ONTEM: Extraction based Alignment Method for Large Ontologies

Zarhouni Mourad[1] and Benslimane Sidi Mohammed[2]

[1] EEDIS Lab., University Djilali Liabes, Sidi Bel Abbes, Algeria
[2]Ecole Supérieure en Informatique, LabRiLaboratory, Sidi Bel Abbès, Algeria

mourad.zerhouni@univ-sba.dz
s.benslimane@esi-sba.dz

Abstract.
The rise of the semantic web and the development of different technologies allow different actors to access knowledge found in different ontologies. This is not always obvious because technical constraints such as data volume and execution time are determining factors in the choice of an alignment algorithm. Among the solutions for scaling, the extraction of ontological entities as well as for partitioning methods can be complementary to alignment techniques, given the reduction in the size of the ontologies to be aligned, and therefore the reduction in execution time. In this article, we propose a new alignment method based on the extraction of concepts and labels as well as the creation of correspondences in an automatic way. Indeed, we emphasize that this method does not require any calculation of similarity distance. The obtained results during the evaluation of our method show its effectiveness and can be a decisive turning point for the different existing alignment methods.

Keywords: Large Ontologies, Alignment, Extraction, Partitionnement.

1 Introduction

Nowadays, ontologies have become one of the most important research orientations, especially with the advent of the Semantic Web. An ontology is defined as the conceptualization of objects recognized as existing in a domain, their properties and the relationships between them [1]. They play a key role in annotating web pages or services by modelling the concepts, attributes and relationships used to annotate resource content. In many application contexts, several ontologies covering the same or related fields are developed independently of each other by different communities, which raises the issue of being able to exchange, integrate and transform data. At this stage, the problem of interoperability arises, allowing heterogeneous systems to communicate and cooperate, and to this end, semantic links must be established between entities belonging to two different ontologies, and the transition to the web is a real challenge that requires researchers to make efforts to optimize content management, which can be constantly enriched and developed. To this end, it is necessary to improve the quality of the organization, structuring, research, identification, access, use, reuse of resources, integration, and automated processing of this content. All alignment techniques are required to scale up to handle large ontologies [2, 3]. However, this is not always obvious because the creation of multiple ontologies, sometimes for the same domain, leads to a heterogeneity between the knowledge expressed within each of them that must be resolved: it is the problem of interoperability.

The objective of our work is to meet the challenge of scaling up alignment method [4]. In particular, we propose an algorithm to extract concepts and labels common to both ontologies for alignment purposes [5]. Our algorithm has been tested on the ontologies in the LargeBio_Track section of the OAEI_2018 campaign. Satisfactory results have been achieved.

This paper is organized as follows: Section 2 is a state of the art that presents the different alignment strategies and focuses on related work. In section 3, we describe our extraction based alignment method for large ontologies. Section 4 is an experimental study that illustrates the results and performance of our method. Section 5 presents a discussion of the results obtained. Finally, Section 6 concludes the document and provides an overview of the directions for future work.

2 Related Work

Alignment consists in determining the set of correspondences between two ontologies by using or implementing solutions to different heterogeneity problems. Several alignment techniques, based on different criteria, are currently proposed in the literature [6] provides a synthesis of alignment techniques.

The choice of one technique or another or the composition of several of them is not an easy task. Several studies complement their alignment results by using WordNet [7] as an external resource, and many alignment methods dedicated to ontologies have emerged in the last decade [8]. However, these methods are designed to align small ontologies. Partitioning [9] and modularization [10] are currently the two main strategies for breaking down large ontologies into blocks or ontology modules, respectively. These methods can only work if the number of concepts at the input of the alignment tool is limited.

One of the solutions for scaling involves the possibility of partitioning ontologies into blocks before performing alignment [11]. The partitioning strategy was proposed by [12] for partitioning into blocks of two large hierarchical classes.

There are several approaches to partitioning. Graph-based approach applies graph-based algorithms to decompose ontology [13]. Logic-based approach uses description logic to partition an ontology [14]. Clustering-based approach consists in creating a partition or a decomposition of this set into sub-parts (clusters) [15]. The modularization strategy was proposed by [16] to deal with large and complex dentistry. It breaks down the problem of large-scale matching into sub-problems by matching at the level of ontology modules [17].

