Boosting Holistic Ontology Matching : an Extended Linear Approach and its Evaluation on Graph Clique-based Relaxed Reference Alignments

Philippe Roussille, Imen Megdiche, Olivier Teste, Cassia Trojahn dos Santos

Boosting Holistic Ontology Matching: Generating Graph Clique-Based Relaxed Reference Alignments for Holistic Evaluation

Philippe Roussille(✉), Imen Megdiche, Olivier Teste, and Cassia Trojahn

Institut de Recherche en Informatique de Toulouse, Toulouse, France
{philippe.roussille,imen.megdiche,olivier.teste,cassia.trojahn}@irit.fr

Abstract. Ontology matching is the process of finding correspondences between entities from different ontologies. Whereas the field has fully developed in the last decades, most existing approaches are still limited to *pairwise matching*. However, in complex domains where several ontologies describing different but related aspects of the domain have to be linked together, matching multiple ontologies simultaneously, known as *holistic matching*, is required. In the absence of benchmarks dedicated to holistic matching evaluation, this paper presents a methodology for constructing *pseudo*-holistic reference alignments from available pairwise ones. We discuss the problem of relaxing graph cliques representing these alignments involving a different number of ontologies. We argue that fostering the development of holistic matching approaches depends on the availability of such data sets. We run our experiments on the OAEI Conference data set.

1 Introduction

Ontology matching is an essential task for the management of the semantic heterogeneity problem in diverse environments. It aims at finding correspondences between entities from different ontologies. Diverse approaches have been proposed in the literature [3] and systematic evaluation of them has been carried out over the last fifteen years in the context of the Ontology Alignment Evaluation Initiative (OAEI) [2] campaigns. Despite the progress in the field, most efforts are still dedicated to *pairwise ontology matching* (i.e., matching a pair of ontologies). However, with the increasing amount of knowledge bases being published on the Linked Open Data, covering different aspects of overlapping domains, the ability of simultaneously matching different ontologies, a task so-called *holistic ontology matching* [12,17], is more than ever required. It is typically the case in complex domains, such as bio-medicine, where several ontologies describing different but related phenomena have to be linked together [14]. As stated in [15], the increase in the matching space and the inherently higher difficulty to compute alignments pose interesting challenges to this task.

https://doi.org/10.1007/978-3-030-03667-6_23

Early works on the field have addressed the problem of holistic schema matching, in particular the works on attribute matching [6,7,18,19]. In [6], a probabilistic framework determines an underlying model capturing the correspondences between attributes in different schemes. For dealing with complex attribute correspondences, the approach in [7] exploits co-occurrence information across schemes and a correlation mining method. This approach has been extended in [19] improving accuracy and efficiency, by reducing the number of synonymous candidates. In [18], the approach aims at incrementally merging 2-way schemes by clustering the nodes based on linguistic similarity and a tree mining technique. Emerging works have addressed the problem of holistic matching of more expressive structures. In [5], the proposal relies on a cross-domain holistic matching approach for aligning large ontologies by grouping concepts in topics that are aligned locally. More recently, a cluster-based distributed holistic approach for data linking has been proposed in [13]. Although novel approaches dedicated to holistic ontology matching have emerged in the literature in the last years, there is however a lack of reference alignments on which these approaches can be systematically evaluated. According to [14], producing such kind of alignments could be potentially useful to support a next generation of semantic technologies. We argue that fostering the development of these approaches depends on the availability of reference alignments.

This paper addresses the problem of holistic ontology matching and the lack of benchmarks in the field. As such, we attempt to study the problem through these main goals, following a methodology we built in different steps:

- we first designed an algorithm as a mean to allow us to build automatically from existing (and depending on) pairwise alignments a way to test and approach what could be considered as holistic (hence we use the term *pseudo-holistic approach*); mainly by our analysis of the concept of "alignment" through the lens of topological graphs, and our work around the concept to produce nuanced views through different levels of relaxation;
- we then applied our algorithm on the OAEI Conference data set, aiming to produce a baseline for our works in order to produce a similar matching task, as there is no current track providing holistic alignment challenges;
- finally, we chose to check the pertinence of the *pseudo-holistic* concept and produced alignments by evaluating the runners-up state-of-the-art tools of the OAEI track on this new task; so that we can discuss the pertinence of having a tool which can evaluate the point of being *holistic*.

For our experiments, we chose to have our evaluation done both ways: by assessing the generated alignments with the existing tools, we want to show that the holistic dimension is not something that can be reduced to a grouping of pairwise matching; and that some tools already using holistic matching, like LPHOM, outperform traditional tools for such a task. This is, for us, a necessary step so we can proceed further towards a full evaluation of the holistic task, by providing competitors that would help to better asses the performance of LPHOM and other holistic matching techniques, while assessing them on peer-reviewed specific alignments.

We organised our work as follows. Section 2 introduces the problem of holistic matching. Section 3 presents our methodology for creating holistic alignments from existing pairwise alignments. In Sect. 4, we discuss the experiments and results. Section 5 presents related works. Finally, Sect. 6 concludes the paper and gives future directions.

