Advanced Engineering Informatics 30 (2016) 687–702

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Full length article

MMOY: Towards deriving a metallic materials ontology from Yago

Xiaoming Zhang [a], Dongyu Pan [a,*], Chongchong Zhao [b], Kai Li [a]

[a] School of Information Science and Engineering, Hebei University of Science and Technology, 26 Yuxiang Street, Shijiazhuang, Hebei 050018, China
[b] School of Computer and Communication Engineering, University of Science and Technology Beijing, 30 Xueyuan Road, Beijing 100083, China

ARTICLE INFO

Article history:
Received 21 December 2015
Received in revised form 3 September 2016
Accepted 11 September 2016

Keywords:
Metallic materials ontology
Yago
Linked Open Data
knowledge base
rule
Materials science and engineering

ABSTRACT

In recent years, materials informatics has been getting more and more attention, and knowledge base plays a critical role for intelligent applications of materials science and engineering. The emerged materials ontologies, most of which are designed manually, provide useful but relatively limited materials knowledge. In contrast, Linked Open Data (LOD) provides some huge open knowledge bases in which lots of materials knowledge is hidden. Therefore, we try to derive an ontology about metallic materials domain from Yago which is a well-known open knowledge base. In this paper, an approach is proposed to generate metallic materials ontology based on the structure of Yago and string matching algorithm. First, we define keywords to initially obtain some metallic materials concepts from Yago by string matching algorithm. Then, based on these obtained concepts, the metallic materials knowledge and related knowledge is extracted according to the structural features of Yago, so as to acquire a metallic materials ontology (named MMOY). The detailed extraction strategy is elaborated as a set of logic rules designed for our approach. The approach is evaluated in light of F1-measure and time performance, and the experimental results demonstrate that the proposed approach can extract metallic materials knowledge and related knowledge from Yago effectively, and the time performance is acceptable. In addition, we have developed a prototype system to visually demonstrate the knowledge structure of MMOY.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

With the development of materials informatics, knowledge base plays a critical role for intelligent applications of materials science and engineering, which could accelerate the design and discover of materials, and meanwhile reduce commercialization cycle for new materials [1–3]. At present, the available volume of knowledge in materials science domain is rapidly growing in different types, which makes it possible to generate knowledge base of materials science by integrating this knowledge in different types. Recently, ontologies and semantic web technologies [4] are used widely in various domains, and semantic representation has become one of the important ways for information integration. So far, in materials domain there have emerged some ontologies [5], e.g. PREMΛP ontology [6] and FreeClassOWL [7], which can provide abstract model of materials science and engineering. However, most of them are designed manually, and provide useful but relatively limited materials knowledge. In contrast, in recent years, Linked Open Data (LOD) [8] cloud has increased significantly and many linked data and ontologies are published in various domains, such as, Yago [9–11], DBpedia [12] and Freebase [13] which are comprised of knowledge spanning varies domains. These huge datasets acting as open knowledge bases contain a lot of materials concepts, and some of them also provide complete knowledge structure that makes it possible for us to generate materials ontology by extracting the knowledge from these huge datasets. DBpeida and Freebase contain a large amount of metallic materials instances, but the classes are relatively less, while Yago contains much more classes than DBpedia and Freebase in metallic materials domain. In addition, the accuracy of Yago has been manually evaluated and the correctness is greater than 95%, which provides a reliable guarantee on data quality. Thus, in this paper, we extract the materials knowledge as well as the related knowledge, especially in metallic materials domain, from Yago, so as to generate a metallic materials ontology (named MMOY).

Yago is a huge dataset which is comprised of more than 10 million entities (e.g., persons, organizations, cities) as well as more than 120 million facts about these entities, and it classifies these entities to more than 350,000 classes [14] by the taxonomies of WordNet [15] and Wikipedia [16] category system. Therefore, Yago contains not only the knowledge structure of metallic materials based on the taxonomies of WordNet and Wikipedia, but also a

* Corresponding author.
E-mail addresses: zxmhebust@163.com (X. Zhang), pdyhebust@163.com (D. Pan), zhaocc@ustb.edu.cn (C. Zhao), mengyao_lk@163.com (K. Li).

http://dx.doi.org/10.1016/j.aei.2016.09.002
1474-0346/© 2016 Elsevier Ltd. All rights reserved.

large amount of metallic materials concepts, for example, alloy, copper and steel. In addition, the basic properties (e.g., chemical, physical, electrical and mechanical property) and the process technologies (e.g., heat treatment) are also contained in Yago. Meanwhile, it also consists of some related entities (e.g., auto company and metalware).

The motivation of our work is to use Yago to facilitate the semantic integration in metallic materials domain. As we know, there are various ways available to access Yago, e.g. SPARQL endpoint, graph browser and thematic dumps. However, there may be still some inconvenience in some cases for the domain-specific semantic integration, if we access Yago directly by SPARQL. For instance, if we measure the similarity of two concepts, we might usually want to know about the paths between the two concepts in Yago to support the computation of the distance of them. Therefore, dumping the required domain knowledge from Yago may be better for our requirement. Nevertheless, there is not appropriate theme available for dumping metallic materials knowledge (together with the related knowledge) from Yago, so we cannot use thematic dump to extract metallic materials knowledge directly. If SPARQL endpoint is used to extract metallic materials knowledge and related knowledge, we should design a large number of SPARQL query statements and analyze results one by one, and meanwhile we also need build the MMOY manually. Obviously, it's a hard work. Hence, we try to design an extraction strategy to generate a metallic materials ontology from Yago (MMOY), which can act as a domain background knowledge for the semantic applications in metallic materials domain.