3 ONTEM APPROACH

In this section, we propose an ONTology Extraction Method (ONTEM) based on an extraction strategy that is, to our knowledge, almost non-existent in ontology alignment work. The proposed method consists of four main steps: 1) Preprocessing, 2) Common entities identification, 3) Mapping generation, 4) Alignment generating. The general architecture of ONTEM is illustrated in Figure 1.

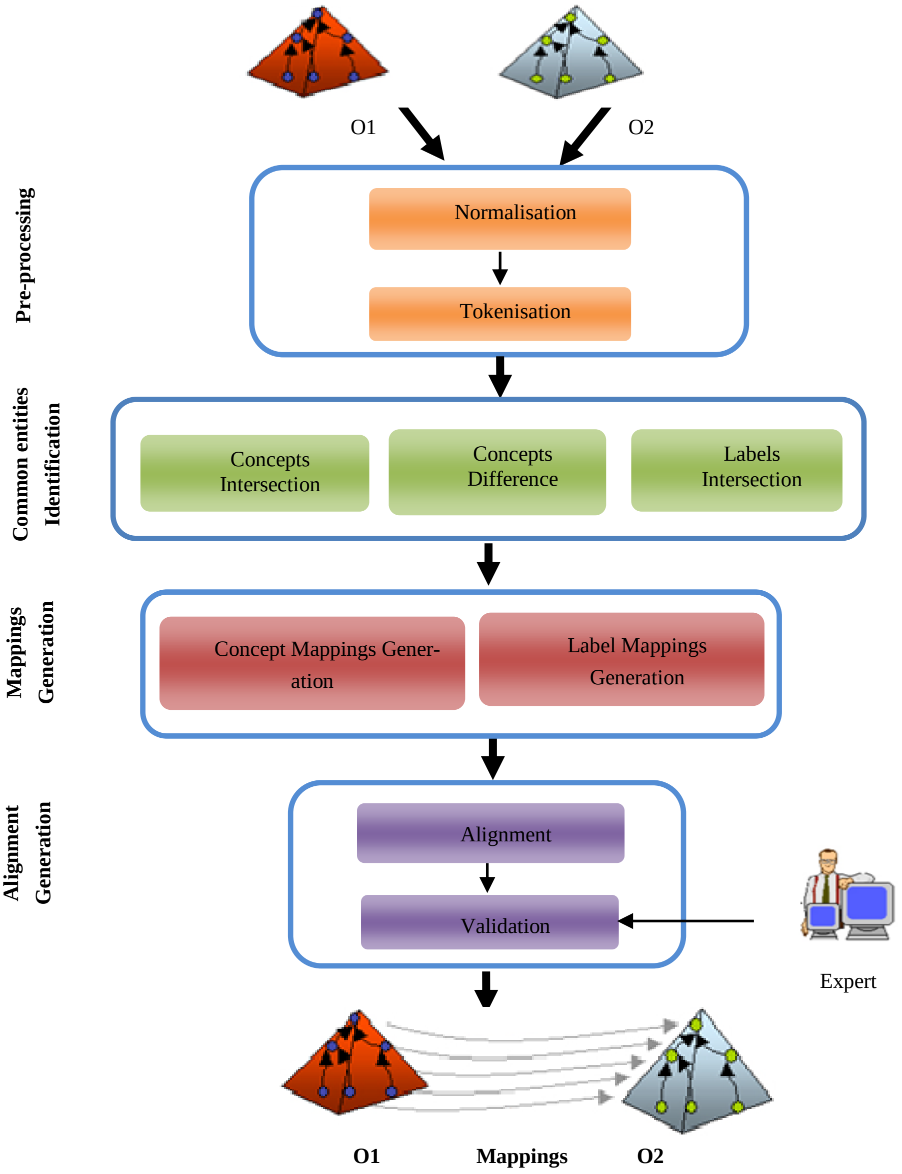

Figure1: Architecture of ONTEM

3.1 Preprocessing

To facilitate the process of comparing the ontological terms labelling classes and their properties (i.e. calculating the distances between their character strings), it is very important to perform a number of pre-processing operations. They significantly improve alignment results. In addition, when aligning based on synonym extraction, pre-processing operations facilitate their recognition by lexical databases and/or synonym dictionaries. The classes of an ontology are extracted after the conceptualization of the targeted domain according to the objectives to be achieved and the application that will use the ontology. We will call anchoring a concept of one ontology matched with a single concept of another ontology, of the same name and meaning as it.We are dealing with two types of anchor pairs: those obtained from the intersection of concepts and those obtained from common labels. The approach exploits the richness of the concept labels. Labels are made up of several words [18]. It assumes that similar concepts share part of their label. It is therefore more appropriate for aligning taxonomies whose concept names are expressions composed of several words because, in this case, these names may share words, which may reveal common points between the concepts concerned.
Two linguistic techniques used are. Normalisation which consists of : 1) transforming all the characters of the ontological terms into lower-case letters, 2) stripping of special characters, spaces, and numbers, 3) elimination of coordinating conjunctions, articles, prepositions, 4) elimination of words designating sets or elements (composition words). Tokenisation, which is a lexical analysis that consists in transforming a character stream into a token stream by an analyzer (Tokenizer) that recognizes punctuation, white characters, etc.

The pre-processing phase is illustrated by Algorithm1

Algorithm1: Preprocessing

```
Inputs:Ontology O
Outputs:List_of_Concepts, List_of_Labels, Index_of_Concepts, Index_of_Labels
Begin
For Each Concept of O do
    CN=Normalize(Concept)
    CT=Tokenize (CN)
    Add(Concept,CT)