2 Problem Statement

Broadly speaking, the matching process takes as input a set of ontologies, denoted Ω, and determines as output a set of correspondences, called *alignment*. The *pairwise ontology matching* process takes as input two ontologies, $\Omega = \{O_1, O_2\}$, and determines as output a set of correspondences denoted as $A_{12} = \{c_1, c_2, ..., c_M\}$. A correspondence c_i can be defined as $<\{e_1, e_2\}, r, n>$, such that: e_1 and e_2 are ontology entities (e.g. properties, classes, instances) of O_1 and O_2, respectively; r is a relation holding between e_1 and e_2 (usually, $\equiv$, $\sqsupseteq$, $\perp$, $\sqcap$); and n is a confidence measure in the [0, 1] range assigning a degree of trust on the correspondence. The higher the confidence value, the higher the likelihood that the relation holds.

We can see the pairwise matching as a special case of holistic ontology matching. The *holistic ontology matching* takes a set $\Omega = \{O_1, ..., O_N\}$ of ontologies with N $\geq$ 2. It consists in determining a set of correspondences as $A_{1...N} = \{c_1, c_2, ..., c_M\}$. Each correspondence c_i is defined as $<\{e_1, ..., e_N\}, r, n>$ such as $\forall j \in [1..N], e_j \in O_j$. For our problem statement, we restrict r to the equivalence relationship between entities.

In case of $N = 3$, each correspondence c_i is defined as a triple correspondence $<\{e_1, e_2, e_3\}, \equiv, n>$ where $e_1 \in O_1$, $e_2 \in O_2$ and $e_3 \in O_3$. Triple correspondences correspond to *cliques* (i.e., a subset of vertices of an undirected graph such that every two distinct vertices in the clique are adjacent) or *Clique-relaxed graphs* as shown in Fig. 1. The main difference between both cases is the value of the confidence value, calculated taking into account the cardinality of the clique (as detailed in Sect. 3.2):

- clique correspondence is $<\{e_1, e_2, e_3\}, \equiv, 1>$
- clique-relaxed correspondence is $<\{e_1, e_2, e_3\}, \equiv, \frac{2}{3}>$.

Fig. 1. Clique-based holistic correspondence (a, left) and clique-relaxed holistic correspondence (b, right).

3 Building Holistic Alignments

In this section, we present a methodology for automatically constructing holistic alignments from available pairwise alignments. This requirement lets us to denote a *pseudo-holistic* approach. This methodology methodology is composed of two main steps:

1. building a graph of all combinations of correspondences in existing pairwise alignments;
2. building the holistic alignments according to different levels of relaxation with respect to complete graphs (cliques): clique-strict method (level 1) and clique-relaxed subgraph method (level 2). In case of level 2, we propose two sub-methods: the first method is a systematic relaxation of cliques, and the second one handles the intra-ontology choice of entities based on ontology relations.

The use of our relaxed methods is to add a complementary idea to the strict clique-approach. While consensus and alignment, on a pairwise basis, can be stated through binary acceptance (this corresponds vs. this does not correspond), we build the relaxed methods on the need to add some blurriness to account for the multiple agreements between ontologies. By using a relaxed consensus, we take into account how holistic agreements can be reached within the group; which is accounted by the first method; and by using a third ontology (WordNet), we find a method to solve consensus from within a group that would appear "natural" from an external onlooker which is here represented by WordNet, acting as a reference and external ambiguity resolver.

3.1 Step 1: Building the Graph of N Pairwise Alignments

This step aims at building a holistic graph $G_H = (V_H, E_H)$ where nodes are entities from the ontologies to be aligned, and edges are correspondences from pairwise alignments, such as:

- $V_H = \{e_{i_k} | e_{i_k} \in \cup_{i=1}^{N} O_i\}$,
- $E_H = \{(e_i, e_j) | \exists < \{e_i, e_j\}, r, n > \in A_{1..N}\}$, with $A_{1..N} = \cup_{k=1,l=k+1}^{N-1} A_{kl}$.

Remark. If we consider $N = 4$, this leads us to the group of $A_{12} \cup A_{13} \cup A_{14} \cup A_{23} \cup A_{24} \cup A_{34}$.

3.2 Step 2: Building Holistic Alignments

This section details the two methods for building the holistic alignments $A_{1...N} = \{c_1, c_2, .., c_i, ..| i \in \mathbb{N}\}$.

Each correspondence should cover the N input ontologies and should be $1:1$ holistic alignment to conserve the $1:1$ requirements of the pairwise alignments. In the following, we explain both levels of methods and the algorithms that we propose to generate the holistic alignments.

Clique-Strict Method (Level 1). The first method concerns the generation of holistic alignments composed of cliques. The cliques are complete graphs extracted from the holistic graph G_H. The algorithm we developed consists in searching complete subgraphs composed of N nodes belonging to the N input ontologies. A clique is considered as the most strongest holistic correspondence, hence it has the confidence value 1.

Remark. To find a clique in the graph G_H, we use the method *find_cliques* from the networkx Python module[1]. The structure of the graph G_H built upon 1:1 pairwise alignments guarantees that each ontology is present only once in the cliques. However, the networkx module can not guarantee the $1:1$ holistic alignments which means that all the N input ontologies are present on the clique results. That's why we check-up if the final selected cliques covers the N input ontologies.

