Universität Rostock

gegründet 1419

Traditio et Innovatio

Investigating the Role of Software in Science by Automatic Knowledge Graph Construction through Natural Language Processing

Dissertation
zur
Erlangung des akademischen Grades
Doktor-Ingenieur (Dr.-Ing.)
der Fakultät für Informatik und Elektrotechnik
der Universität Rostock

vorgelegt von

M.Sc. David Schindler, geb. am 30.05.1994 in Burglengenfeld

Santa Marta, Kolumbien, 12.04.2024

Abstract

Software has become increasingly important in data-driven research and plays an integral role in today's science, contributing to all steps of scientific investigations. Knowledge of how software is applied in scientific investigations is essential to the scientific community in terms of provenance, because knowledge of utilized software is a prerequisite for reproducibility and software influences the outcome of scientific research. Furthermore, tracking the usage and impact of software provides a central feedback mechanism for the development of research software, a task often performed by researchers themselves. However, knowledge of software application in research is sparse, with a majority of information on software usage being informally described in the textual descriptions of scientific research, leading to a lack of software visibility. This thesis, develops a method to systematically analyze software usage in scientific publications at a large scale. For this purpose, I created the high-quality ground-truth dataset SoMeSci, which is based on manual annotation covering all relevant knowledge of software usage in scientific publications, such as software metadata (e.g., versions or developers), contextual information, and unique identifiers. Findings on this dataset revealed that providers of bibliographic data currently use unsuited representation formats for formal software citation, hampering systematic analyses of software citations. I therefore developed an automatic information extraction pipeline for software mentions in scientific articles, solving the problem with a high recognition rate of 86.6% in terms of FScore. However, my results also demonstrate the complexity of the task, highlighting the challenges of generalization and the complexity of the required Entity Disambiguation. The established extraction pipeline was applied on a large-scale dataset of 3.2 M articles from PubMed Central and the resulting data was formally modeled in the Research Knowledge Graph SoftwareKG, to allow a FAIR publication and reuse by the scientific community. Finally, I performed large-scale analyses to provide insights on software usage over time, between domains, in dependence of article impact, and concerning the state of open source software, as well as specific analyses tracking the publication of software, extendable software architectures, and the relation between software and article retraction. Overall, the resources of SoMeSci and SoftwareKG—made available in the scope of this work—build the bases for further analyses of software in science and can serve as a starting point to implement applications required by the scientific community, such as software impact measures or a software recommendation system.