This version of the article has been accepted for publication, after peer review and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/ 10.1007/978-3-031-39141-5_17

Scoring Ontologies for Reuse: An Approach for Fitting Semantic Requirements *

Sabrina Azzi[1], Ali Assi[1], and Stéphane Gagnon[2]

[1] The House of Commons, Ottawa, Canada
sabrina.azzi@parl.gc.ca
ali.assi@parl.gc.ca
[2] Université du Québec en Outaouais, Gatineau, Canada
stephane.gagnon@uqo.ca

Abstract. The process of reusing ontologies is still challenging for the ontological community. One of the challenging efforts is to select the most relevant ontology from a set of candidates that needs a deep consideration. After the step of finding the candidates, many of them can be more appropriate than others as they fit better to the ontology requirements expressed by competency questions. First, we develop a mathematical formalisation based on Set Theory and we design the problem as an optimization problem to assist the knowledge engineer in selecting ontologies. Then, we provide formal steps to make well-founded comparison across a set of candidate ontologies. At last, we propose metrics to quantify the decision during the selection step.

Keywords: Competency questions · Ontology engineering · Ontology reuse · PNADO · Semantic requirements.

1 Introduction

An ontology is *a formal and explicit specification of a shared conceptualization*[1]. It consists of resources (see Section 3) that describe a domain of interest (*e.g.*, finance, parliament, pneumonia diagnosis, *etc.*).

Ontology building process includes the definition of its domain and its scope. Indeed, the knowledge engineer must identify the functional requirements that the ontology must satisfy. Therefore, Competency Questions (CQs) were first introduced in [2] for ontology engineering and formulated in natural language.

Ontology reuse is the process of selecting existing ontologies that meet some of the requirements defined in the conception [3]. This process significantly reduces the development time and effort.

Reusing ontologies is not a trivial process. It can be applied either in hard or soft mode. While, the former reuses the whole selected ontology. The latter imports only a set of its concepts. This process starts by selecting the candidate ontologies according to four main criteria [4]. These ontologies could be found in public repositories. Then, by pruning them to select the most appropriate ones.

* Supported by the House of Commons of Canada.

To choose the most relevant ontology, [5] presents an approach based on the notion of preference between candidate ontologies. However, the approach omits cases of reuse and which reuse to consider, *i.e*, soft or hard. As well, in the context of selecting foundational ontologies as starting points for ontology engineering, [6] implements a selection algorithm scoring ontologies by number of categories and end-user CQs they can satisfy. Then, the authors formulate a representative description along preset ontology attributes, such as Ontological Commitments. However, this approach presupposes that CQs are formulated within the scope of the ontology, and that ontology attributes are the end-users decision criteria.

In this paper, we propose an approach that facilitates the ontology reuse. This approach has been applied during the project about developing an ontology covering pneumonia diagnosis (PNADO)[7] and has proven its efficiency. In contrast to the existing approaches, ours permits to guide the knowledge engineer to select relevant ontologies for reuse. Indeed, it covers all the cases of reuse and recommend the corresponding mode. We develop a mathematical formalisation based on Set Theory and we design the problem as an optimization function. The novelty of our approach is to refine the decision criteria for ontology reuse objectives. To the best of our knowledge, there is no approach that leverages the decision criteria during ontology selection for reuse.

2 Related work

Selecting the appropriate ontology for reuse from a set of candidates is still challenging. Many works have recognized the challenge of making a well-founded comparison with a reasonable amount of effort [3, 8]. [9] affirms that choosing ontology for reuse is a highly subjective task carried out by experienced ontology engineers, who choose ontologies that intuitively respond to the requirements. Whether implemented as a tool or presented as a guideline, most works used some kind of calculation over a set of subjectively-weighted criteria conducting to approaches that are too superficial to account for the candidates' semantics.

Among the aspects affecting the reusability of an ontology is the perception of its quality according to some evaluation criteria. The only work that has tackled this challenge and addressed the assessment the candidates' semantics was done by Katsumi and colleagues [5]. This work introduced for the first time the notion of preference between ontologies and provided a definition that allows the knowledge engineer to make a well-founded comparison between ontologies, with respect to their semantic requirements. One of the major limitations of this work is not considering all the cases of comparison between candidate ontologies. Also, the approach does not provide any measures that allow to select accurately the ontologies.

