Analyzing the Citation Rate of a Scientific Paper Using ML and Graph Algorithms

Vladimir Zakharov, Sergey Fedchin, Artem Chebykin

Fall 2024

1 Introduction

1.1 Paper Overview

Citation rate is an important metric to know about a scientific paper. It is often referred to as a KPI for researchers. Governmental agencies and scientists pay attention to the citation rate, when evaluating the quality of the paper. Hence, it is important to understand what exactly affects the citation of your work. In our research, we investigate the connection between various properties of scientific papers and their citation rates, using machine learning and graph algorithms.

1.2 Field Overview

There are quite a few different papers that focus on predicting the citation rate. Most of them use only simple ML methods to obtain results, while we focus more on implementing graph algorithms to perform this task. The studies we have analyzed are:

Predicting High Impact Academic Papers Using Citation Network Features , that focused primarily on the graph features of the network, disregarded any meta-information about the paper.

Predicting the future success of scientific publications through social network and semantic analysis on the other hand, took into account the semantic structure of papers. However, this study used graph information only to create vertex features (betweenness, degree, etc.)

Learning to Measure Influence in a Scientific Social Network considered not only the author of a publication but also their co-authors, thus creating a more informative structure; yet only simple ML methods were used to conduct

the research.

2 Data Acquisition and Preprocessing

The amount of data and its quality is the basis for any data science research. To explore the dependencies between the paper citation rate and its inner properties, we have chosen the dataset: HepTh dataset .The dataset represents the scientific paper citation graph in the field of high-energy physics in the years 1991 to 2003

An oriented edge between vertices a→b shows, that paper a cites paper b. Apart from the graph information itself, we are also provided with various features of each paper, which are, however, not standardized. Some features only appear in few papers, while others are present for each. We have decided to keep and work with the following features present for each paper: date of publication, authors, title + abstract.

The task of processing the dates was particularly tricky one since that part of the data was stored in a multitude of different formats, and they had to be standardized. (We have chosen yyyy-mm-dd format for it). Furthermore, the graph included some edges that led from an older paper to a more recent one, which didn't make sense; hence, those edges were removed from our data. We have also transformed the list of authors from a string into an actual Python list and parsed the abstract from the metadata files provided in the dataset.

After all the described manipulations, the data was ready for further exploration and analysis.