

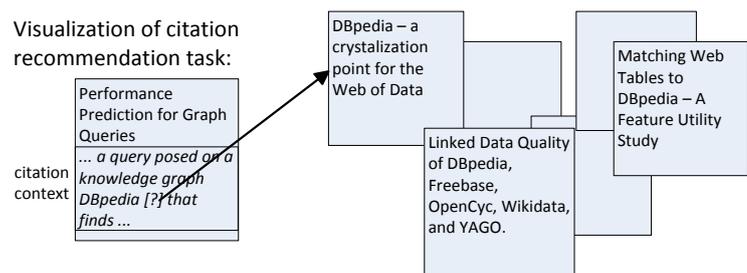
Call for Master Thesis

“Recommending Scientific Publications for Texts Using Neural Networks”

What is the topic?

Citation recommendation refers to the task of recommending (scientific) publications for a given text. Due to the vast amount of publications nowadays, citation recommendation can significantly help researchers in their daily work.

Visualization of citation recommendation task:



The Microsoft Academic Graph

(MAG) [1] is a large knowledge graph containing information about millions of scientific publications from various domains and from millions of researchers [2]. Due to its large size and noise-free representation, the MAG is in particular suitable to be used as knowledge repository for citation recommendation.

The goal of the proposed thesis is to develop a **citation recommendation system based on the MAG**. This means that a model is trained in a supervised machine learning setting in order to allow for selecting suitable publications for a given citation context out of the set of all available publications (typically hundreds of thousands or millions). More precisely, this thesis encompasses the following sub-steps:

1. (Optional) Given the publications represented in the MAG, **retrieve the body texts** of the publications as far as possible from the Web.
2. Learn **embeddings for all publications** represented in the MAG graph (and maybe for additional entities such as venues), using both *textual* information and the *links* within the MAG graph. The learned embedding vectors for papers can then already be used for calculating similarities among papers.
3. Learn **joint embeddings of both citation contexts and publications**. An example of joint learning can be found in [3]. It might be necessary to train separate models for different scientific domains, as the citation behavior might vary.
4. Given the joint embeddings, a **citation recommendation system** can be built, which mainly consists of applying the learned model of joint embeddings to new incoming citation contexts.
5. Evaluate the developed citation recommendation system automatically (by removing citations and re-predicting them) and manually, if necessary.

A thorough introduction to the topic and necessary literature will be provided by the supervisor.

[1] <https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/>

[2] <https://www.microsoft.com/en-us/research/project/academic/articles/march-2018-graph-update/>

[3] “Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation”, LREC 2016. <https://arxiv.org/abs/1601.01343>

Which prerequisites should you have?

- Interest in machine learning (word embeddings, neural networks, etc.)
- Programming skills for machine learning (e.g., using Python).
- Willingness to work with large data sets (on servers/clusters).

Keywords: Implementation, knowledge graph, scholarly data, embeddings, neural networks, machine learning, natural language processing, big data.

Contact: **Michael Färber**
michael.farber@cs.uni-freiburg.de