Journal of Informetrics 11 (2017) 46–62

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

Discovering discoveries: Identifying biomedical discoveries using citation contexts

Henry Small [a,*], Hung Tseng [b,1], Mike Patek [c]

[a] SciTech Strategies, Inc., 105 Rolling Road, Bala Cynwyd, PA 19004, USA
[b] National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, 6701 Democracy Boulevard, Bethesda, MD 20892, USA
[c] SciTech Strategies, Inc., 58 Russell Street, Keene, NH 03431, USA

ARTICLE INFO

Article history:
Received 25 August 2016
Received in revised form 12 October 2016
Accepted 6 November 2016
Available online 19 November 2016

Keywords:
Discovery
Biomedicine
Citation contexts
Citances
Machine learning
Pubmed central

ABSTRACT

A procedure for identifying discoveries in the biomedical sciences is described that makes use of citation context information, or more precisely citing sentences, drawn from the PubMed Central database. The procedure focuses on use of specific terms in the citing sentences and the joint appearance of cited references. After a manual screening process to remove non-discoveries, a list of over 100 discoveries and their associated articles is compiled and characterized by subject matter and by type of discovery. The phenomenon of multiple discovery is shown to play an important role. The onset and timing of recognition of the articles are studied by comparing the number of citing sentences with and without discovery terms, and show both early onset and delays in recognition. A comparative analysis of the vocabularies of the discovery and non-discovery sentences reveals the types of words and concepts that scientists associate with discoveries. A machine learning application is used to efficiently extend the list. Implications of the findings for understanding the nature and justification of scientific discoveries are discussed.

1. Introduction

Discoveries are the *sine qua non* of science. They are how scientists learn new things about the world, make sense of reality, and advance the boundaries of the known into the realm of the unknown. They are also what scientists strive to make, what advances their careers and status, and what they fight over for priority (Merton, 1957). Discoveries can be solutions to known problems or solutions to problems that only later become manifest. But not every discovery pans out. Like polywater, cold fusion, or N-rays, some fail to gain acceptance and fall by the wayside. However, some are so compelling that they command immediate assent and even astonishment, like Archimedes jumping out of the bath tub exclaiming "Eureka!" (Koestler, 1964, p. 106) or Jim Watson saying that the double helix model of DNA is too beautiful not to be true (1968, p. 205). But what are the hallmarks of scientific discovery and how do we know when a discovery is made?

Philosophers of science have drawn the distinction between the context of discovery and the context of justification (Losee, 1972, p. 115; Reichenbach, 1949). The context of discovery, or nascent moment as Holton (1973, p. 17) calls it, can be governed by chance, erroneous information, and even dreams. The context of justification, on the other hand, is where

* Corresponding author.
E-mail addresses: hsmall@mapofscience.com (H. Small), tsengh@mail.nih.gov (H. Tseng), mpatek@gmail.com (M. Patek).
[1] Hung Tseng's views expressed here are personal and do not represent those of the NIAMS/NIH.

http://dx.doi.org/10.1016/j.joi.2016.11.001
1751-1577/

cooler heads must evaluate cold facts. Popper (1959, p.31) argued that philosophy can only deal with the latter period where hypotheses are put to the most stringent tests. The crucial point is that discovery is more than just an insight, inspiration, or lucky guess. It must also pass some initial threshold of justification and survive a process of ongoing challenges. This second stage may involve corroboration, confirmation by others, and demonstrating consistency with existing experiment and theory (Stent, 1972). In this paper we will be primarily concerned with the context of justification, not the "aha" moment of initial inspiration, and with the process by which the scientific community comes to label a finding as a "discovery", although in the end we will call the separation of these contexts into question.

