Similarity-based Ontology Mapping in Material Science Domain

Wei Lin, Changjun Hu, Yang Li and Xin Cheng
School of Computer and Communication Engineering, University of Science and Technology Beijing, No.30 Xueyuan Road, Haidian District, Beijing, China

Keywords: Ontology, Mapping, Similarity Measure, Material Scientific Data.

Abstract: How to accurately retrieve data for users from massive, distributed and relational complex material databases is a major challenge in the domain of material science. Ontology mapping is regarded as a solution provider to the problem addressed. The number of material ontologies that are publicly available and accessible increases dramatically, so does the need for establishing semantic mapping among them to ensure interoperability. In this paper, we proposed a compositive similarity measure for ontology mapping. The material ontologies are generated from relational databases schemas based on rules. Then they are compared from concept name, structure and individuals. Finally, we describe a set of experiments on material science domain and show that our method propose highly accurate ontology mapping.

1 INTRODUCTION

The specialities of scientific data are the barrier of data sharing. Generally, scientific data has the following features: 1) Massive: China's scientific data size has reached PB scale. 2) Distribution: different types of data locate in heterogeneous databases. 3) Relevance: the relevance among scientific data is complex.

Data retrieval from massive, distributed and relational complex scientific databases remains an issue: 1) Traditional keyword search is unsatisfactory as it only considers the literally matching of data. 2) Efficiency and precision of data retrieval is extremely low due to distribution of material scientific data.

Accordingly, semantic retrieval based on ontology mapping is recommended. However, ontology research is inadequacy in scientific domain due to the mentioned features of scientific data. In this paper, we propose a compositive similarity measure for ontology mapping. The material ontologies are generated from relational databases schemas at first. Then the similarity of concept name, structure and individuals between different ontologies are calculated separately and composited to a compositive similarity. Finally, we prove in experiments that the compositive similarity measure achieves high precision ratio and recall ratio and is more appropriate for ontology mapping in the domain of material science.

The remainder of this paper is structured as follows. Section 2 will provide the reader with related research while section 3 will explain the details of our proposed measure. The experimental data and analysis will be presented in section 4. Finally, the conclusion of this paper and remarks on future work are given in section 5.

2 RELATED RESEARCH

At present, there are quite a lot ontology mapping systems. Doan et al. (2003) classified the individuals of concepts based on machine learning method and implemented GLUE ontology mapping system. Ehrig and Staab (2004) optimized existing ontology mapping measures and realized a quick ontology mapping system QOM. OLA ontology mapping system was developed specifically for OWL ontologies (Euzenat and Valtchev, 2004). Falcon-AO ontology mapping system was implemented by Hu et al. (2005) based on a graph matching algorithm. These are powerful systems, but suffer from a few drawbacks. The QOM system only computes the similarity of a restricted subset of candidate concept pairs. It's considered as a way to tradeoff between quality and efficiency of the mapping generation algorithms. GLUE system doesn't take into account structural information of

Lin W., Hu C., Li Y. and Cheng X..
Similarity-based Ontology Mapping in Material Science Domain.
DOI: 10.5220/0004350902650269
In *Proceedings of the 9th International Conference on Web Information Systems and Technologies* (WEBIST-2013), pages 265-269
ISBN: 978-989-8565-54-9

ontology concepts while OLA system and Falcon-AO system ignore the influence of individual information. Since results of all these systems are one-sided, a similarity measure with better performance is needed.

Concept similarity computing directly relates to both quality and reliability of the mapping process. Name similarity computing of ontology concepts is relevance to word similarity computing. A lot of scholars have done much work on word similarity computing based on semantic dictionary. Budanitsky and Hirst (2006) evaluated five lexical semantic relatedness measures by comparing their performance in detecting and correcting real-word spelling errors. Mahapatra et al. (2010) computed similarity of two sentences by calculating a weighted sum of the word similarity and the semantic similarity. All of the measures use WordNet as their central resource.

All these above measures are just for word similarity computing. Concept similarity computing of ontologies should also consider other affecting factors such as concept structure and concept individual.

3 METHODOLOGY

According to the features and requirements of material scientific data, we constructed ontologies from relational databases schemas. A compositive measure is then proposed for ontology mapping based on their similarity.

3.1 Mapping from Databases to Ontologies

We define five rules to transform relational databases to OWL ontology base on the features of material scientific data and OWL description language (Cullot et al., 2007).

