Extending Ontologies in the Nanotechnology Domain using Topic Models and Formal Topical Concept Analysis on Unstructured Text

Huanyu Li, Rickard Armiento, and Patrick Lambrix

The Swedish e-Science Research Centre & Linköping University, Linköping, Sweden
firstname.lastname@liu.se **

Abstract. In the data-driven workflows in the materials science domain, much of the data and knowledge is stored in different heterogeneous data sources maintained by different groups. This leads to a reduced availability of the data and poor interoperability between systems in this domain. Ontology-based techniques are an important way to reduce these problems and a number of efforts have started. In this paper, we use a phrase-based topic model approach and formal topical concept analysis on unstructured text in this domain to suggest additional concepts and axioms for the ontology that should be validated by a domain expert.

1 Introduction

More and more researchers in materials science have realized that data-driven techniques could accelerate the discovery and design of materials. Therefore, a large number of research groups and communities have developed data-driven workflows including data repositories (for an overview see [4]) and data analytics tools for particular purposes. Taking nanotechnology as an example, [6] states that there exists a gap between data generation and shared data access. The domain lacks standards for collecting and systematically representing nanomaterial properties. To solve these challenges, it is proposed that ontologies and ontology-based techniques can play a significant role in the data-driven materials science as ontologies provide a formal and explicit representation of knowledge of a domain which will enable reproduction, sharing and integration of data.

However, developing ontologies is not an easy task and often the resulting ontologies are not complete. In addition to being problematic for the correct modelling of a domain, such incomplete ontologies also influence the quality of semantically-enabled applications such as ontology-based search and data integration. Incomplete ontologies when used in semantically-enabled applications can lead to valid conclusions being missed. For instance, in ontology-based search, queries are refined and expanded by moving up and down the hierarchy of concepts. Incomplete structure in ontologies influences the quality of the search results.

We propose a novel method for extending existing ontologies by detecting new concepts that should be included in the ontologies. We do this by presenting an approach, formal topical concept analysis, that integrates a variant of topic modeling and formal concept analysis.

2 Approach

Our approach for extending ontologies, shown in Fig. 1, contains the following steps. In the first step *creation of a phrase-based topic model* documents related to the domain of interest are used to create topics (upper part in Fig. 1). We use the phrases-based topic model in the ToPMine system [1]. Given a corpus of documents and the number of requested topics, representations of latent topics in the documents are computed. The phrases as well as the topics are suggestions that a domain expert should validate or interpret and relate to concepts in the ontology. In the second step the (possibly validated and updated) topics are used in a *formal topical concept analysis* which returns suggestions to the domain expert regarding relations between topics and thus concepts in the ontology. We define a new variant of formal concept analysis (e.g., [2]) and use this new variant on topics (lower part in Fig. 1). These topics can come directly from the previous step or can be a modified version of the topics of the previous step, where non-relevant topics or phrases are removed. Both steps lead to the addition of new concepts and (subsumption) axioms to the ontology.

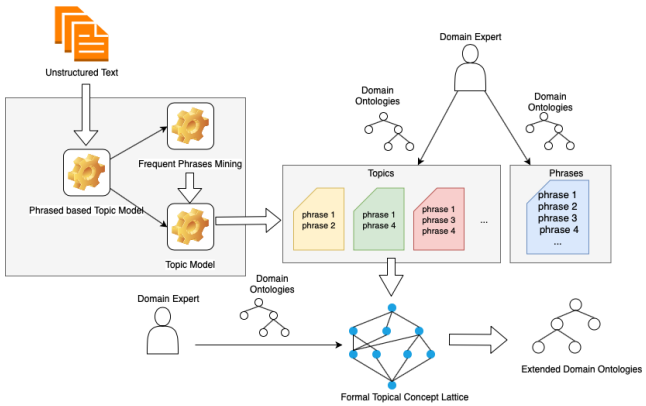

Fig. 1. Approach.

As shown in Fig. 1, a domain expert is involved in the different steps in our approach to validate and interpret the results of the phrase-based topic model and the formal topical concept analysis in terms of **interpreting all phrases** appearing in all topics, **interpreting topics** using the representative phrases or a subset of representative phrases in a topic. The outcome can be divided into

Table 1. Results

	label	Nanoparticle	eNanoMapper
# of general concepts	No-g	14	12
# of existing concepts	EXIST(-m)	46	52
# of new concepts	ADD(-m)	35	32
# of new axioms	ADD(-m)	42	37
# of queries	Q	49	49
# of influenced concepts	-	72	37

existing knowledge (EXIST(-m)), and new knowledge ADD(-m), where the '-m' qualifier denotes that a modified version of the label exists or should be added (e.g., *core-shell nanoparticle* for *core shell*). Further, the label could refer to a too general concept for the ontology (No-g) or not represent relevant knowledge (No). For the topic interpretation there is an additional outcome (Q) referring to too specific concepts for the ontology, but such concepts could be defined using concepts in the ontologies and OWL constructs. Finally, the domain expert **interprets the formal concept analysis-based lattice** and may find new concepts and subsumption axioms.