Clique-Relaxed Subgraph Method (Level 2). The clique-strict method is too strict because we are faced most commonly to incomplete graphs that should be part of the solutions of holistic alignments. To concretely expose the idea, we notice that the left subgraph in Fig. 2(b) is part of the solution of the complete graph of Fig. 2(a). Hence, we can infer from the subgraph of Fig. 2(b) a holistic alignment with a lower level of confidence corresponding to its incompleteness with respect to the clique.

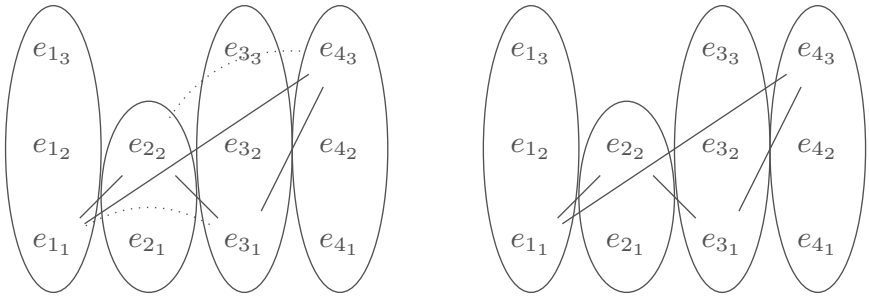

Fig. 2. (a) Example of a clique subgraph; (b) The clique-relaxed subgraph.

In order to compute the confidence of the clique-relaxed subgraph, we define the notion of *clique-likeness*, which is the geometric distance of a subgraph compared to a clique; for instance, the level of confidence of the graph of Fig. 2(a) is $\frac{2}{3}$. The formula is as the following for a subgraph denoted $G_i = (V_i, E_i)$:

$$clique\_likeness(G_i) = \frac{2 * |E_i|}{|V_i| * (|V_i| - 1)}$$

[1] https://networkx.github.io/documentation/networkx-1.9.1/reference/generated/networkx.algorithms.clique.find_cliques.html#networkx.algorithms.clique.find_cliques.

We rely on the *connected_component_subgraphs* method from the networkx Python module[2] to search all the subgraphs of G_H with respect to two conditions, namely that all ontologies should be represented by at least one node, and that each subgraph G_i is maximal. Based on the content of these subgraphs, we provide two methods to generate the holistic alignments.

Method 1: Clique-Relaxed Holistic Alignment Algorithm. This method is a systematic relaxation of cliques, which means that the subgraphs are incomplete cliques composed exactly of one node from the N input ontologies. This method is explained in Algorithm 1.

Algorithm 1. Clique-relaxed holistic alignment algorithm: Method 1

Data: G_H
Result: $A_{1..N}$
$A_{1..N} \leftarrow \emptyset$;
foreach $G_i \in connected\_component\_subgraphs(G_H)$ **do**
  $//G_i = (V_i, E_i)$ is a subgraph of G_H
  $onto \leftarrow \emptyset$;
  foreach $e_{j_k} \in V_i$ **do**
    $onto \leftarrow$ onto $\cup$ {j} ;
  end
  if $|onto| = N$ *and* $|V_i| = N$ **then**
    $A_{1..N} \leftarrow A_{1..N} \cup \{< V_i, \equiv, clique\_likeness(G_i) >\}$;
  end
end

Method 2: Clique-Relaxed Subgraphs Based on Intra-ontology Relations. This method handles the case when the subgraphs are composed of one or several nodes from ontologies O_i, for some or all $i \in [1, N]$. The proposed method will then select only one tuple of nodes based on the intra-ontology relations and the best confidence value of *clique_likeness*.

By taking the example of Fig. 3, we notice that the subgraph have two nodes from O_1, noted e_{1_1} and e_{1_2}, so we have to choose either the solution 1, composed of the clique-relaxed $= \{e_{1_1}, e_{2_1}, e_{3_2}, e_{4_3}\}$ or solution2, composed of the clique-relaxed $= \{e_{1_2}, e_{2_1}, e_{3_2}, e_{4_3}\}$.

- For solution 1 (a), the $clique\_likeness(G_i) = \frac{1}{3}$.
- For solution 2 (b), we propose that we can use the relationship between e_{1_1} and e_{1_2} to infer new mappings for e_{1_2}. As the $e_{1_2} \subseteq e_{1_1}$ (subclassof relation) and $<\{e_{1_1}, e_{2_1}\}, \equiv, 1>$ thus we can infer the pairwise mapping $<\{e_{1_2}, e_{2_1}\}, \equiv, 1>$. Therefore, the $clique\_likeness(G_i) = \frac{1}{2}$.

Based on the *clique_likeness* score, we choose the solution 2 because of its higher confidence value (Fig. 4).

[2] https://networkx.github.io/documentation/networkx-1.9.1/.