Kuhn (1962, p. 52) said that discovery is not possible without a paradigm which sets our expectations. When an expectation is violated, a problem is born which we can then attempt to solve. The solution to the anomaly may require a revolution or revamping of our understanding, which then opens up new questions. Problems that arise within the context of a paradigm are called puzzles. When DNA became recognized as critical to inheritance (Dubos, 1976), the natural question arose "What is the molecular structure of DNA and how does it enable inheritance?" When questions crystallize within a community, a competition among scientists can ensue to find a solution. Of course, the recognition of an unsolved problem or open question requires a deep understanding of the current state of knowledge. Scientists may even lack the framework to ask a question such as "How does gravity affect time?", a question which would be unlikely to come up without relativity theory. Thus, earlier discoveries can set the stage for later problems and discoveries. As Olby (1974, p. 426) said about Watson and Crick's double helix structure of DNA, it was not just that it fit with the known facts about DNA but that it opened up new questions and set the framework for future work. This has been variously described as the fruitfulness of a theory (Kuhn, 1977, p. 322).

Others have attempted to model the discovery process in computer programs, conceiving all problems as puzzles whose solutions could be found by some kind of heuristic search (Langley, Simon, Bradshaw, & Zytkow, 1987). Another research tradition sees discovery as the finding of novel combinations. In 1964 Arthur Koestler introduced his ideas on "bisociation" – the joining of two frames of reference to arrive at a novel synthesis. Swanson's work (1986) is in a similar vein, involving the connecting of previously unconnected areas of biomedical knowledge – or more accurately, indirectly connected areas – to gain new knowledge. More recently Foster, Rzhetsky and Evans (2015) explore scientists' problem choices and show that risky choices that pay off result in greater recognition than conservative choices that remain within the paradigm. They operationalize this on a network of chemical entities that have been connected in article abstracts and look for novel combinations, the more risky and unlikely the combination, the greater the surprise and the reward.

Perhaps all discoveries involve a degree of surprise, and the source of this may be the unexpected convergence between the conjecture and the evidence, or, as Ziman (1968, p. 48) describes it, as the falsification of a preconceived or vague notion. However, because all new knowledge is tentative and subject to revision, it can take a period of time and contributions by many researchers until the initial conjecture comes to be regarded as a "discovery" by the community. Hence all discoveries are retrospective designations even though some lags may be very short and others quite long.

Despite efforts to construct a theory, the ability to systematically identify discoveries has remained elusive. No comprehensive inventory has been created. The usual approach is to rely on the pronouncements and press releases of scientists themselves or their interpretation by science writers. It would appear at first glance that citation analysis, where we can observe the impact of scientific articles over time, is an ideal tool for identification, and indeed simple citation counts do identify many scientific discoveries (Garfield, 1979). However, there are numerous reasons for citation, and highly cited lists tend to be dominated by methods, reviews and data compilations.[2] Thus, simple citation counts do not provide enough information for a definitive identification. In this paper we propose a method that augments citation counting with the language used by citing authors, namely words that explicitly label referenced items as discoveries. This method, together with machine learning to omit false positives, greatly improves our ability to automate discovery identification. Once an accurate list is in hand we can begin to work backwards to find common characteristics that can shed light on the nature of discovery.

2. Data and methods

Fortunately, we are now gaining access to an expanding corpus of machine readable scientific articles in full text and we can use this resource to study the contexts in which articles are cited, so called citation context analysis. An important source of curated full text for the analysis of scientific papers is PubMed Central® (PMC). This open repository was created in 2000 and includes papers that were required to be publically available under the National Institutes of Health public access policy and legislative mandates.

We limited our study of biomedical discoveries to the full text from PubMed Central called the "open access subset". This subset includes 1.1 million full texts of primarily biomedical articles covering publications mainly in the most recent period but also some coverage extending back several decades. The oldest article found in the subset was from 1896, but 90% of articles are from the last 13 years (counting through mid-2015). Over the time period, the coverage rapidly expanded from 4500 articles in 2000 to about 200,000 in 2014.

[2] Of the 100 most cited articles in Pubmed Central only about four are discoveries and the remaining 96 either methods, reviews or data compilations.