3 Experiments and Results

Data and Experiments. The corpus that we use is based on reports on nanoparticles from the Nanoparticle Information Library (`http://nanoparticlelibrary.net`). For each nanoparticle report, we take the text in 'Research Abstract' as well as the abstracts (or only the titles if there is no abstract) from the publications in 'Related Publications'. The final corpus contains 627 abstracts (or titles). The ontologies that we extend are the Nanoparticle ontology [5] (1904 concepts and 81 relations) and the eNanoMapper ontology [3] (12,531 concepts and 4 relations). Both ontologies are available via BioPortal (`https://bioportal.bioontology.org/`).

Results and discussion of results. In Table 1 we show the results regarding the interpretation of phrases, topics and lattice nodes from the experiments.

We show the number of general concepts (No-g), existing concepts (EXIST(-m)), new concepts (ADD(-m)), new axioms (ADD(-m)), queries (Q) and influenced concepts by adding the new axioms. Our results showed that the approach generated many EXIST(-m) cases. This provides a sanity check for our approach as it shows that existing concepts can be found. Further, the approach found 35 and 32 new concepts for the NanoParticle ontology and the eNanoMapper ontology respectively, as well as 42 and 37 new axioms. In addition to the new concepts and new axioms, also other concepts are influenced. Indeed, for a new axiom A is-a B, the sub-concepts of A receive B and all its super-concepts as its super-concepts (and thus inherit their properties), and all super-concepts of B receive A and its sub-concepts as sub-concepts (and thus all instances of these concepts are also instances of B and its super-concepts). In this experiment, 72 concepts from NanoParticle ontology were influenced by the new axioms. Therefore, the quality of NanoParticle ontology-enabled applications is improved whenever one

of the 35 new or 72 influenced concepts is used. For the eNanoMapper ontology the number of influenced existing concepts by adding new axioms is 37.

For the experiments we have currently used few resources, i.e. circa 600 abstracts and less than 10 hours for each of three experts (a domain expert and two knowledge engineering experts). As the quality of the ontologies and their use is raised, it is clear that the effort for extending the ontologies was worth-while.

4 Conclusions

In this paper we have used a phrase-based topic model approach and our own variant of formal concept analysis for extending ontologies. A domain expert interprets the results which are phrases, topics and a lattice. This leads to the confirmation of ontological concepts (EXIST(-m)) or to the addition of new concepts and axioms (ADD(-m)). The latter is the actual extension of the ontologies. Also, concepts from more general or other domains may be found, as well as very specific concepts in the domain that need not be added to the ontology. We have shown the usefulness of the approach by extending two ontologies in the nanotechnology domain using approximately 600 abstracts.

One issue that the domain expert noted was that it was not always easy to decide which level of granularity to use during the interpretation. In the future we will investigate how to help the domain expert dealing with this issue. In particular, the lattice appears to help refining topics into concepts that are more general and meaningful in the domain. This may be a useful step forward towards a higher level of automation in the process of extracting ontology information out of unstructured text. Furthermore, we will investigate the scalability of our approach by experimenting with more documents. Another possible direction is to investigate synergy possibilities between the topics and the ontology concepts, e.g., by using the ontologies to generate the corpora, or by iterating between topic generation and interpretation.

References

1. El-Kishky, A., Song, Y., Wang, C., Voss, C.R., Han, J.: Scalable topical phrase mining from text corpora. Proc VLDB Endowment **8**, 305–316 (2014)
2. Ganter, B., Wille, R.: Formal concept analysis: mathematical foundations. Springer Science & Business Media (2012)
3. Hastings, J., Jeliazkova, N., Owen, G., Tsiliki, G., Munteanu, C.R., Steinbeck, C., Willighagen, E.: eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. Journal of Biomedical Semantics **6**, 10:1–10:15 (2015)
4. Lambrix, P., Armiento, R., Delin, A., Li, H.: Big semantic data processing in the materials design domain. In: Encyclopedia of Big Data Technologies (2019)
5. Thomas, D.G., Pappu, R.V., Baker, N.A.: Nanoparticle ontology for cancer nanotechnology research. Journal of Biomedical Informatics **44**, 59–74 (2011)
6. Tropsha, A., Mills, K.C., Hickey, A.J.: Reproducibility, sharing and progress in nanomaterial databases. Nature nanotechnology **12**, 1111–1114 (2017)