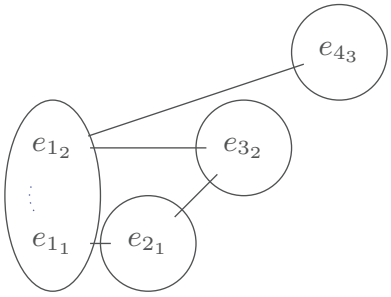

Fig. 3. Example of intra-ontology multiple choice. The circled elements belongs to the same ontologies, the black vertices shows the extra-ontological links while the blue dotted vertices shows intra-ontological links. (Color figure online)

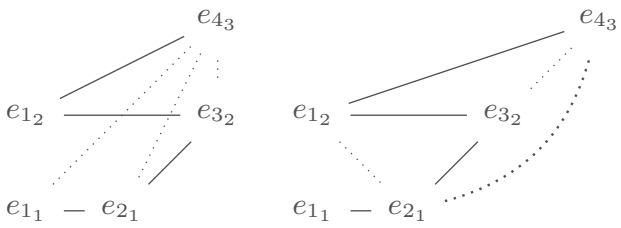

Fig. 4. (a) Solution 1 and (b) solution 2 from Fig. 6.

Algorithm 2 implements *method 2* and complements *method 1*. It can be used by pairwise tools for constructing their holistic alignments, based on the generated pairwise alignments. In this algorithm, the function named *score* calculates a score from a set of entities. For each entity, we normalize its name (lowering case, removing camel case and snake case) so that it can be seen as a sentence (or a word). Through POS, we find the most important word in such sentence, removing duplicates if any ("Conference Paper" and "Paper" are both seen as "Paper" duplicates). We then compute all the hypernyms of all the synonyms of this word, using WordNet. Then, we compute the intersection of the hypernyms of the entities of one candidate; the score being its cardinality.

In the example of Fig. 5, we illustrate the case of $N = 4$ ontologies from the OAEI Conference Track (cmt, conference, iasted and edas). We notice two possible solutions that can be proposed for the subgraph composed of the entities "Submission" (iasted), "Submitted_contribution" and "Paper" (conference), "Paper" (edas), and "Paper" (cmt). In order to find the alignment, we compute the score of the two potential clique-relaxed subgraphs which contains either the entity "Paper" or "Submitted_contribution" (conference). The retained holistic alignment is solution 1, which has the highest score; its confidence value is $\frac{3}{6} = 50\%$.

Algorithm 2. Select clique-relaxed subgraphs based on intra-ontology relations: Method 2

Data: G_H
Result: $A_{1..N}$
$A_{1..N} \leftarrow \emptyset$;
foreach $G_i \in connected\_component\_subgraphs(G_H)$ **do**
  $//G_i = (V_i, E_i)$ is a subgraph of G_H
  $onto \leftarrow \emptyset$;
  foreach $e_{j_k} \in V_i$ **do**
    $onto \leftarrow$ onto $\cup$ {j} ;
  end
  if $|onto| = N$ **then**
    if $|V_i| = N$ **then**
      $A_{1..N} \leftarrow A_{1..N} \cup \{¡V_i, \equiv, clique\_likeness(G_i) >\}$;
    end
    else
      for $j \leftarrow 1$ **to** N **do**
        $E_j \leftarrow \emptyset$;
        foreach $e_{j_k} \in V_j \cap V_i$ **do**
          $E_j \leftarrow E_j \cup \{e_{j_k}\}$
        end
      end
      $s_{max} \leftarrow 0$;
      $c_{max} \leftarrow \emptyset$;
      foreach $clique \in \prod_{j=0}^{N} E_j$ **do**
        if $score(clique) \geq s_{max}$ **then**
          $s_{max} \leftarrow score(clique)$;
          $c_{max} \leftarrow clique$;
        end
      end
      $G_{max} \leftarrow (c_{max}, \{(e_j, e_k) | \exists e_j \in c_{max}, \exists e_k \in c_{max}, \exists (e_j, e_k) \in E_i\})$;
      $A_{1..N} \leftarrow A_{1..N} \cup \{< c_{max}, \equiv, clique\_likeness(G_{max}) >\}$;
    end
  end
end

4 Experiments

4.1 Materials and Methods

Data Set. Our holistic reference data set has been constructed from the OAEI Conference data set[3], which provides real-world and expressive ontologies covering the conference organisation domain [22]. This data set is composed of 16 ontologies and a subset of 21 pairwise reference alignments involving 7 ontologies (ra1). We have applied the 3 methods described above for generating the

[3] http://oaei.ontologymatching.org/2017/conference/.