Rule1: join-tables in the databases convert into object properties (OWL:ObjectProperty).
Rule2: other tables convert into classes or subclasses (OWL:Class or OWL:SubClass).
Rule3: referential constraints of tables convert into object properties (OWL:ObjectProperty).
Rule4: columns of tables convert into data properties (OWL:DataProperty).
Rule5: rows of tables convert into individuals. (OWL:Individual).

Figure1 shows the specific rules.

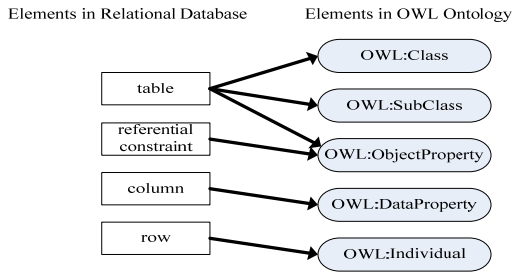

Figure 1: Transformation from databases to OWL ontology.

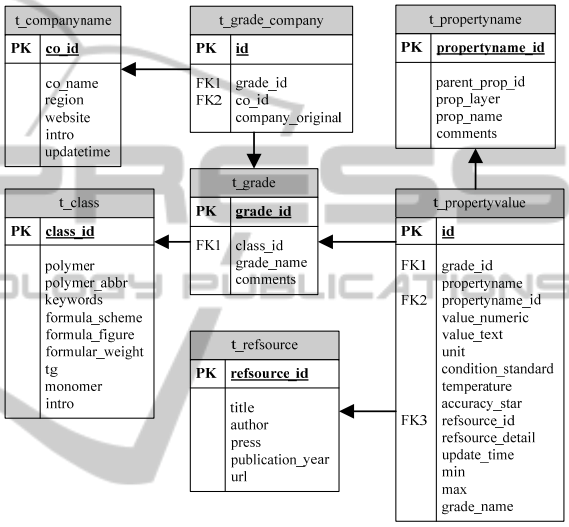

Figure 2: Relations of tables.

For example, with table t_grade in the logical structure shown in figure2, our transformation is as follows: Convert table t_grade into OWL class t_grade; Convert foreign key constraint FK_class_id into object property class_id; Columns of table t_grade would be converted into data properties and each row of table t_grade would be converted into individuals.

3.2 Similarity Computing

Similarity computing is the direct basis for creating semantic relations between concepts. At present, similarity computing mainly based on concept name, definition, structure and so on. Since each measure has its limitations, the result of each single measure is one-sided and can not fully reflect the relations among concepts. For material ontologies, concept name can reflect information of the concept to a certain extent. For example, concept name 'thermoplastic' means a kind of plastic material. Likewise, concept structure reflects the hierarchical

information of ontology. For instance, if a material's parent concept is plastic, we can infer that it must be a branch of plastic. Each individual represents a specific material in material ontologies. Therefore, we conclude a compositive measure according to the above-mention features of material ontologies.

3.2.1 Name Similarity

In traditional similarity measures, name similarity depends on the matching degree of name string. However, this measure only considers literally matching when calculating similarity.

Therefore, similarity computing based on semantic dictionaries improves accuracy of similarity. In general, for two given concepts, the shorter the path in WordNet tree, the higher the similarity. However, the depth of a concept in the hierarchy tree is ignored.

According to the features of material ontologies and word organization rules of WordNet, we propose a similarity measure based on WordNet. Both depth and distance of concept pairs are considered in this measure, the formula is shown as the following:

$$\mathrm{Sim_n}(C1, C2) = \frac{2 * \mathrm{dep}(\mathrm{com}(C1, C2))}{2 * \mathrm{dep}(\mathrm{com}(C1, C2)) + \mathrm{dist}(C1, C2)} \quad (1)$$

Note that com(C1,C2) represents the least common node of C1 and C2, dep(com(C1,C2)) represents the depth of com(C1,C2) in the hierarchy tree, dist(C1,C2)represents the shortest distance of C1 and C2 in the hierarchy tree.

3.2.2 Structural Similarity

For material ontologies, concept structure reflects the hierarchical information of ontology. For two concepts in material ontologies, if their parent concept is similar, they are similar to a certain extent. Likewise, if their son concept or brother concept is similar, they are possibly similar. The formula is shown as the following for calculating structural similarity:

$$Sim_s(C1, C2) = \frac{\omega_1 Sim_p(C1, C2) + \omega_2 Sim_{so}(C1, C2) + \omega_3 Sim_b(C1, C2)}{\omega_1 + \omega_2 + \omega_3} \quad (2)$$

Note that Sim_p(C1,C2) represents similarity of C1's parent nodes and C2's parent nodes and is calculated via Equation (1).