However, there exist many difficulties in the process of extracting metallic materials knowledge, e.g. (1) the knowledge structure of metallic materials in Yago is implicit and (2) the naming of metallic materials concepts is unknown. Hence, in order to extract the knowledge from such a huge and complex dataset accurately, our approach combines similarity algorithm with the Yago structure [11]. The contributions of our work can be summarized as follows:

(1) An approach is proposed to derive MMOY from Yago. First, candidate keywords are defined and string matching algorithm is used to initially identify the metallic materials concepts in Yago. Then, based on the matching results, both hierarchical structure and non-hierarchical structure of Yago are used to acquire the domain knowledge structure as completely as possible, and the former is for metallic materials knowledge and the latter is for the related knowledge. In the proposed approach, just a small number of keywords are required, which can speed up the matching process, and taking full advantage of Yago structure makes up the limitation of the string matching strategy.
(2) In our proposed approach, a set of rules is designed to extract the metallic materials knowledge and related knowledge according to the features of Yago structure, and each rule is represented in predicate logic language.
(3) We evaluate our approach using precision, recall, F1-measure and time performance. The experimental results demonstrate that our method returns expected precision, recall and F1-measure. Furthermore, with the increasing of scale of the datasets, the time cost has not significantly increased. Thus, the proposed approach can extract the metallic materials knowledge and related knowledge from Yago effectively, and the time performance is acceptable.
(4) A prototype system is designed to visually display the knowledge structure of MMOY. In this system, the relations between concepts can be displayed obviously and users can have a better understanding of the metallic materials concepts and knowledge structure.

Although MMOY contains a lot of metallic materials concepts and a comparatively comprehensive knowledge structure, it lacks specific digital description for concepts. Based on these features, MMOY can be used in the following aspects. (1) In the materials community, lots of materials knowledge is hidden in non-structure and semi-structure data (e.g., text files and web pages), and due to the materials knowledge scatters in the natural language text, it is hard to exploit them. Since MMOY contains most of the terms in metallic materials domain as well as the relations between them, it can be utilized to facilitate the identification of high-value materials knowledge. (2) Traditional materials datasets (e.g., relational databases and Excel documents) focus more on the data description and value in a specific aspect, and the PSPP (Proces sing-structure-property-performance) linkages [17] are implicit or very weak. If MMOY is associated with traditional materials datasets, the relations in MMOY can give the materials datasets a semantic enrichment which can provide a better support for materials experts to do their research work such as materials selection [18]. Moreover as a domain ontology, MMOY's reasoning ability can help traditional datasets check consistency and discover new knowledge hidden in the datasets. (3) MMOY has a complete knowledge structure with rich hyponymy relations, so it could act as a background knowledge base in materials science domain to support materials ontology matching [19] which is the key point for resolving the heterogeneity across different materials information sources. In addition, (4) Due to lack of semantic information, traditional keyword-based search methods for metallic materials domain have their limitation. MMOY holds a domain-specific knowledge graph, so it can be used to expand user's query [20] so as to improve the retrieval effectiveness.

The rest of the paper is organized as follows: in Section 2 we discuss related work. Section 3 defines problems and introduces the approach and process in this paper. Following that Section 4 introduces detailed implementation method. In Section 5, the experimental evaluation is given and discussed. Section 6 describes a prototype system. Finally, Section 7 provides the conclusion.

2. Related work

For extracting or retrieving domain knowledge from LOD, the recognition of target concepts is one of the main challenges. Currently, there exist some approaches or strategies to identify target concepts. For example, Calegari and Pasi [21] use "bags of words" related to the users' interests to identify the similar entities in Yago with exact string matching and partial string matching so as to generate purpose ontology. In KFM [22], the string matching scores of predicate sets are used to find similar domain knowledge in LOD cloud. Indeed, there are many classical string similarity algorithms (e.g. n-gram [23], Jaccard [24] and SMOA [25]) which are usually used to define the relation between two concepts. The choice of the string similarity algorithms is based on the features and the application requirements [26]. n-gram is usually used to extract features of large datasets. Based on this feature, Sanchez-Pi et al. [27] utilize n-gram algorithm to generate concept synonyms list for classification improvement. Jaccard is a statistical algorithm for calculating similarity of sample sets. Harispe et al. [28] use Jaccard algorithm to define the relations between the biomedical concepts which are usually compound words. Additionally, SMOA algorithm reduces the influence of different substrings on the similarity calculation. Hu et al. [29] employ SMOA to generate a set of candidate anchors before ontology matching. The precision of string similarity algorithms is relatively low because most of them focus on the string structure rather than the sense.

In addition, taking advantage of the external datasets is another important way to recognize domain concepts. LOD, WordNet and

Download English Version:

https://daneshyari.com/en/article/6679626

Download Persian Version:

https://daneshyari.com/article/6679626

Daneshyari.com