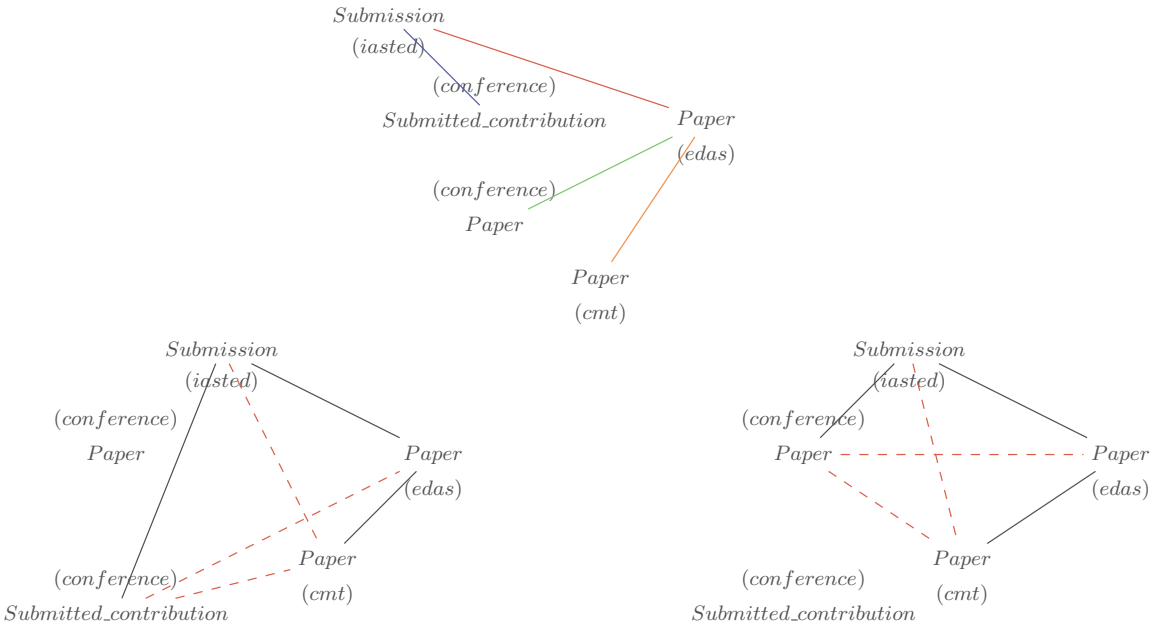

Fig. 5. (a) Original extracted subgraph, (b) method 1, (c) method 2

holistic reference alignments on the basis of ra1. Although transitive closure computed alignments for this track have been constructed and manually checked by evaluators (ra2), they are not available. The version of our holistic data set is hence based on the publicly available original alignments (ra1). All the generated alignments and the code for generating them are available online[4].

Tools. We have applied our methodology to generate holistic alignments from the available results of OAEI 2017 participating tools[5] and compared their results with the LPHOM holistic approach [12]. The available results for the following tools were considered: ALIN, AML, KEPLER, LogMap, LogMapLt, ONTMAT, POMap, SANOM, WikiV3 and XMap. Even though these tools were not developed for that purpose, their results were the only available for a baseline comparison. To the best of our knowledge very few holistic systems are available. We have run the AML-Compound tool[6], but it was not able to generate any alignment for this data set.

Evaluation Metrics. The results are discussed in terms of precision, recall and F-measure. We compare the correspondences from the reference alignment to the correspondences generated by the matchers considering an exact match. For all evaluated alignments we do not take into account their confidence.

Execution Environment. All the experiments have been run on a 32 GB RAM available, 7CPU x64 @3.6 Ghz machine. While LPHOM takes some seconds for generating the alignments, we could not compare its runtime performance with the OAEI tools (only alignments are available).

[4] https://github.com/PhilippeRoussilleIRIT/EKAW-2018-holistic.
[5] http://oaei.ontologymatching.org/2017/conference/eval.html.
[6] https://github.com/AgreementMakerLight/AML-Compound.

4.2 Results and Discussion

In this section, we first provide an analysis of the data set, comparing the behaviour of the different methods for generating the reference alignments. This comparison takes into account the overall results of the matching tools in each setting. Figure 6 gives an overview of the results' distribution reflecting the complexity of each method for each $N > 2$ (we have intentionally hidden tool names). We can clearly distinguish two types of behaviours:

- for the trend of $N = 3$ and $N = 4$, with few exceptions, we can observe that the clique-strict method results are closer to both relaxed methods. It shows that with few number of ontologies, regardless the kind of method, the correspondences generated by the methods are close. The structural difference given a clique compared to a relaxed clique is smaller the fewer nodes in the sub-graph.
- for the second trend of $N = 5$ and $N = 6$, we can observe that the method clique-strict is better than both relaxed methods. These cliques allows for identifying the common entities shared across the ontologies. In the case of the Conference data set, by manually examining the outputs, the clique-strict alignments are composed of exact matches. It corroborates the intuition that increasing the number of nodes in a subgraph, increases the differences between their structures (cliques and relaxed cliques structures).

Second, we compare the performance of the holistic alignments generated from the pairwise ones coming from the OAEI tools, with respect to the generated holistic reference one. Although this evaluation setting may introduce a bias in the evaluation, in the lack of available fully holistic tools, it is the material

