution of references across the IMRaD structure of citing papers have found that method sections contain fewer and older references than other sections (Bertin, Atanassova, Gingras, 2016). Another similar study not explicitly looking at IMRaD sections found a consistent text location (measured in character centiles) for highly cited papers which the authors inferred was the location of the methods section but otherwise did not examine the nature of the cited work (Boyack, van Eck, Colavizza & Waltman, 2018).

Recently the journal *Nature* published a study of the most cited 100 papers with data obtained from the Web of Science (Van Noorden, Maher & Nuzzo, 2014). The authors begin by pointing out that some of the landmark discoveries of the 20th century do not appear in the top 100 papers, and on the contrary ". . . the vast majority describe experimental methods or software that have become essential in their fields." They claimed: "To make exciting advances, researchers rely on relatively