In our approach, we consider more than two ontology candidates and compare them based on the semantic requirements of the ontology expressed by CQs. We take into account all the cases of reuse and recommend the corresponding mode. We provide a mathematical formalisation based on Set Theory and we design the problem as an optimization problem.

3 Preliminaries and definitions

In this section, we present preliminary definitions about ontologies and the order set between the candidate ontologies.

Definition 1. ***(Ontology).*** *An ontology is a set of triplets in the form of* `<subject, predicate, object>` $\in \mathbf{R} \times \mathbf{P} \times (\mathbf{R} \cup \mathbf{L})$ *where R is the set of resources, P is the set of predicates and L is the set of literals.*

In the following, we refer to the set of *relevant* concepts (*i.e.*, those that respond to the CQs) by *intended models*, denoted by R. This set is defined by a domain experts during the first phase of the ontology development.

Indeed, we refer to the set of candidate ontologies by $O = \{i : 1 \leq i \leq n\}$. An ontology $i \in O$ consists of a set concepts denoted by O_i. According to the CQs, we split the concepts for each candidate ontology $i \in O$ into two sets R_i and $\bar{R}_i$, *i.e.*, $O_i = R_i \cup \bar{R}_i$ where $R_i \cap \bar{R}_i = \emptyset$. R_i represents the set of concepts in i that also belongs to R, *i.e.*, $R_i \subseteq R$. Thus, R_i represents the relevant concepts in ontology i. Similarly, $\bar{R}_i$ consists of the remaining concepts in O_i *i.e.*, $O_i - R_i$. Thus, $\bar{R}_i \not\subset R$ which means that $R \cap \bar{R}_i = \emptyset$.

Based on R_i and $\bar{R}_i$, we measure the *meaningful* of each ontology $i \in O$ by the number of concepts in $R - R_i$ and $\bar{R}_i$. The former quantifies the concepts corresponding to the one CQ that are omitted in i. The latter measures the number concepts included in i which are not relevant to any CQ. We call such concepts by superfluous concepts or noisy concepts. The more the number of the omitted concepts in i is small, the more i covers the needed concepts defined by the CQs. Indeed, the more the number of the concepts in $\bar{R}_i = O_i - R_i$ is small, the more i does not include concepts related to any CQ.

Definition 2. ***(Optimal Ontology).*** *An ontology $j \in O$, where O is the set candidate ontologies, is called optimal if j verifies: $R \cap R_j = R$ (which means that $R_j = R$) and $\bar{R}_j = \emptyset$.*

We define the following objective function that globally scores the suitability of an ontology $i \in O$:

$$score(i) = |R - R_i| + |\bar{R}_i| \tag{1}$$

Our approach is motivated by the minimization problem:

$$\arg\min_{i^* \in O} score(i^*) = |R - R_{i^*}| + |\bar{R}_{i^*}| \tag{2}$$

where O is the set of candidate ontologies and $i^* \in O$ is the ontology that has the smallest meaningful *score* value.

Definition 3. ***(Meaningful Ontology).*** *We call an ontology $i \in O$ more meaningful than ontology j, denoted by $i \preceq j$, when $|R - R_i| \leq |R - R_j|$ and $|\bar{R}_i| \leq |\bar{R}_j|$ where O is the set of candidate ontologies.*

Definition 4. ***(Meaningfully Equivalent Ontology).*** *We call an ontology i is meaningfully equivalent to ontology j, denoted by $i = j$, iff $|R - R_i| = |R - R_j|$ and $|\bar{R}_i| = |\bar{R}_j|$.*

Lemma 1. *$\preceq$ is a partial ordering over the set of candidates O.*

Proof. We need to prove that $\preceq$ is reflexive, transitive and antisymmetric.

