CSC-RUB PhD Project Proposal

Title: Text mining of scientific literature for materials discovery and knowledge generation

Sector of research: Computational Materials Science, Knowledge Generation, Discovery Informatics

Degree awarded: Dr.-Ing. (Phd)

Keywords: concept learning, knowledge acquisition, text mining, scientific discovery, hypothesis generation, metal alloys, defects

Supervisor of PhD project: Prof. Dr.-Ing. Markus Stricker, e-mail: markus.stricker@rub.de

Research focus of supervisor: My group, Materials Informatics and Data Science at the Interdisciplinary Centre for Advanced Materials Simulation (ICAMS), applies algorithms and methods from data sciences and machine learning to materials science problems. Our background is in Computational Materials Science with a focus on metals and alloys. Specific applications currently include data fusion from simulation and experiment to improve interpretability, the development of interatomic potentials based on neural networks for alloys which are not well described by analytical formulations, and data mining in large defect simulations to improve the understanding of their collective behavior in order to inform coarse-grained models.

Publications:
The choice for relevant publications is based on the objective of the research proposal: code development, defects in metals, machine learning techniques in physical metallurgy

(1) **Stricker, M.** et al. Machine learning for metallurgy II. A neural-network potential for magnesium Phys. Rev. Materials, American Physical Society, 2020, 4, 103602

(2) **Stricker, M.** & Curtin, W. A. Prismatic Slip in Magnesium J. Phys. Chem. C, American Chemical Society, 2020, 124, 27230-27240

(3) Musil, F. ... **Stricker, M.** et al. Efficient implementation of atom-density representations The Journal of Chemical Physics, 2021, 154, 114109

(4) **Stricker, M.** et al. Irreversibility of dislocation motion under cyclic loading due to strain gradients Scripta Materialia, 2017, 129, 69 – 73

(5) Roters, F. ... **Stricker, M.**; et al. DAMASK – The Düsseldorf Advanced Material Simulation Kit for modeling multi-physics crystal plasticity, thermal, and damage phenomena from the single crystal up to the component scale Computational Materials Science, 2019, 158, 420 - 478

Summary of research plan:

Background: Being able to keep an overview of the amount of scientific literature published each year poses a great bottleneck for scientific progress. The knowledge contained therein might not reach everyone and so a huge potential is not being used or possible connections are not being made. But issue also extends into the past: there might be findings which were neglected or overlooked and therefore never became part of the body of knowledge. It is essentially a issue of balancing knowledge generation with integrative analysis and ultimately controls the speed of progress of science. Internet search engines and specialized search engines for scientific literature provide one means to find possibly relevant literature, but exhaustively scanning the relevant literature manually is often tedious if not impossible. Recent advances in text mining technology

provide a path to automated analysis of knowledge contained in text form. Further, graph based algorithms provide the framework to generate hypotheses purely based on mined knowledge.

Study objective: The main objective of this project is to develop a system to autonomously mine materials science knowledge in text form (peer-reviewed journal articles) and generate a model which connects the knowledge as well as generate hypotheses for materials synthesis and design. Ideally, these hypothesis are tested with simulations as well as experiments through collaborations. The specific topics to be mined will be decided based on the applicants specific background but are ideally related, but not limited, to metals, mechanics, metallurgy, and crystal defects.

Expected Results: Several algorithms will be developed, applied and combined for text mining, knowledge acquisition, hypothesis generation, which all work in a federation to generate future discovery paths based on published literature; publication of results in peer-reviewed journals; participation and presentation of the results at international conferences.

Methods: The methods used include but are not limited to clustering techniques, graphical representations of knowledge, decision trees, graph diffusion; journal access is provided through RUB agreements with publishers; computational infrastructure is available through powerful personal workstations at ICAMS and through access to a high performance computing cluster.

Candidate Requirements:

- An excellent Master's degree in Engineering or Materials Science with a strong computational focus
- Good command of interpreted programming languages, ideally Python
- A high level of spoken and written English (IELTS band score of 6.5 or higher)

Motivation for CSC application (max 250 words): The successful candidate will work at the Interdisciplinary Centre for Advanced Materials Simulation (ICAMS), a central facility for Scale-Brigding Materials Modeling at Ruhr-University Bochum. She or he will have access to high performance computing facilities. He or she will benefit from existing collaborations and work groups of Prof. Stricker's group at RUB and with international collaboration partners. She or he further has the possibility to participate in the ICAMS Graduate School for Scale-Bridging Materials Modeling which includes interdisciplinary lectures, a graduate seminar, and soft skill training such as scientific presentation and writing.