    Add((Concept,CT),INDEX_CONCEPTS)
    For Each Label ofConcept
            LN=Normalize (Label)
            LT=Tokenize (LN)
            Add(Label,LT)
            Add((Label,LT),INDEX_LABELS)
    EndFor
EndFor
Return (List_of_Concepts, List_of_Labels, Index_of_Concepts, Index_of_Labels)
End
```

3.2 Common entities identification

Since the ontologies treated concern the same field, it is obvious that they share common elements. This observation leads us directly to consider the common elements to both ontologies. The elements common to both ontologies can obviously be concepts, labels, properties, as well as relationships. Given two voluminous ontologies, the objective is to match the concepts of the first ontology with the concepts of the second ontology. To do this, we will use the "Intersection" operation to determine the common concepts to both ontologies.

3.2.1 Intersection of common entities

Given a domain D and two ontologies O1ϵ D and O2ϵ D.

Let LCO1 and LCO2 the list of concepts of Ontology 1 and Ontology 2 respectively.

Let LLO1 and LLO2 the list of labels of Ontology 1 and Ontology 2 respectively.

The set LIC (List of Intersection of Concepts) will form the list of common concepts to both ontologies.

LIC=LCO1∩ LCO2.

In the same way, the set LIL (List of Intersection of Labels) will form the list of common labels to both ontologies.

LIL=LLO1∩ LLO2.

The intersection of common entities (concepts and labels) is performed by Algorithm2 for concepts and Algorithm3 for labels.

Algorithm2: Intersection of Concepts

```
Inputs: LCO1, LCO2
Outputs: LIC
Begin
LIC = Intersection (LCO1,LCO2)
Return (LIC)
End
```

Algorithm3: Intersection of Labels

```
Inputs: LLO1, LLO2
Outputs:  LIL
Begin
LIL = Intersection (LLO1,LLO2)
Return (LIL)
End
```

3.2.2 Difference of concepts

In this phase, we retain only the non-composed concepts. We will only select concepts that are not composed and do not belong to LIC. A compound concept being a name that contains at least one character" _".

Let LNCCO1 the list of non-composed concepts of Ontology 1.
Let LCSSO1, list of Concepts for Searching Synonyms of Ontology 1.
The difference on concepts is performed by Algorithm 4.

Algorithm4: Difference of concepts

```
Inputs: LNCCO1, LIC
Outputs:  LCSSO1
Begin
LCSSO1 = Difference (LNCCO1,LCI)
Return (LCSSO1)
End
```

3.3 Mapping generation

The mapping discovery process, called ontology alignment, is a function *f* that applies to two ontologies O1 and O2, with a set of parameters *p* (weights, thresholds, etc.) and a set of external resources *r*, and produces a set of mapping *A*.