Sim_{so}(C1,C2) represents similarity of C1's son nodes and C2's son nodes. For concept C1 and C2 and their son concepts {C11,C12…,C1i,…C1m}and {C21,C22…,C2i,…C2n}, we calculated son similarity with the following steps:

1). For each son concept of C2, calculates the similarity with each son concept of C1 and records as Sim(C1i,C2j) ;

2). For 0<j<=n, set Sim(C1i,C2) = max[Sim(C1i,C2j)];

3). For 0<i<=m, set $\mathrm{Sim_{so}}(C1,C2) = \frac{\sum_{0<i\leq m} \mathrm{Sim}(C1i,C2)}{m}$.

Sim_b(C1,C2) represents similarity of C1's brother node and C2's brother node. The calculation of Sim_b(C1,C2) is similar to that of Sim_{so}(C1,C2).

ω 1, ω2 and ω3 are calculated via the sigmod function σ(x) =1/(1+e^(-5(x-α))) separately (Tang et al., 2006).

Note that X represents similarity of each measure and α represents the central point of sigmod function. In this experiment, we set α=0.5.

3.2.3 Individual Similarity

Each individual in material ontologies represents a specific material. In general, individual similarity computing based on the following rules: If the individuals of two concepts are similar, the concepts are similar. The individual similarity computing based on joint distribution is also available for ontology mapping in material domain (Doan et al., 2003).

The formula is shown as follows:

$$\mathrm{Sim}_i(C1, C2) = \frac{P(c1, c2)}{P(c1, \overline{c2}) + P(\overline{c1}, c2) + P(c1, c2)} \quad (3)$$

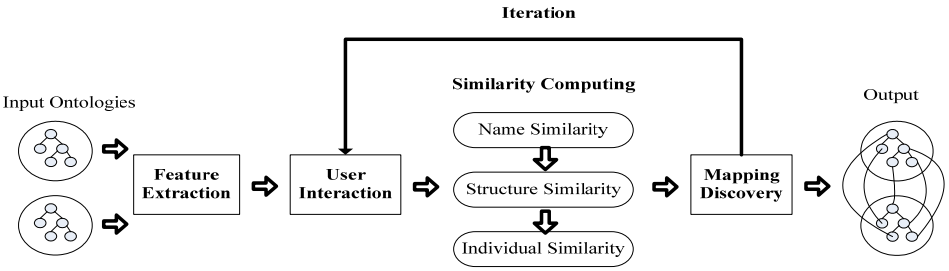

Figure 3: Ontology mapping process.

Note that $P(c1, c2)$ represents the probability of individuals belong to both C1 and C2. $P(c1, \overline{c2})$ represents the probability of individuals belong to C1 but not to C2. $P(\overline{c1}, c2)$ represents the probability of individuals belong to C2 but not to C1.

3.2.4 Compositive Similarity

Based on the three calculations of similarity, we propose the compositive similarity measurement:

$$\mathrm{Sim}(C1, C2) = \frac{\omega_1 Sim_n(C1, C2) + \omega_2 Sim_s(C1, C2) + \omega_3 Sim_i(C1, C2)}{\omega_1 + \omega_2 + \omega_3} \quad (4)$$

ω 1, ω2 and ω3 are calculated via the sigmod function mentioned in section 3.1.2.

3.3 Ontology Mapping

Ontology mapping is the process to create semantic relations between source ontology and target ontology. Figure 3 shows the iterative process of ontology mapping which mainly includes feature extraction, user interaction, similarity computing and mapping discovery. Figure 3 shows the mapping process.

Feature extraction: We utilized Jena API to handle OWL ontologies and extracted information such as concepts, individuals and relations from two heterogeneous material ontologies.

User interaction: Some concept pairs are mapped manually by users before the mapping process. The fault mappings will be corrected or deleted after the mapping process.

Similarity computing: Similarities of concept pairs are calculated by the compositive measure.

Mapping discovery: The threshold is set to 0.7. The concept pairs with similarity greater than 0.7 will be mapped.

Iteration: The accuracy of ontology mapping will be improved by the iteration of the process.

4 EXPERIMENTAL EVALUATION

According to the mentioned measure, we realized the transformation from relational databases to ontologies in material domain in our experiment. Ontology mapping was done based on the proposed compositive similarity measure. In this section, the effect of the compositive measure would be validated compare to single measure.

Table 1: Ontology information.

Ontology Name	Concepts	Properties	Individuals
plastic	26	15	204
rubber	16	15	182

Table 1 shows the information of ontologies generated from relational databases.

Our development tool is MyEclipse 6.0.1. Jena API and JWI are used in the program. Part of compositive similarity is shown in Table 2.