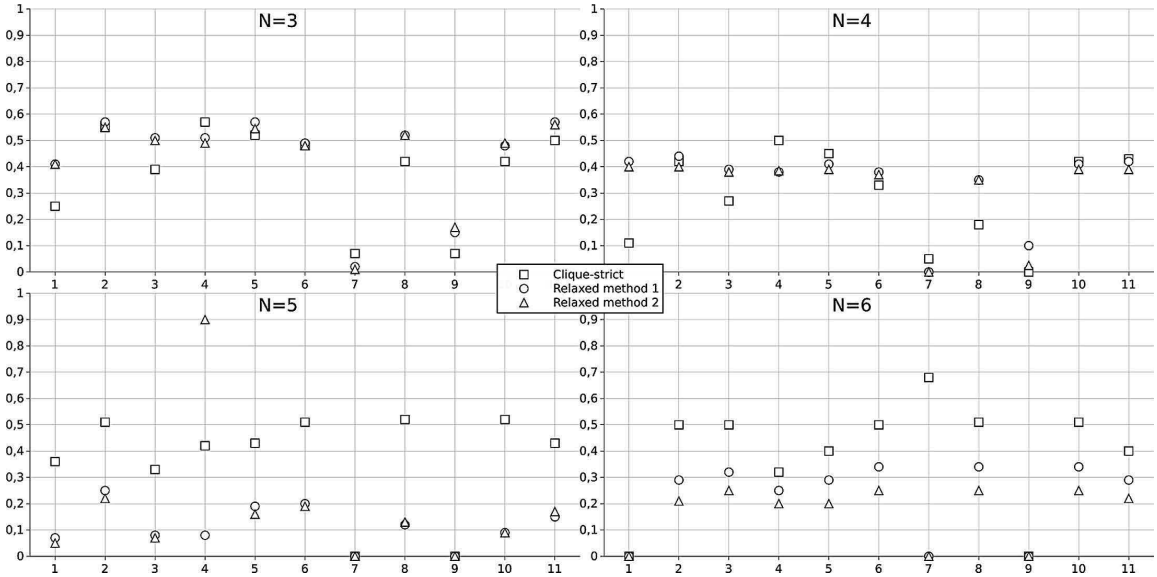

Fig. 6. Comparison of methods according to the number of input ontologies, with the number of ontologies (in abscissa) and the f-measure (in ordinate).

we have for comparison. Table 1 shows the results applying the different proposed methods for generating holistic reference alignments, varying the number of input ontologies (N). Looking at first to the holistic tool, we can observe that, although the LPHOM holistic approach does not perform very well for a small number of ontologies, it is in the top-3 (f-measure) for $N = 6$ ontologies (for all methods). As expected, the tools specifically designed for the pairwise task better perform for $N = 2$. Their performance however mostly decreases with the increasing of N (some are not able to generated alignments for $N = 6$), while LPHOM relatively maintains its performance.

Overall, as Fig. 7 shows, in terms of precision, LPHOM (.56) is of the top-4 systems (AML and ALIN .59, XMap .58 and PopMap .57). The holistic approach privileges precision in detriment of recall (.35), with coherent generated alignments. In terms of F-measure, the given results are intermediate, about .10 points (.42) compared to the best system, which is AML (.52). However, we have to keep in mind that our approach here is *pseudo*-holistic, and thus heavily influenced by the number of ontologies. As the number of ontologies increases, reaching up to 6, the F-measure decreases, showing that there are room for improvements. This can be explained due to the structures of the tasks and the way the tools work: as the matching structures differ from a strict clique approach (which, in a pairwise context, is kept all the time as pairwise alignments are cliques), the limits between matches become blurrier. Most tools will easily find a similarity between two entities, and two groups of entities, but the transient aspect of the pseudo-holistic relaxation cannot be easily translated in terms of strictness. As such, when trying to assess all ontologies at once, only the main and nearly exact matches remain; while when computed pairwise, this information cannot be extrapolated as the similarity matrix does not incorporate the new similarities. Finally, we are ware that the performance of the different matchers compared to LPHOM are not as significant as if our experiments were ran using specifically holistic matchers. However, they are significant enough to show that the holistic matching task has inherent properties.

Table 1. Evaluation results on F-measure. N indicates the number of input ontologies. Higher is better.

Method	Clique-strict					Clique-relaxed: method 1					Clique-relaxed: method 2				
N	*2*	*3*	*4*	*5*	*6*	*2*	*3*	*4*	*5*	*6*	*2*	*3*	*4*	*5*	*6*
ALIN	.30	.24	.11	.36	.00	.42	.42	**.43**	.06	.00	.41	.42	**.41**	.05	.00
AML	**.62**	**.55**	.43	**.52**	.50	**.71**	**.56**	**.44**	**.25**	.29	**.71**	**.54**	**.41**	**.23**	.22
LPHOM	.47	.38	.17	.34	**.50**	.58	.50	.40	.10	**.33**	.58	.49	.38	.09	**.25**
KEPLER	.54	**.56**	**.49**	.43	.33	.59	.50	.39	.10	.25	.59	.48	.37	.10	.20
LogMap	**.61**	**.53**	**.44**	.44	.40	**.67**	**.56**	.41	**.19**	.29	**.67**	**.54**	.38	**.16**	.20
LogMapLt	.54	.47	.35	.52	.50	.59	.49	.38	**.19**	**.33**	.58	.48	.36	**.18**	**.25**
OntoMap	.22	.06	.03	.00	**.67**	.03	.00	.00	.00	.00	.03	.00	.00	.00	.00
PopMap	.50	.42	.15	**.53**	**.50**	.62	.52	.36	.13	**.33**	.62	.50	.34	.12	**.25**
Sanom	.28	.06	.00	.00	.00	.37	.14	.03	.00	.00	.37	.13	.03	.00	.00
Wikiv3	.49	.44	**.43**	**.53**	**.50**	.57	.48	.41	.09	.33	.57	.47	**.39**	.09	.25
XMap	**.59**	.50	.43	.45	.40	**.69**	**.56**	**.42**	.15	.29	**.69**	**.54**	.39	.14	.22