A=f(O1,O2,p,r).

Alignment consists of several steps [19]: extracting the data to be reconciled, selecting the pairs of elements to be compared, calculating a similarity for each selected pair, deducing the alignment from the previously calculated similarity measurements. Each method of calculating a similarity measure corresponds to the execution of a particular alignment technique. Several classifications of these techniques have been proposed in the literature [20, 21, 22].We find that each of the concepts in the first ontology directly points to their corresponding concepts in the second ontology. The anchor pair is saved in the LCC list. The algorithm for generating direct concept mappings is represented by the Algorithm 5.

Algorithm5: Generating direct concept mappings

```
Inputs:   LIC, ICO1, ICO2
Outputs: LCC
Index_Mappings
Begin
For Each Concept€ LIC do
    Entity_Name_1=Read(Concept, ICO1)
    Entity_Name_2=Read(Concept,ICO2)
    Add (Entity_Name_1,LCC)
    Add(Entity_Name_2,LCC)
    Add (Entity_Name_1,Entity_Name_2, Index_Mappings)
EndFor
Return(LCC)
End
```

The algorithm for generating concept matches from the list of common labels is represented by the Algorithm 6.

Algorithm6: Generating mapping concepts based on common labels

```
Inputs :LIL, ILO1, ILO2, ICO2, Index_Mappings
Outputs :LLC, Index_Mappings /* Updated */
Begin
For Each (Label ϵ LIL) do
/* browse the labels of LIL */
   Entity_Name_1=Read(Label, ILO1)
   Entity_Name_2=Read(Label, ILO2)
Found=0
   While  (Entity_Name_2  hasValues Is True) && (Found==0)
/* Retrieving  Entity_name2 not yet used */
      IF (Entity_Name_1,Entity_Name_2) not found  in  Index_Mappings
           Add(Entity_Name_1,LLC)
           Add(Entity_Name_2,LLC)
            /* insert Entity_Name_1 and Entity_Name2 in LLC */
           Add(Entity_Name_1,Entity_Name_2, Index_Mappings)
            /*Update of the Index_Mappings */
            Found=1;
      Else
         Entity_Name_2=Read(Label, ILO2)
      EndIf
   EndWhile
EndFor
Return(LCL, Index_Mappings)
End
```

3.4 Alignment generation

The alignment will be created automatically from the LCC, LCL and LCW lists. The algorithm for generating the alignment is illustrated by the Algorithm 7.

Algorithm7: Alignment Generation

```
Inputs: LCC, LCL, LCW
Outputs: F-ALIGNE
For Each Concept ϵ LCC do
    Entity_Name_1=Read(Concept, LCC)
    Entity_Name_2=Read(Concept, LCC)
    Add(Entity_Name_1,F-ALIGNE)
    Add(Entity_Name_2,F-ALIGNE)
EndFor
For Each Concept ϵ LCL do
    Entity_Name_1=Read(Concept, LCL)
```

```
        Entity_Name_2=Read(Concept, LCL)
        Add(Entity_Name_1,F-ALIGNE)
       Add(Entity_Name_2,F-ALIGNE)
    EndFor
    For Each Concept ϵLCW do
        Entity_Name_1=Read(Concept, LCW)
        Entity_Name_2=Read(Concept, LCW)
        Add(Entity_Name_1,F-ALIGNE)
        Add(Entity_Name_2,F-ALIGNE)
    EndFor
    Return (F-ALIGNE)
    End
```

ONTEM is a fully automatic method and requires no user intervention during the alignment process. However, the expert who is a knower of the field can confirm, suggest other alignments

4 Experimentation

To highlight the validity of our method, we will compare the alignments that ONTEM has produced with a reference alignment contained in the Oaei_LargeBio_Track_2018 section [23].

The ONTEM prototype was developed on the Eclipse Helios platform, using the java and APIjena2.4 programming language, as well as the SPARQL semantic graph reading language.The machine on which the work was performed has an Intel® Core TM(2) Duo CPU E7500 2.93 Ghz 2.94 Ghz, 4.0 GB RAM, 32bit operating system, Windows 7 Professional N.The evaluation metrics Precision, Recall and F-measure were used to compare our ONTEM method with other pioneering methods in the field, namely: AML, FCAMapX, LogMapBio, LogMap, LogMapLt, XMap, POMAP++[24], DOME, ALDO2Vec, KEPLER[25]. The results are illustrated in Figures 2 and 3.

