Journal of Informetrics 9 (2015) 860–871

Contents lists available at ScienceDirect

Journal of Informetrics

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

Assessing the impact of software on science: A bootstrapped learning of software entities in full-text papers

Xuelian Pan[a,b], Erjia Yan[c,*], Qianqian Wang[d], Weina Hua[a]

[a] School of Information Management, Nanjing University, Nanjing, China
[b] Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing, China
[c] College of Computing and Informatics, Drexel University, Philadelphia, USA
[d] College of Humanities, Jinling Institute of Technology, Nanjing, China

ARTICLE INFO

Article history:
Received 27 May 2015
Received in revised form 30 July 2015
Accepted 30 July 2015
Available online 10 September 2015

Keywords:
Entity extraction
Information extraction
Software
Software citation
Citation analysis
Bootstrapping

ABSTRACT

Although software has helped researchers conduct research, little is known of the impact of software on science. To fill this gap, this article proposes an improved bootstrapping method to extract software entities from full-text papers and assess their impact on science. Evaluation results show that the proposed entity extraction system outperforms three baseline methods on extracting software entities from full-text papers. The proposed method is then used to learn software entities from all papers published in *PLoS ONE* in 2014. More than 2000 unique software entities are obtained which accounted for more than 20,000 mentions and more than 7000 citations. The paper finds that software is commonly used in the scientific community along with a substantial uncitedness.

1. Introduction

The science of science community has a tradition of using publications to assess the productivity and impact of authors, institutions, and regions (e.g., Yan & Guns, 2014; Yan & Sugimoto, 2011). Publications form the foundation for scholarly communication and shape the epistemological cultures in science (Cronin, 2008; Hyland, 2004; Klein, 1996). The pursuit of publications by scientists has resulted in a fast volume growth in virtually every scientific discipline and has brought forward the so-called “publish or perish” phenomenon (Fanelli, 2010; Nature Editorial, 2010).

Publications also play a distinctive role in research evaluations, particularly for funded research, as there is the need for an accountability for government sponsored R&D expenditures to justify the investment benefits of knowledge production and innovation (e.g., Fortin & Currie, 2013; Jacob & Lefgren, 2011; Wang, Liu, Ding & Wnag, 2012). The general public should be informed of the short- and long-term impact of these expenditures (Lane, 2009).

Recent years have witnessed a changing landscape in scholarly publication and communication: open access, open data, open source, preprint, and social media have deeply changed the way scientists communicate science (Ball & Duke, 2012; Borgman, 2011; Kraker, Leony, Reinhardt, Gü, & Beham, 2011; Starr & Gastl, 2011; Sugimoto et al., 2013; Vision, 2010). For instance, more and more scientists choose preprint archives as the ultimate publication outlet (Larivière et al., 2014); in the

[*] Corresponding author at: College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104-2875, USA. Tel.: +1 215 895 1459.
E-mail address: ey86@drexel.edu (E. Yan).

http://dx.doi.org/10.1016/j.joi.2015.07.012
1751-1577/© 2015 Elsevier Ltd. All rights reserved.

meantime, social media such as Twitter, blogs, and TED talks are becoming popular choices for scientists to disseminate their research outcomes (Haustein, Peters, Sugimoto, Thelwall, & Larivière, 2014; Haustein et al., in press; Sugimoto et al., 2013). In addition, more scholars have made their articles, datasets, and software publicly available on the Internet (Kraker et al., 2011). These sharing activities, which benefit the researchers and the public at large, are a part of open science. Although open science has attracted a lot of attention, open source as a component of open science did not to the same extent as its counterparts open access and open data. Tensions exist: on the one hand, more open source software packages are delivered and many of them are used in the scientific community; on the other hand, little is known of the use and impact of open source software in science.

Another vital change on the horizon is the definition of research outputs: publications have been long seen as the end research outputs—this notion has become more transient in recent years as digital outputs such as software can be the end products in many contemporary scientific inquiries. While some journals have made the attempt to associate publications to certain software (Candela et al., 2015), the more common practice has been that software is referenced in unsystematic ways in scientific literature. They can be embedded in documents by digital object identifiers (DOIs), hyperlinks, and featured on dedicated websites or simply be mentioned in paragraphs. Therefore, a clear gap exists in how to incorporate digital outputs, such as software, as an integral component in studies of science of science. A study to assess the impact of software on science is thus imperative, because it will complement the current publication-driven science of science research and help build an open, transparent, and inclusive scientific reward system.

One cornerstone of this endeavor is the design of text-based methods to identify software entities in full-text corpora because these entities are largely mentioned in the text rather than formally cited in the way as their publications counterpart. This paper will serve this purpose by the design and evaluation of a bootstrapping method to automatically extract software entities from a full-text data set. Specifically, as an exploratory study, we intend to examine:

- The method to extract software entities from full-text corpora: what text-based method can be proposed to extract software entities; how effective is the proposed technique; how does it compare with prior information extraction techniques.
- The popularity of pieces of software in science: how prevalent is software use in science; what are the most frequently used pieces of software in science.
- Software use and citation impact: what are the most highly cited pieces of software; what is the relationship between software use and citation.

This study is part of a large effort to examine software attribution and citation (see NSF, 2014). The answers to the above questions will help survey the current status of software use in science and lay a solid foundation for succeeding research in this vein. These efforts will inform our understanding of software reference contexts help build an open, transparent, and inclusive scientific reward system.

2. Literature review

2.1. Rule-based information extraction

Information extraction has attracted attention from both researchers and practitioners for years (Chiticariu & Reiss, 2013). Rule-based approaches have become popular, because they are interpretable (Chiticariu & Reiss, 2013) and adaptable by incorporating domain knowledge (Kluegl, Atzmueller, & Puppe, 2009; Grimes, 2011). The development of rule-based information extraction systems is the process of improving extraction coverage and accuracy and in the meantime reducing time complexity and human supervision (Huang, 2014).

Early in rule-based information extraction, researchers used hand-coded rules to identify information from text (Hearst, 1992; Appelt, Hobbs, Bear, Israel, & Tyson, 1993). Although the experimental results showed that these systems performed well, the process of creating rules was time-consuming. To reduce human supervision, some information extraction systems were designed to automatically generate rules using an annotated training corpus, such as AutoSlog (Riloff, 1993), PALKA (Kim & Moldovan, 1993), and CRSTAL (Soderland, Fisher, Aseltine, & Lehnert, 1995); it was found that these corpus-based methods extracted information more effectively than the previous hand-coded approaches.

Nonetheless, annotating a training corpus is still a demanding task. To further reduce human involvement, researchers developed weakly supervised methods for information extraction. Riloff's research group proposed the bootstrapping method to automatically extract information from an unannotated text corpus using heuristics (Riloff & Jones, 1999; Riloff, 1996; Thelen & Riloff, 2002). These approaches required a small number of seed terms and an unlabeled text corpus as input. The seed terms were used to generate contextual patterns and the top-ranked patterns were used to identify new terms. Then, the learned terms were used to create patterns that can identify more entity terms in an iterative way. To improve precision, researchers developed several measures, such as pattern accuracy and confidence, to discard bad-performing patterns and entities (Lin, Yangarber, & Grishman, 2003; Yangarber, Lin, & Grishman, 2002). The caveat is, however, that discarding patterns below certain threshold can result in low coverage in a small data set. Alternatively, Gupta and Manning (2014a) focused on learning good-performing patterns by predicting labels of extracted entities. This system required the use of external domain dictionaries and was incapable of identifying positive entities from unlabeled entities extracted by top-performing patterns. All these weakly-supervised methods relied on the assumption that “relevant patterns should