- $\preceq$ **is reflexive**. We must prove that $O_1 \preceq O_1$ where $O_1 \in O$. Since $O_1 \preceq O_1$, we have $|R - R_1| = |R - R_1|$ and $|\bar{R}_1| = |\bar{R}_1|$ where R_1 is the set of the relevant concepts in O_1. Hence, $\preceq$ is reflexive.
- $\preceq$ **is transitive**. Let $O_1 \in O,\ O_2 \in O$ and $O_3 \in O$. We need to prove that if $O_1 \preceq O_2$ and $O_2 \preceq O_3$ then $O_1 \preceq O_3$. Since $O_1 \preceq O_2$, we have $|R - R_1| \leq |R - R_2|$ and $|\bar{R}_1| \leq |\bar{R}_2|$. Similarly, since $O_2 \preceq O_3$ then $|R - R_2| \leq |R - R_3|$ and $|\bar{R}_2| \leq |\bar{R}_3|$. Therefore, $|R - R_1| \leq |R - R_3|$ and $|\bar{R}_1| \leq |\bar{R}_3|$. Hence, $O_1 \preceq O_3$.
- $\preceq$ **is antisymmetric**. We need to prove that if $O_1 \preceq O_2$ and $O_2 \preceq O_1$, then $O_1 = O_2$. Since $O_1 \preceq O_2$ and $O_2 \preceq O_1$ then $|R - R_1| \leq |R - R_2|$ and $|R - R_2| \leq |R - R_1|$ then $|R - R_1| = |R - R_2|$. Similarly, $|\bar{R}_1| \leq |\bar{R}_2|$ and $|\bar{R}_2| \leq |\bar{R}_1|$ then $|\bar{R}_2| = |\bar{R}_1|$. Therefore, $|R - R_1| = |R - R_2|$ and $|\bar{R}_1| = |\bar{R}_2|$. Hence, $O_1 = O_2$.

Based on the Lemma 1, the candidate ontologies in O are now comparable. Thus, the best ontology that can be used from O is simply the ontology that has the smallest meaningful score value. If two ontologies $i \in O$ and $j \in O$ posses the same meaningful score, we prioritize i over j and vice-versa. If i has more relevant concepts than j, *i.e.*, $|R_i| > |R_j|$, then i is more precise than j. Formally:

Definition 5. ***(Precise Ontology).*** *An ontology $i \in O$ is more precise than ontology $j \in O$ if for $score(i) = score(j)$ we have $|R - R_i| < |R - R_j|$ where O is the set candidate ontologies. We denote this relation between i and j by $i \rhd j$.*

Lemma 2. *$\rhd$ is a strict order over the set of candidate ontologies O.*

Proof. We need to prove that $\rhd$ is irreflexive, transitive and asymmetric.

- $\rhd$ **is irreflexive**. We must prove $i \rhd i$ does not hold for any i in O. Let say that $i \rhd i$ holds. That means $|R - R_i| < |R - R_i|$ which is incorrect since $R_i = R_i$. Therefore, $\rhd$ must be irreflexive.
- $\rhd$ **is transitive**. Let $O_1 \in O,\ O_2 \in O$ and $O_3 \in O$. We need to prove that if $O_1 \rhd O_2$ and $O_2 \rhd O_3$ then $O_1 \rhd O_3$. Since $O_1 \rhd O_2$, we have $|R - R_1| < |R - R_2|$. Similarity, since $O_2 \rhd O_3$, we have $|R - R_2| < |R - R_3|$. Thus, $|R - R_1| < |R - R_2| < |R - R_3|$. Therefore, $\rhd$ must be transitive.
- $\rhd$ **is asymmetric**. We need to prove that if $i \rhd j$, then $j \rhd i$ does not hold. Let consider that $i \rhd j$ (*i.e.*, $|R - R_i| < |R - R_j|$) and $j \rhd i$ (*i.e.*, $|R - R_j| < |R - R_i|$) hold for $i \neq j$. Thus, $|R - R_i| = |R - R_j|$. This equality leads to a contradictory since $|R - R_i| \neq |R - R_j|$ for $i \neq j$.

Algorithm 1 Selecting relevant ontologies

Input: $O = \{o_1, o_2 \ldots, o_n\}$ set of (n) candidate ontologies that respond to one QC.
R set of relevant concepts.
Output: $O_s = \{o_i, o_j \ldots, o_k\}$ set of ontologies $(m \leq n)$ to be reused.