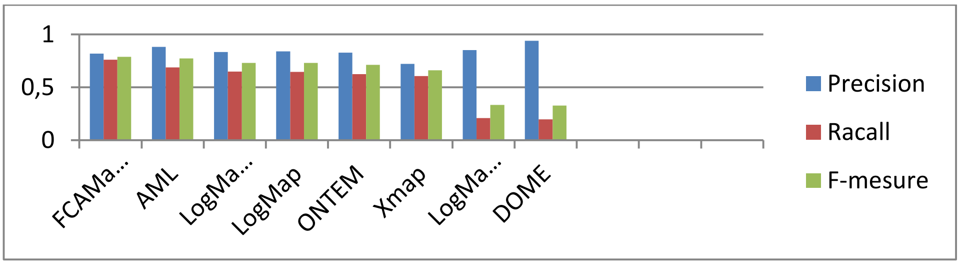

Figure 2. Comparative results of FMA Whole Ontology with SNOMED large fragments

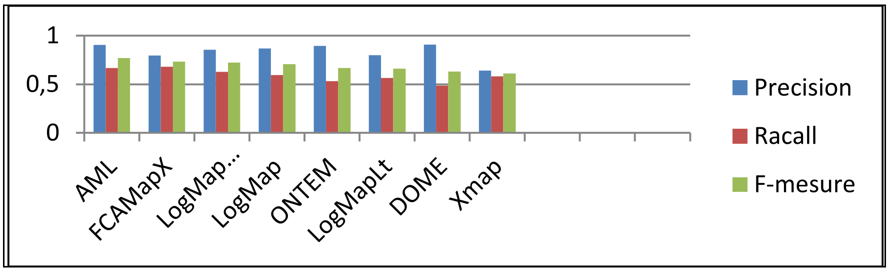

Figure 3. Comparative results of NCI Whole ontology with SNOMED large fragments

Although ONTEM does not perform any similarity calculations, the results we have achieved are more than satisfactory. They have shown promising results for its first comparison with the reference alignments contained in the LargeBioTrack 2018 section. We show through ONTEM that even if the ontologies are very large, their effectiveness increases. Therefore, it is not necessary to limit the size of the concept sets at the input of the alignment tool. We are convinced that ONTEM will provide a new basis for all other methods based on similarity calculations. Indeed, these calculations will only concern concepts not processed by our method.

5 Conclusion and perspectives

In this article, we have focused on the issue of large-scale ontology alignment for the Semantic Web. Indeed, the variety of ontologies of the same domain in the semantic web has led to heterogeneity and therefore to the development of ontology alignment methods. For more than a decade, ontology alignment methods have been attempting to solve heterogeneity and ontology matching problems. Today in many real applications such as in the medical field, the size of ontologies is very large and current alignment methods are faced with many challenges such as lack of memory and long processing times. We have shown that our ONTEM method stands out from the crowd of existing methods by its originality. It makes it possible to build new architectures based on existing methods that will boost ONTEM for much better results, because the purpose of all the work is to be able to make ontology-based information systems interoperable. To this end, this document provides an appropriate solution to this type of problem.A prototype has been set up to support the proposed approach. With this realization, we were able to evaluate our comments and compare them with other recognized methods in this field such as the OAEI_LargeBio_Track section of the 2018 campaign.

In our future work, we plan to consolidate our method to better support the alignment of full-scale large ontologies. We have already started to address this issue, but updating the test database poses other challenges, in terms of the ontological languages used and the evolving semantic description formalisms.

