Procedia Computer Science

Volume 80, 2016, Pages 386–397

ICCS 2016. The International Conference on Computational Science

A Hybrid Human-Computer Approach to the Extraction of Scientific Facts from the Literature

Roselyne B. Tchoua[1], Kyle Chard[2], Debra Audus[3], Jian Qin[4], Juan de Pablo[5], and Ian Foster[1,2,6]

[1] Department of Computer Science, The University of Chicago, Chicago, IL, USA
roselyne@uchicago.edu
[2] The Computation Institute, The University of Chicago and Argonne, Chicago, IL, USA
[3] The National Institute of Standards and Technology, Gaithersburg, MD, USA
[4] Department of Chemical Engineering, Stanford University, Stanford, CA 94305
[5] Institute for Molecular Engineering, The University of Chicago, Chicago, IL, USA
[6] Math and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA

Abstract

A wealth of valuable data is locked within the millions of research articles published each year. Reading and extracting pertinent information from those articles has become an unmanageable task for scientists. This problem hinders scientific progress by making it hard to build on results buried in literature. Moreover, these data are loosely structured, encoded in manuscripts of various formats, embedded in different content types, and are, in general, not machine accessible. We present a hybrid human-computer solution for semi-automatically extracting scientific facts from literature. This solution combines an automated discovery, download, and extraction phase with a semi-expert crowd assembled from students to extract specific scientific facts. To evaluate our approach we apply it to a challenging molecular engineering scenario, extraction of a polymer property: the Flory-Huggins interaction parameter. We demonstrate useful contributions to a comprehensive database of polymer properties.

Keywords: Crowdsourcing, Information Extraction, Classification, Flory-Huggins, Materials Science

1 Introduction

The amount of scientific literature published every year is growing at a prolific rate. Some studies count more than 28,000 scientific journals and 1.8 million articles published annually [19]. As a result, the amount of information (e.g., experimental results) embedded within the literature is overwhelming. It has become impractical for humans to read and extract pertinent information. This problem hinders the advancement of science, making it hard to build on existing results buried in the literature. It also makes it difficult to translate results into applications

 Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2016

doi:10.1016/j.procs.2016.05.338

and thus to produce valuable products. In materials science and chemistry, for example, difficulties discovering published materials properties directly affect the design of new materials [6]. Indeed, despite the many publications in this domain, the process of designing new materials is still one of trial and error. Access to a structured, queryable database of materials properties would facilitate the design and model validation of new substances, improving efficiency by enabling scientists and engineers to more quickly discover, query, and compare properties of existing compounds. At the very least, it would transform an avalanche of publications into a machine-accessible and human-consumable source of knowledge.

Historically, materials properties have been collected in human-curated review articles and handbooks (e.g., the *Physical Properties of Polymers Handbook* [7], the *Polymer Handbook* [18]). However, this approach is laborious and expensive, and thus such collections are published infrequently. We contend that a better approach is to leverage information extraction techniques to process thousands of papers and output structured content for human consumption. To this end, we have developed a semi-automated system, χDB, which, with moderate input from humans, can extract materials properties for the scientific community.

We initially target extraction of a fundamental thermodynamic property called the Flory-Huggins interaction (or χ) parameter, which characterizes the miscibility of polymer blends. We chose to work with this property as a test case as it is particularly challenging to extract, due to the fact that it is published in heterogeneous data formats (e.g., text, figures, tables) and is represented in several different temperature-dependent expressions. To address these challenges, we developed a workflow consisting of an automated Web information extraction phase followed by a crowdsourced curation phase. The output of this workflow is a high quality human- and machine-accessible *digital handbook* of polymer properties. We show that we are able, using only a small group of students, to create a high quality database of properties with more χ values than in other notable handbooks. We expect that our approach is likely also to work well for other materials properties and in other scientific domains.

The rest of this paper is organized as follows. Section 2 presents background information related to Flory-Huggins theory and polymer science. Section 3 discusses related approaches that support automated extraction. Section 4 describes the χDB architecture. Section 5 presents the data collected via crowdsourcing. Section 6 explores the application of machine learning algorithms to improve the automatic selection of χ-relevant publications. Finally, we conclude and discuss future work in Section 7.

2 Application Background

The initial focus of our work is the extraction of properties of particular polymers blends (e.g, χ parameter and glassification temperature). Although highly curated properties database exist for hard [8] and metallic [17] materials, no equivalent exists for polymers blends. However, there is a clear need for a trusted, up-to-date, and easily accessible databases of properties within the soft matter community.

Polymers are large molecules (macromolecules) composed of many repeating units. Since polymeric materials are both ubiquitous and typically consist of several polymeric components, which are generally incompatible, the χ parameter represents a key property in the design of next-generation materials. A database of χ values would allow researchers to make informed judgments as to which χ values and thermodynamic analysis to use when predicting and understanding the phase behavior of multi-component polymeric materials. However, while there are thousands of published χ parameters, there is little consensus regarding the values. Different measurement methods yield different values, and different groups have at times reported