Knowledge mining from scientific articles and reports

Background/ Motivation

The number of scientific articles and reports produced every year is increasing. Simple search on web of science for articles published under "aquatic toxicity", for example, gives about 2000 articles for year 2019. The number of papers written about the topic within 10-year-period reaches 20 000. The broader the term of the search topic, the higher number of articles are found by the search. The knowledge contained in the papers can be utilized for various purposes such as modeling, explanation or clarification of obtained results, and identification of future research questions. However, due to the amount and diversity of the existing information, existing knowledge can be overlooked, difficult to find, or not properly combined yet.

The aim of the current study is to extract knowledge, so called prior knowledge, from open literature in the last 20 years related to environmental, health and safety aspects of chemicals and later apply this knowledge in *in-silico* prediction models. The prior knowledge can refer to properties of the estimated functional relationship between target and predictor variables (e.g., hazard property and molecular structure) or input/output information [1]. Such kind of knowledge, obtained by data mining techniques, heuristics or generated from theoretical observations or provided by experts in the field, have been proven to increase the performance of data mining prediction models, if they are properly incorporated into it.

Variable constraints, predictors significance, correlation between predictors, information about the input data such as noise, outliers, way of addressing missing data can be the possible areas of knowledge exploration and generation with the aim to be further utilized in the prediction process. The knowledge can be incorporated in form of additional constraint functions or applied to choose or specify data mining model parameters.

Objective:

The overall objective of the project is to develop a procedure for extracting and systematizing knowledge from scientific articles and reports. This project utilizes natural language processing (NLP) tools for extracting available for selected molecular properties knowledge out of the literature of the last 20 years.

Tools/Databases:

Literature search: CitNetExplore, Iris.ai, Google scholar, Web of Science, Scifinder, PubMed

Text mining: Python & python packages for NLP

Prerequisites:

The applicant should have a background in chemical engineering with interest or previous experience in data science methods and applications or, vice versa, a background in computer science with interest or previous experience in chemical engineering applications.

Supervisors:

Gulnara Shavalieva gulnara.shavalieva@chalmers.se
Stavros Papadokonstantakis stavros.papadokonstantakis@chalmers.se
Chalmers University of Technology, Division of Energy Technology

References:

[1] S. Chen, C. Gao, and P. Zhang, "Enhancing Transparency of Black-box Soft-margin SVM by Integrating Data-based Prior Information," 2017.