Journal of Informetrics 12 (2018) 59–73

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

Characterizing in-text citations in scientific articles: A large-scale analysis☆

Kevin W. Boyack[a,*], Nees Jan van Eck[b], Giovanni Colavizza[c], Ludo Waltman[b]

[a] SciTech Strategies, Inc., Albuquerque, NM, USA
[b] Centre for Science and Technology Studies (CWTS), Leiden University, The Netherlands
[c] Digital Humanities Laboratory, École Polytechnique Fédérale de Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 9 October 2017
Received in revised form
21 November 2017
Accepted 21 November 2017

Keywords:
In-text citations
Citation position analysis
Field-level analysis
Reference age
Citation counts

ABSTRACT

We report characteristics of in-text citations in over five million full text articles from two large databases – the PubMed Central Open Access subset and Elsevier journals – as functions of time, textual progression, and scientific field. The purpose of this study is to understand the characteristics of in-text citations in a detailed way prior to pursuing other studies focused on answering more substantive research questions. As such, we have analyzed in-text citations in several ways and report many findings here. Perhaps most significantly, we find that there are large field-level differences that are reflected in position within the text, citation interval (or reference age), and citation counts of references. In general, the fields of *Biomedical and Health Sciences*, *Life and Earth Sciences*, and *Physical Sciences and Engineering* have similar reference distributions, although they vary in their specifics. The two remaining fields, *Mathematics and Computer Science* and *Social Science and Humanities*, have different reference distributions from the other three fields and between themselves. We also show that in all fields the numbers of sentences, references, and in-text mentions per article have increased over time, and that there are field-level and temporal differences in the numbers of in-text mentions per reference. A final finding is that references mentioned only once tend to be much more highly cited than those mentioned multiple times.

1. Introduction

The increasing availability of full text from scientific articles in machine readable electronic formats is a development with the potential to greatly impact citation analytics and to significantly improve the accuracy of models of the structure of science. Full text contains information not only on the exact locations of in-text citations within articles, but also on the context in which a citation to previous work is made. Specific problems that can be addressed using full text data include classification of in-text citations by type and function, and improving measures of impact by weighting of citations based on polarity, typology, function, citing location, and perhaps other features as well. Weighting of citations also has the potential to impact our knowledge of the structure of science, in that document clustering (and the resulting maps) could be based

☆ The peer review process of this paper was handled by Vincent Larivière, Associate Editor of Journal of Informetrics.
* Corresponding author.
E-mail addresses: kboyack@mapofscience.com (K.W. Boyack), ecknjpvan@cwts.leidenuniv.nl (N.J. van Eck), giovanni.colavizza@epfl.ch (G. Colavizza), waltmanlr@cwts.leidenuniv.nl (L. Waltman).

https://doi.org/10.1016/j.joi.2017.11.005
1751-1577/

on a more accurate measure of the relatedness between documents. These applications, although beyond the scope of this paper, motivate the current work, which studies the characteristics of in-text citations (and associated features) in two large full text databases. A solid understanding of the characteristics of in-text citations is required before advanced applications of full text data, such as those mentioned, can be most fruitfully pursued.

Study of in-text citations and related text from scientific documents using full text sources has a long history. Although both positional (the location of references) and semantic (the meaning of references) studies have been pursued, here we focus primarily on the positional aspect. The terminology used in previous studies of in-text citations is not consistent. Thus, to avoid confusion, we define our terminology here. A *reference* is an item in the bibliography or reference list of a document. An *in-text citation* is a *mention* of a reference within the full text of a document. A reference can be mentioned one or more times in a document. Each mention is an in-text citation. We use the terms *in-text citation* and *mention* interchangeably in this article.

Our work examines distributions of in-text citations for two large full text datasets – the PubMed Central (PMC) Open Access subset and a large portion of the Elsevier full text corpus. Using these large and disciplinarily broad datasets, we will show that there are significant variations in the distributions that have not been reported before. We specifically investigate field-level dependencies and report citation count distributions as a function of text progression.

The paper proceeds as follows. We first review relevant literature and then describe our datasets and analysis methods. Results are then reported along with key observations. The paper concludes with a summary, mention of limitations and suggestions for additional work.

2. Background

Over the years, many studies of the location or position of in-text citations have sought to identify the relative value of citations as a function of the position or the number of mentions. Implicit among many of these studies is the assumption that references that are more related to the citing article are the more valuable or essential references for that article. In reviewing prior work, we focus on those studies that explicitly include mention location in their analysis. As will be shown below, regarding distributions of in-text citations, there is a consensus among previous studies that mentions tend to be more concentrated at the beginnings (e.g., introduction and related work) and endings (e.g., discussion and conclusion) of articles than in the middle sections. There is also a rough consensus that references that are mentioned outside the introductory sections tend to be the most valuable.

Early studies were necessarily done by hand with small datasets. In one of the first studies, Voos and Dagaev (1976) examined citations to a set of four highly cited articles, two from biology, one from medicine and one from physics. Despite their very small sample, their findings suggested that a) most mentions come from introduction sections, b) the location and the number of mentions – which early studies often referred to as öp. cit.¨– were both important in determining the value of a citation, c) time was important, and d) different disciplines had different citation patterns. Bonzi (1982) used a set of nearly 500 citations from 31 articles and found that the number of times a work is cited in the text "shows promise of predicting relatedness between citing and cited works".

Cano (1989) sought to study citation function and utility while also examining position. Using 344 references that were coded by function and utility by their authors, they found that references that were mentioned in a perfunctory and negational way were most often peripheral or of low utility. They also found that references classified as organic, conceptual, operational or evolutionary were more typically essential or of higher utility, and that mentions were more concentrated in the first 15% of an article. Hooten (1991) examined 417 citing contexts and found that references with larger numbers of in-text citations seemed more related to the citing paper, and thus more essential, than those with only one in-text citation.

McCain and Turner (1989), using a set of 11 highly cited papers, created an index based on citing location, number of in-text citations, citation utility from citation contexts, and self-citation, finding that papers with a later citation peak (at six years) were more broadly useful than those with an early citation peak. Citations to papers with a later citation peak were more often for methodological advances rather than for experimental results or theoretical concepts. Maričić, Spaventi, Pavičić, and Pifat-Mrzljak (1998) examined citation contexts as well as locations using 11% of the mentions to a set of 357 articles, and suggested that references should be valued differently based on the section of the citing article in which they appear. They found references with relatively low "meaning" (or value) to be mentioned predominantly in the introduction, while those mentioned in other sections had higher meaning.

Bornmann and Daniel (2008) examined a set of 350 in-text citations to a set of articles written by grant applicants. Using the IMRaD (introduction, methods, results and discussion) structure, they found that while more mentions appeared in the introduction and discussion sections of citing articles, the methods and results sections were slightly enriched with mentions to articles with higher citation counts. In perhaps the most detailed comparison with ground truth data available, Tang and Safer (2008) surveyed authors of 49 articles in biology and 50 articles in psychology who assessed the mentions in their articles for importance, reason for citation, and relationship to the cited author. They found that reference importance increased proportionally with numbers of mentions and more detailed discussion of the cited document. In addition, the authors considered references mentioned in the methods and results sections to be most important, while those mentioned in the introduction section only were less important than those mentioned in other sections. Hou, Li, and Niu (2011) studied 651 biochemistry papers and found that references that shared at least 10 references with the citing paper had, on average,