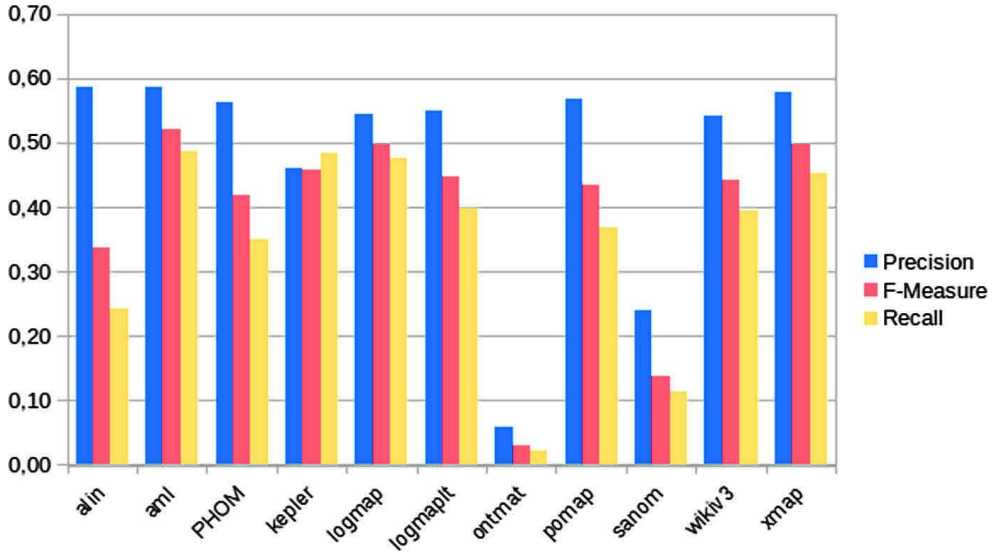

Fig. 7. Average results (precision, recall and f-measure)

5 Related Work

In this section we describe the main related work on (i) holistic approaches and (ii) holistic ontology reference alignments.

Holistic Approaches. As stated in Sect. 1, most works on holistic matching give special attention to attribute matching [6,7,19]. These approaches handle simple attributes compared to the more structured schemes of ontologies. Under a different perspective, a cross-domain holistic approach for matching large ontologies has been proposed in [5]. In [15], the holistic AML-Compound system extends the pairwise AML system adapting WordNet similarities and Jaccard indexes. Recently, an instance-based distributed holistic approach is presented in [13], which is based on a clustering of entities representing the same real-world object. Differently from [5–7,19], LPHOM is not restricted to attributes, while we do not perform cross-domain holistic matching as [5]. Compared to [7], LPHOM can also return simple and multiple correspondences and it is extensible to new constraints, differently from [15]. As some pairwise matchers [10,11], we adopt constraints that reduce the possibility of generating incoherent alignments. With respect to the matching strategies we apply, while the selection strategy in [21] is based on paths in the graph, we reduce the selection to the maximum-weighted bipartite graph matching (MWGM) problem like OLA [4] and we adopt a different structural similarity strategy from [8]. Compared to OLA we do not compute structural similarities but encode structural properties as linear constraints. As CODI [9], we perform both structural matching (without additional structural similarity computation) and alignment extraction phases. Unlike CODI whose pairwise approach is reduced to a NP-Hard problem, our solution extends a polynomial problem in both pairwise and holistic versions [12]. In a holistic and monolingual setting, we apply a combinatorial optimisation problem using linear programming, as done in [16] in pairwise. The

constraints proposed by [16] for multiple correspondences, can be simply added to our model to enhance the matching of multiple correspondences in the relaxed version of our model.

Holistic Reference Alignments. While systematic evaluation of matching approaches has been dedicated to pairwise systems[7], there is a lack of reference alignments on which these approaches can be systematically evaluated. We argue that fostering the development of these approaches depends on the availability of such data sets. Current holistic approaches are (manually) evaluated on data sets used in the context of the tool development. The closer approach to ours is from [15]. The authors propose to exploit OBO cross-products to create ternary compound alignments between ontologies, in order to create a benchmark. They have created a set of seven cross-products collections each with at least 100 definitions corresponding to ternary compound correspondences. Differently from [15], our correspondences do not involve any logical construction and are not limited to ternary composition of ontologies. This could be rather seeing as generating complex correspondences [20]. Finally, in [13], a reference alignment for multi-source clustering of large data sets from the geographic and music domains has been proposed. They evaluate the efficiency and scalability of the distributed holistic clustering for large data sets with millions of entities from the two domains. While they handle larger data sets focusing on linking discovery, our approach is limited to schema matching [1].

6 Concluding Remarks and Future Work

This paper has proposed a methodology for constructing holistic alignments from existing pairwise alignments. The approach relies on graph cliques involving a different number of ontologies. We applied our approach for generating holistic reference alignments from the original Conference reference alignments. These alignments have been the basis for evaluating alignments from a specific designed matcher and from OAEI matchers, in a holistic setting. Although we propose a pseudo-holistic approach, it is a first step towards the holistic ontology matching evaluation, open new challenges in the field.