1: $O_s \leftarrow \{\}$;
2: **for all** $o_i \in O$ **do**
3: $score(o_i) \leftarrow |R - R_{o_i}| + |\bar{R}_{o_i}|$
4: **end for**
5: Rank $o_i \in O$ by their meaningful $score(o_i)$.
6: Select the most meaningful ontologies $o_i^* \in O$ with the minimum $score(o_i^*)$.
7: $O_s \leftarrow \cup\{o_{i^*}\}$
8: **if** $|O_s| = 1$ **then**
9: **if** $R_{o_i} = 0$ **then**
10: **return** O_s ▷ Hard reuse of o_i
11: **else**
12: **return** O_s ▷ Soft reuse of o_i
13: **end if**
14: **else if** $|O_s| > 1$ **then** ▷ More than 1 ontology has the same score
15: $O_p \leftarrow \{\}$;
16: Rank $o_j \in O$ by their precise score $|R - R_{o_j}|$.
17: Select ontology o_j with the minimal precise value $|R - R_{o_j}|$.
18: $O_p \leftarrow \cup\{o_j\}$
19: **if** $|O_p| = 1$ **then**
20: **if** $R_{o_j} = 0$ **then**
21: **return** O_p ▷ Hard reuse of o_j
22: **else**
23: **return** O_p ▷ Soft reuse of o_j
24: **end if**
25: **end if**

4 Our approach

The proposed approach focuses on reuse of ontologies that best respond to some of semantic requirements or intended models. Its main idea is to use CQs to select only the ontologies that would be interesting for the knowledge engineer (*i.e.*, the relevant ontologies). In following, we present an algorithm that returns these ontologies based on Definitions 3 and 5. For each ontology chosen and according to its meaningfulness score, we recommend hard of soft reuse.

For each CQ, we calculate the score of its $o_j \in O$ candidate ontologies (line 3). Then, we rank the candidates according to their meaningful scores (line 5). Two cases are considered. The first one is where only one ontology is returned (line 8). In this case, the ontology is considered as the most meaningful ontology and it is selected for reuse. Then, its precision score is calculated. Based on this score, the hard or soft mode will be recommended. If is a precise ontology, *i.e*, no superfluous models then we recommend hard reuse. Otherwise, soft reuse is considered. The second case is where many ontologies are returned (line 14). In

this case, we rank them according to their precise score $|R - R_{o_j}|$ and we select the most precise ontology having the minimal precise score.

5 Conclusion

We presented in this paper a new formalized approach for selecting relevant ontologies to be reused during ontology development. The approach was drawn from lessons learned during the project of developing a pneumonia diagnosis ontology (PNADO). We developed a mathematical formalisation based on Set Theory and we designed the problem as an optimization problem. In this perspective, we proposed an efficient algorithm based on this formalisation. We also proposed metrics to quantify the decision during selection ontologies.

An important future work would be to implement this algorithm and evaluate it in real use cases. It also be interesting to propose an extension of this approach that leverages parallel computation frameworks to handle ontology reuse in big data era.

References

1. Rudi Studer, V Richard Benjamins, and Dieter Fensel. Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1-2):161–197, 1998.
2. Michael Grüninger and Mark S Fox. The role of competency questions in enterprise engineering. In *Benchmarking—Theory and practice*, pages 22–31. Springer, 1995.
3. Megan Katsumi and Michael Grüninger. What is ontology reuse? In *FOIS*, pages 9–22, 2016.
4. Sabrina Azzi. *Nouvelle méthodologie de construction d'ontologies médicales: cas d'étude: diagnostic de la pneumonie.* PhD thesis, Université du Québec en Outaouais, 2021.
5. Megan Katsumi and Michael Grüninger. Choosing ontologies for reuse. *Applied Ontology*, 12(3-4):195–221, 2017.
6. Zubeida Khan and C Maria Keet. Onset: Automated foundational ontology selection and explanation. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 237–251. Springer, 2012.
7. Sabrina Azzi, Wojtek Michalowski, and Michal Iglewski. Developing a pneumonia diagnosis ontology from multiple knowledge sources. *Health Informatics Journal*, 28(2):14604582221083850, 2022.
8. Reham Alharbi, Valentina Tamma, and Floriana Grasso. Characterising the gap between theory and practice of ontology reuse. In *Proceedings of the 11th on Knowledge Capture Conference*, pages 217–224, 2021.
9. G Cota et al. The landscape of ontology reuse approaches. *Appl. Practices Ontol. Des., Extraction, Reason*, 49:21, 2020.