The concept pairs whose compositive similarity greater than 0.7 would be mapped.

An evaluation criterion is the mapping between plastic ontology and rubber ontology created manually. Recall ratio and precision ratio is used to evaluate experimental results. Recall ratio is defined as:

$$\text{Recall} = \frac{\text{right mappings found}}{\text{all standard mappings}} \quad (5)$$

Precision ratio is defined as:

$$\text{Precision} = \frac{\text{right mappings found}}{\text{all mappings found}} \quad (6)$$

Note that both precision ratio and recall ratio of compositive measure are higher than 90%, far exceed single measure. The experiment reveals that compositive measure does improve the uality of ontology mapping in material domain. Thus it can be

Table 2: Compositive Similarity.

	PMMA	polybutene	rubber	polyacrylonitrile
ABS_PMMA	0.9	0.42	0.42	0.22
polybutylene	0.3	0.84	0.56	0.53
polymer	0.33	0.52	0.57	0.42
acrylonitrile	0.27	0.49	0.55	0.72

Table 3: Experimental Result.

	Precision ratio (%)	Recall ratio (%)
Name Similarity	44.44	80
Structure Similarity	50	90
Individual Similarity	27.78	100
Compositive Similarity	94.44	94.44

seen that, compositive measure we proposed performs well in heterogeneous ontology mapping in material domain.ontology mapping in material domain.

5 CONCLUSIONS

The experiments show that compositive measure achieves higher precision ratio and recall ratio compared to single similarity measure. We can conclude that compositive similarity measure is suitable for heterogeneous ontology mapping in material domain. In addition, our work will be developed further in the future: In the Ontology building section： Rules should be improved in order to realize automatic mapping between relational database and OWL ontology in material domain. In the ontology mapping section：For the word which is not included in WordNet, word segmentation should be done before similarity computing. Firstly, in ontology construction section: rules should be improved in order to realize automatic mapping between relational database and OWL ontology in material domain. Secondly, in the ontology mapping section: the word which is not included in WordNet, word segmentation should be done before similarity computing.

ACKNOWLEDGEMENTS

Supported by the Key Science-Technology Plan of the National ‘Eleventh Five-Year-Plan’ of China under Grant No. 2011BAK08B04, the R&D Infrastructure and Facility Development Program under Grant No. 2005DKA32800, and the 2012 Ladder Plan Project of Beijing Key Laboratory of Knowledge Engineering for Materials Science under Grant No. Z121101002812005.

REFERENCES

Budanitsky A., Hirst G. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics, 32(1), 13-47.

Cullot N., Ghawi R., Yétongnon K. 2007. DB2OWL: A Tool for Automatic Database to Ontology Mapping In: Ceci et al., eds. Proceedings of the *15th Italian Symposium on Advanced Database Systems*, Torre Canne, Italy. SEBD, 491-494.

Doan A., Madhavan J., Dhamankar R., Domingos P., Halevy A. 2003. Learning to match ontologies on the Semantic Web. *The VLDB Journal*, 12(4) pp.303-319.

Ehrig, M., Staab, S. 2004. QOM - Quick Ontology Mapping In: S. A. Mcllraith, D. Plexousakis, F. V. Harmelen, eds. *3th International Semantic Web Conference*, Hiroshima, Japan, November 7-11, 2004. Berlin: Springer Berlin Heidelberg, 683-697.

Euzenat J, Valtchev P．2004. Similarity-based ontology alignment in OWL-lite In: R. Lopez and L. Saitta, eds. *Proceedings of the European Conference on Artificial Intelligence*, Valencia, Spain, August 22-27, 2004. Amsterdam: IOS Press, 333-337.

Hu W., Jian N. S., Qu Y. Z., Wang Y. B. 2005. GMO: A graph matching for ontologies In: Ashpole, Benjamin and Ehrig et al., eds. Proceedings *Of the KCAP 2005 workshop on Integrating Ontologies*, Canada, October, 2005. Germany, CEUR-WS.org, 43-50.

Mahapatra L., Mohan M., Khapra M. 2010. OWNS: Cross-lingual Word Sense Disambiguation Using Weighted Overlap Counts and Wordnet Based Similarity Measures In: Proceedings of the 5th *International Workshop on Semantic Evaluation, Sweden*, 2010. PA, USA: Association for Computational Linguistics Stroudsburg, 138-141.

Tang J., Li J. Z., Liang B. Y. 2006. Using Bayesian decision for ontology mapping. *Journal of Web Semantics, 4*(4), 243-262.