Citation Analysis

Olga Dorabiala, Marta Wolfshorndl, Yang Zhou

In this project we will integrate the DBLP data with the citation data from arnet miner to produce a comprehensive analysis of citation attributes. We will download the data into postgresql and format it into a predetermined schema. We are interested in answering questions such as: 1) A self citation analysis determining the measurement of self-citation per paper for each author. It was recently reported in a paper published in 2019 that self-citation is a hidden problem in many scientific fields, artificially inflating an author's citation statistics. We will determine if this is the case in the DBLP data, who the worst offenders are in terms of authors and publications, and how many self-citations to other citations there are per paper. 2) How many citations away from a Turing award winner is the average author (in different fields)? We think it would be interesting to determine how many citations or coauthors each author is away from famous individuals in their field or in other fields. 3) What are the most influential papers and journals based on metrics such as the H-index? We will add more questions once we explore the data further. Possible topics include an analysis of which topics are more represented over time, is citation inflation happening in more recent papers, and is there a delayed recognition of influential papers? We have access to the data and tools we need as we will mostly be using postgresql. One challenge will be in working with such a large data set and in the data cleaning. Our final product will be a series of plots and images representing the answers to the queries we have made.

References:

Ioannidis, John PA, et al. "A standardized citation metrics author database annotated for scientific field." PLoS biology17.8 (2019): e3000384.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). pp.990-998.

Van Noorden, Richard, and Chawla D. Singh. "Hundreds of extreme self-citing scientists revealed in new database." Nature 572.7771 (2019): 578.