REFERENCES

1. Furst F., Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationnalisation, thèse de doctorat, Université de Nantes ,2004.
2. Diallo G. An effective method of large scale ontology matching. J Biomed Semant. 2014; 5(1):44.
3. Jimenez-Ruiz, E., Cuenca Grau, B., Zhou, Y., Horrocks, I.: Large-scale interactive ontology ´ matching: Algorithms and implementation. In: European Conf. Artif. Intell. (ECAI). (2012).
4. Ivanova, V, Lambrix, P, and Aberg, J., Requirements for and evaluation of user support for large-scale ontology alignment. In Proceedings of the European Semantic Web Conference, pages 3–20, 2015. doi: 10.1007/978-3-319-18818-8 1.
5. Oliveira, D., Pesquita, C., Improving the interoperability of biomedical ontologies with compound alignments. Journal of biomedical semantics 9(1) (2018)
6. Euzenat J, Shvaiko P. Ontology Matching, 2nd edition, Springer; 2013.
7. Miller G., al., Introduction to WordNet: An on-line lexical database, MIT Press, 1993.
8. Fatima Ardjani, DjelloulBouchiha, MimounMalki, Ontology-Alignment Techniques: Survey and Analysis, IJMECS, vol.7, no.11, pp.67-78, 2015.
9. Pereira, S., Cross, V., Jimenez-Ruiz, E.: On partitioning for ontology alignment. In: International Semantic Web Conference (Posters & Demonstrations). (2017)
10. Emanuel Santos, Daniel, CatiaPesquita, and Francisco M Couto. Ontology alignment repair through modularization and confidence-based heuristics. PLoS ONE, 10(12):e0144807, 2015.
11. Hu, W., Qu, Y., Cheng, G.: Matching Large Ontologies: A Divide-and-Conquer Approach. Journal on Data and Knowledge Engineering, vol. 67(1), p.140--160. 2008
12. Hu Wei, Zhao Yuanyuan and QuYuzhong. Partition-Based Block Matching of Large Class Hierarchies. In: Proceedings of the First Asian Semantic Web Conference, 3-7, 2006, Beijing, China.
13. Silva, F.B., Tabbone, S., Torres, R.d.S.: Bog: A new approach for graph matching. In: Pattern Recognition (ICPR), 2014 22nd International Conference on. pp. 82–87. IEEE (2014)
14. Jiménez-Ruiz, E. and Grau, B.C. (2011). Logmap: Logic-based and scalable ontology matching. The Semantic Web–ISWC 2011. Springer
15. Soraya Setti Ahmed, MimounMalki, Sidi Mohamed Benslimane. Ontology Partitioning: Clustering Based Approach. International Journal of Information Technology and Computer Science, 2015, 06, 1-11
16. Ernesto Jimenez-Ruiz, Bernardo Cuenca Grau, Yujiao Zhou, and Ian Horrocks. 2012. Large-scale interactive ontology matching: Algorithms and implementation. Frontiers in Artificial Intelligence and Applications, 444–449.
17. Oliveira, D., Pesquita, C.: Improving the interoperability of biomedical ontologies with compound alignments. Journal of biomedical semantics 9(1) (2018)
18. Ruder, S., Vuli_c, I., S_gaard, A.: A survey of cross-lingual word embedding models.arXiv preprint arXiv:1706.04902 (2017)
19. M. Ehrig, Ontology Alignment: Bridging the Semantic Gap, Semantic Web and Beyond Computing for Human Experience 4, Springer, (2007), pp. 1-250.

20. Zhao M, Zhang S. Identifying and validating ontology mappings by formal concept analysis. In: ISWC Conference on Ontology Matching (OM), vol. 1766: 2016. p. 61–72. Online: www.CEUR-WS.org.
21. Diallo G. An effective method of large scale ontology matching. J Biomed Semant. 2014; 5(1):44. -47-Horridge M, Bechhofer S. The OWL API: A java API for OWL ontologies. Semantic Web. 2011; 2(1):11–21.
22. Stoilos, G., Geleta, D., Shamdasani, J., Khodadadi, M.: A novel approach and practical algorithms for ontology integration. In: Proceedings of ISWC (2018).
23. http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/2018/result
24. Laadhar, F. Ghozzi, I. Megdiche, F. Ravat, O. Teste, F. GargouriPOMap: An Effective Pairwise Ontology Matching System 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KEOD'17), Funchal (Madeira, Portugal) 2017
25. Kachroudi, M., Diallo, G., Yahia, S.B.: OAEI 2017 results of KEPLER. In: 12th International Workshop on Ontology Matching. (2017) 138–145