A Clustering-based Approach to Ontology Alignment

Songyun Duan[1], Achille Fokoue[1], Kavitha Srinivas[1], and Brian Byrne[2]

[1] IBM T. J. Watson Research Center, Hawthorne, New York, USA
{sduan, achille, ksrinivs}@us.ibm.com
[2] IBM Software Group, Information Management, Austin, Texas, USA
byrneb@us.ibm.com

Abstract. Ontology alignment is an important problem for the linked data web, as more and more ontologies and ontology instances get published for specific domains such as government and healthcare. A number of (semi-)automated alignment systems have been proposed in recent years. Most combine a set of similarity functions on lexical, semantic and structural features to align ontologies. Although these functions work well in many cases of ontology alignments, they fail to capture alignments when terms or structure varies vastly across ontologies. In this case, one is forced to rely on manual alignment. In this paper, we study whether it is feasible to re-use such expert provided ontology alignments for new alignment tasks. We focus in particular on many-to-one alignments, where the opportunity for re-use is feasible if alignments are *stable*. Specifically, we define the notion of a cluster as being made of multiple entities in the source ontology $\mathcal{S}$ that are mapped to the same entity in the target ontology $\mathcal{T}$. We test the *stability hypothesis* that the formed clusters of source ontology are stable across alignments to different target ontologies. If this hypothesis is valid, the clusters of an ontology $\mathcal{S}$, built from an existing alignment with an ontology $\mathcal{T}$, can be effectively exploited to align $\mathcal{S}$ with a new ontology $\mathcal{T}'$. Evaluation on both manual and automated high-quality alignments show remarkable stability of clusters across ontology alignments in the financial domain and the healthcare and life sciences domain. Experimental evaluation also demonstrates the effectiveness of utilizing the stability of clusters in improving the alignment process in terms of precision and recall.

1 Introduction

Ontology alignment is an important problem for the linked data web, as more and more ontologies get published for specific domains such as government and healthcare. A number of (semi-)automated alignment systems have been developed in recent years (*e.g.,* Lily [16], ASMOV [8], Anchor-Flood [11], RiMOM [13]). Most systems combine a large set of similarity functions on lexical, semantic and structural features to align ontologies (for surveys, see [2], [14]). While these similarity functions are important and effective for many cases of ontology alignments, there are also cases where none of the similarity functions adequately capture the nature of the alignment; this is particularly true when the two ontologies of extremely different modeling granularities are involved. For instance, one alignment exercise frequently conducted by IBM consultants in the field is to align models that describe assets at an IT level (*e.g.,* the IBM Information FrameWork model used to describe IT assets in the banking industry) to models that describe the same assets at a business level (*e.g.,* the IBM Component Business Model for banking). Because the two models describe the same assets in different terms and different structures, traditional approaches to automated ontology alignment fail abysmally (the mapping precision we have measured can be as low as 1% in these cases). In fact, the only alternative in such cases is to rely on a domain expert who can provide the alignment between these types of models. However, if the expert has actually done the hard work of mapping the models once, is it feasible to re-use these high-quality manual ontology alignments, to improve the alignment process for new alignments when the two models evolve, or when the same model needs to be aligned to new models? This is the research focus of our paper.

For the purpose of investigating the re-use of manual mappings, we direct our attention in this paper to many-to-one (or conversely, one-to-many) mappings, because this is where mapping re-use can be readily applied while similarity functions fail to produce valuable information for alignments. In many-to-one mapping scenarios, multiple entities in one ontology $\mathcal{S}$ get mapped to a single entity in a target ontology $\mathcal{T}_1$. The grouping of multiple entities in $\mathcal{S}$ can be viewed as user-specified clustering of source entities. In principle, there is a chance that prior ontology alignments can provide some guidance for the current alignment task in hand, if there is some *stability* in mapping certain entities in one ontology to the same entity in target ontologies. Put it another way, the question is whether the user-specified clusters based on the alignment of $\mathcal{S}$ to $\mathcal{T}_1$ tend to appear when $\mathcal{S}$ is aligned with ontology $\mathcal{T}_2$ different from $\mathcal{T}_1$. If the user-specified clustering in $\mathcal{S}$ is in fact stable, then the clustering information can be exploited when alignment needs to be performed from $\mathcal{S}$ to $\mathcal{T}_2$. Specifically, a mapping provided by an expert on one of the entities in a cluster of $\mathcal{S}$ can be automatically generalized to map other members of this cluster.

