S03-1 (02/26 13:00-13:40)

Experiments with large-scale citation networks

Thorsten Koch
Zuse Institute Berlin, Germany

Abstract

Journals were initially invented to distribute articles efficiently to those interested in a particular topic. Since articles can now be distributed easily everywhere electronically, classification and evaluation remain the most important functions of journals, as looking at all articles at once was impossible. However, as research output increases and research topics become evermore interdisciplinary, the effort to upkeep the existing system increases while performance deteriorates. On the other hand, our computational possibilities, both regarding hardware and algorithms, have increased dramatically during the last decades. The question arises whether it would not be better to look at all articles together. We conducted several experiments using the WOS, Crossref, and OpenALEX large-scale citation networks. Each has over 50 million articles and nearly a billion references. A major question in citation analysis is whether some observations are due to genuine characteristics of science or due to some underlying properties of the data set. We performed comparisons and mergings, gaining some insights we will report. Another question is how to classify articles or cluster them by topic. As science becomes increasingly interdisciplinary, it is clear that only multi-label classification or fuzzy clustering can correctly describe the relationship between articles. Many recent results in machine learning demonstrate that sticking to classic or historical human-devised viewpoints, in this case, human-selected topics, can possibly be a dead end. Therefore, we want to achieve an automated multi-label clustering based only on the available data itself. This -- as expected -- turned out to be quite challenging. In the presentation, we will report our results and the questions they induce.