Extracting Metal-Organic Framework Knowledge from Scholarly Big Data

Xintong Zhao

Drexel University

DREXEL UNIVERSITY Metadata Research Center College of Computing & Informatics

Introduction

Within the past decade, data-driven approaches have become the fourth paradigm of scientific knowledge discovery and hence have greatly accelerated the discovery process. Materials science, which has profound impact on people's life, has been benefited from big data and deep learning research with no exception [3].

Recently, studies extracting knowledge from scholarly data have gained significant attention from Metal-Organic Framework (MOF) researchers for two main reasons [1-2, 4]: (1) materials literature such as peer-reviewed conference/journal articles contain rich and diverse knowledge resources (e.g., experiment data, materials chemical structures, properties, synthesis methods and so forth); (2) the volume of material literature is massive, and it is still growing rapidly. Researchers urge to explore knowledge automatically instead of reading every piece manually.

To this end, we conducted a two-prompts research to explore the intellectual structure of MOF research community and to extract synthesis information from texts.

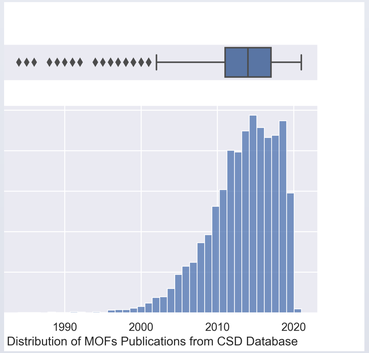

Figure 1. MOF Publication Distribution in Cambridge Database

Objectives

We address the following questions in our study:

1 – What are the major research topics in MOFs over the time?

2 – What and how to extract MOF synthesis information from text?

3– How to structure the extracted data to make sure they are informative and easy to retrieve?

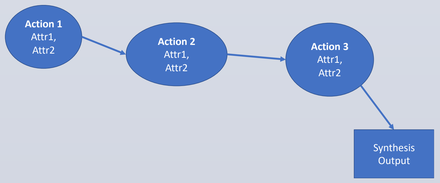

Figure 2. Modeling Synthesis Procedure as a Graph

Methods & Current Progress

I. Communication between Domain Researchers and Data Scientists

The flowchart shown at right provides a high-level demonstration of how domain and data researchers work together.

First, domain scientists provide a scientific question then identify data source/types. Second, once the research goal is clear, data scientists start to design machine learning models to extract target knowledge from data sources. Domain scientists help with validating the performance of designed models. If models perform as expected, we apply models in large scale and generate data output.

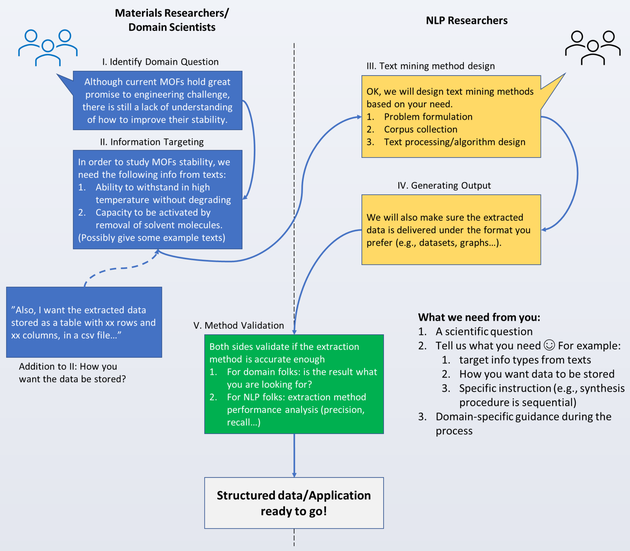

Figure 3. Demonstration of Domain/Data Communication

II. Exploring Research Trends in MOF Area

Research trends are research topics that have gain great attentions at different time periods. To discover research trends, we analyzed MOF research papers from 1980 to 2022. We used topic modeling techniques on 5353 paper abstracts and categorized them into different 5-year periods. We also analyzed their index terms to find out possible trends.

While the exploration is still ongoing, our current data suggest that gas separation and catalytic performance are two major applications of metal-organic framework structures. Figure 4 shows the change of topic mentioned over time.

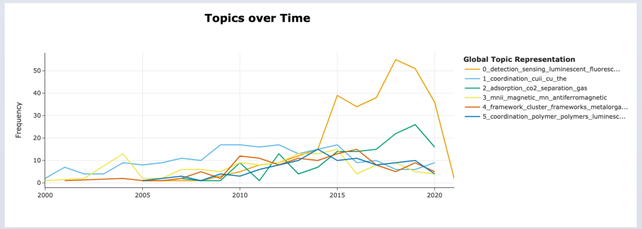

Figure 4. Research Topics Over Time

III. Extracting Synthesis Information from Unstructured Texts

Input Sentence:

Isonicotinic acid (2.00 g , 16.24 mmol) was put into a 250-mL , two - necked flask with toluene (100 mL) , under an atmosphere of nitrogen , and the mixture was stirred for ten minutes after addition of triethylamine (2.72 mL , 19.49 mmol) .

Output:

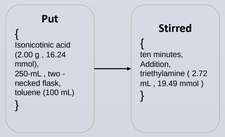

Figure 5. Output Design For Synthesis Extraction

We formulate the synthesis extraction as a two-step task: 1) named entity recognition and 2) relation extraction. As the first step, it is necessary to identify important entities (e.g., synthesis method, synthesis action, numerical descriptors and so forth) from unstructured text.

After extracting target entities, it is also important to recognize their relations – for example, which synthesis action is this numerical descriptor trying to describe? What is the sequence of synthesis procedures? The questions above matter to domain scientists. To address these questions, we perform relation extraction as the second step.

While the experiment is still ongoing, our output design is demonstrated at Figure 5.

Future Work

To advance our study in the future, we plan to integrate all extracted data into a knowledge graph, which accelerates knowledge search and reasoning for domain scientists.

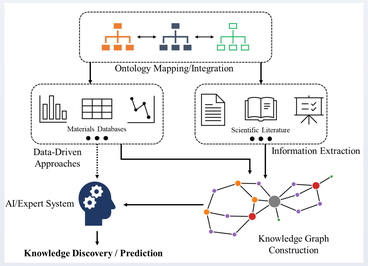

Figure 6. An Idea of Materials Knowledge Discovery and Integration

Another thing interested us is the set of existing materials ontologies – recently, there are increasing number of ontologies uploaded to the platform called MatOnto. It is interesting to analyze the use of these ontologies in MOF area.

Figure 7. HIVE4MAT – An Ontology-based Indexing System

References

[1] Nandy, Aditya, et al. "MOFSimplify, machine learning models with extracted stability data of three thousand metal–organic frameworks." Scientific Data 9.1 (2022): 74.

[2] Gubsch, Kristian, et al. "DigiMOF: A Database of MOF Synthesis Information Generated via Text Mining." (2022).

[3] Luo, Yi, et al. "MOF synthesis prediction enabled by automatic data mining and machine learning." Angewandte Chemie International Edition 61.19 (2022): e202200242.

[4] Park, Hyunsoo, et al. "Mining Insights on Metal–Organic Framework Synthesis from Scientific Literature Texts." Journal of Chemical Information and Modeling 62.5 (2022): 1190-1198.

Acknowledgment

This work is supported by NSF ID4 Institute for Data Driven Dynamical Design.