As future work, we plan to extend the evaluation of this data set with a manual verification as well as to work on the transitive closure computed alignments. We intend as well to work on other kind of relation than equivalences. This work also opens additional perspectives in the field, once current solutions to manage and evaluate ontology matching (i.e., Alignment API) and weighted and semantic precision and recall measures are limited to deal with pairwise matching.

Acknowledgement. This research received financial support by the SmartOccitania project from the France's Strategic Investment Program (Programme d'investissements d'avenir - PIA) and the French Environment & Energy Management Agency (ADEME).

[7] http://oaei.ontologymatching.org/.

References

1. Berro, A., Megdiche, I., Teste, O.: A linear program for holistic matching: assessment on schema matching benchmark. In: Chen, Q., Hameurlain, A., Toumani, F., Wagner, R., Decker, H. (eds.) DEXA 2015. LNCS, vol. 9262, pp. 383–398. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22852-5_33
2. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., Trojahn, C.: Ontology alignment evaluation initiative: six years of experience. J. Data Semant. **15**, 158–192 (2011)
3. Euzenat, J., Shvaiko, P.: Ontology Matching, 2nd edn. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38721-0
4. Euzenat, J., Valtchev, P.: Similarity-based ontology alignment in OWL-Lite. In: Proceedings of the 16th European Conference on Artificial Intelligence, pp. 333–337 (2004)
5. Gruetze, T., Böhm, C., Naumann, F.: Holistic and scalable ontology alignment for linked open data. In: Proceedings of the 5th Linked Data on the Web Workshop at the 21th WWW (2012)
6. He, B., Chang, K.C.-C.: Statistical schema matching across web query interfaces. In: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, SIGMOD 2003, pp. 217–228. ACM 92003)
7. He, B., Chang, K.C.-C., Han, J.: Discovering complex matchings across web query interfaces: a correlation mining approach. In: Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining, pp. 148–157 (2004)
8. Hu, W., Jian, N., Qu, Y., Wang, Y.: GMO: a graph matching for ontologies. In: K-Cap 2005 Workshop on Integrating Ontologies 2005, pp. 43–50 (2005)
9. Huber, J., Sztyler, T., Nößner, J., Meilicke, C.: CODI: combinatorial optimization for data integration: results for OAEI 2011. In: Proceedings of the 6th International Workshop on Ontology Matching (2011)
10. Jean-Mary, Y., Shironoshita, E., Kabuka, M.: Ontology matching with semantic verification. Web Semant. Sci. Serv. Agents World Wide Web **7**(3), 235–251 (2009)
11. Jiménez-Ruiz, E., Cuenca Grau, B.: LogMap: logic-based and scalable ontology matching. ISWC 2011. LNCS, vol. 7031, pp. 273–288. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25073-6_18
12. Megdiche, I., Teste, O., Trojahn, C.: An extensible linear approach for holistic ontology matching. In: Groth, P. (ed.) ISWC 2016. LNCS, vol. 9981, pp. 393–410. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46523-4_24
13. Nentwig, M., Groß, A., Möller, M., Rahm, E.: Distributed holistic clustering on linked data. In: Panetto, H., et al. (eds.) On the Move to Meaningful Internet Systems. OTM 2017 Conferences. OTM 2017. Lecture Notes in Computer Science, vol 10574, pp. 371–382. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69459-7_25
14. Oliveira, D., Pesquita, C.: Compound matching of biomedical ontologies. In: Proceedings of the International Conference on Biomedical Ontology, ICBO 2015, Lisbon, Portugal, 27–30 July 2015 (2015)
15. Pesquita, C., Cheatham, M., Faria, D., Barros, J., Santos, E., Couto, F.M.: Building reference alignments for compound matching of multiple ontologies using obo cross-products. In: Proceedings of the 9th International Workshop on Ontology Matching, pp. 172–173 (2014)
16. Prytkova, N., Weikum, G., Spaniol, M.: Aligning multi-cultural knowledge taxonomies by combinatorial optimization. In: Proceedings of the 24th International Conference on World Wide Web, pp. 93–94. ACM (2015)

17. Rahm, E.: Towards large-scale schema and ontology matching. In: Bellahsene Z., Bonifati A., Rahm E. (eds.) Schema Matching and Mapping. Data-Centric Systems and Applications. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-16518-4_1
18. Saleem, K., Bellahsene, Z., Hunt, E.: PORSCHE: Performance ORiented SCHEma mediation. Inf. Syst. **33**(7–8), 637–657 (2008)
19. Su, W., Wang, J., Lochovsky, F.: Holistic schema matching for web query interfaces. In: Ioannidis, Y., et al. (eds.) EDBT 2006. LNCS, vol. 3896, pp. 77–94. Springer, Heidelberg (2006). https://doi.org/10.1007/11687238_8
20. Thiéblin, É., Haemmerlé, O., Hernandez, N., Trojahn, C.: Towards a complex alignment evaluation dataset. In: OM Workshop at ISWC, pp. 217–218 (2017)
21. Xiang, C., Chang, B., Sui, Z.: An ontology matching approach based on affinity-preserving random walks. In: Proceedings of the 24th International Conference on Artificial Intelligence, pp. 1471–1477 (2015)
22. Zamazal, O., Svtek,V.: The ten-year ontofarm and its fertilization within the onto-sphere. Web Semant. **43**(C), 46–53 (2017)