Analyzing 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 thing to know about a scientific paper. It is often referred to, as a KPI for a researcher. Governmental agencies and other scientists pay attention to the citation rate, while evaluating the quality of the paper. Hence it is important to understand, what exactly affects citation of your work. In our research we will investigate the connection between various properties of the scientific paper and it's citation rate, using machine learning and graph algorithms.

1.2 Field overview

There are quite many different papers, which focused on predicting the citation rate. Most of them were using mostly ML methods to obtain results, while we will focus more on implementing graph algorithms to perform this task. The studies we analyzed are:

Predicting High Impact Academic Papers Using Citation Network Features

Predicting the future success of scientific publications through social network and semantic analysis

Learning to Measure Influence in a Scientific Social Network

2 Data acquiring and preprocessing

The amount of data and its quality is a fundamental basis for any research. To explore the dependencies between the paper citation rate and its

inner properties we have chosen the dataset: Citation dataset.The dataset represents the scientific paper citation graph in the field of high energy physics in the years 1991-2003

Oriented edge between vertecies a-¿b shows, that paper a citates paper b. Apart from the graph information itself, we are also provided with various features of each paper, which are, however, not standardized. Some of the features are only present in a few paper, and others are madatory for each. We have decided to keep and work with the following features, present for each paper: date of publication, authors, title + abstact.

The task of processing the dates was particularly tricky one, since that part of data was presented in a multitude of different formats and they had to be converted to uniform one. (We have chosen yyyy-mm-dd). Furthermore, the graph included some edges which 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 to an actual Python list and parsed the abstract from the meta information files, provided in the dataset.

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