To evaluate the *stability hypothesis*, we define two novel metrics to measure the similarity of clusters constructed for an ontology $\mathcal{S}$ based on its alignment results to different ontologies. These metrics are conceptually similar to Levenshtein and Jaccard measures of string similarity. We also design a mapping strategy that utilizes the clustering information for new alignments and study the effectiveness of this strategy in terms of the classical mapping quality metrics such as precision and recall. Furthermore, we characterize mapping *efficiency* in terms of the amount of saving in human effort required in the alignment task with and without the clustering information. We apply these metrics to compare two independent alignments that were performed by IBM consultants in the field. The first alignment involved the mapping of the IBM Component Business Model (CBM), a flat model of business functions expressed in business terms to Information FrameWork (IFW), a structured and detailed model of enterprise processes described from an IT perspective. The second alignment involved a very different version of the CBM model which was aligned to a mostly unchanged IFW model. The alignment process was conducted about a year apart, by different consultants. As mentioned earlier, applying any of the standard similarity functions to either model pair fails to detect any meaningful alignments. Manual mappings produced by IBM consultants had most CBM entities mapped to multiple IFW entities. Using these expert created mappings as reference, we tested whether the user defined clusters of IFW entities stayed stable when it was mapped to a very different version of CBM. Our evaluation of the previously defined metrics showed remarkable stability of clustering of IFW entities (the average similarity of clusters is 0.89, within the range of [0, 1]). For the same dataset, the improvement in mapping precision is 0.4, and the efficiency is 0.95 within a scale of 0 to 1; the higher the better. For repeatability purposes, we evaluated these same metrics for 2312 ontology comparisons publicly available on the BioPortal web site[3] with again remarkable stability of clustering of source ontology entities (the average similarity of clusters is 0.84). From these positive stability results, the opportunity for re-use is quite clear: clustering information generated from existing alignments is very helpful for new alignment tasks. For instance, if entities a, and b in ontology $\mathcal{S}$ are mapped to entity c in ontology $\mathcal{T}_1$, and a is mapped to entity d in another ontology $\mathcal{T}_2$, we know b should be mapped to d in $\mathcal{T}_2$ as well.

Our main contributions in this paper are as follows:

- We present a novel technique to uncover, from existing many-to-one (or conversely, one-to-many) alignments, internal structures of related entities (*i.e.,* clusters of entities) in ontologies.
- We show the stability of those clusters across alignments in two different domains (finance and healthcare & life sciences) and on both manually created mappings and automatically generated high-quality mappings.
- We describe how clusters discovered in existing many-to-one and one-to-many alignments can be exploited for performing new alignments, and evaluate the impact on both mapping quality (precision/recall) and mapping efficiency (saving in human effort).

The remainder of the paper is organized as follows. In the next section, we present an overview of our clustering-based ontology alignment approach and the fundamental stability hypothesis it relies on. In Section 3, we describe cluster similarity measures needed to validate the stability hypothesis. The evaluation results on many-to-one alignments are presented in Section 4. Finally, after discussing related work in Section 5, we conclude in Section 6.

2 Overview of Clustering-based Ontology Alignment

In many-to-one alignment scenarios, multiple entities in the source ontology $\mathcal{S}$ get matched to the same entity in the target ontology $\mathcal{T}$. One way to interpret

[3] http://bioportal.bioontology.org

the alignment result of $\mathcal{S} \rightarrow \mathcal{T}$ is that the entities in $\mathcal{S}$ are partitioned into clusters (*i.e.,* groups) such that each cluster of entities are matched to the same entity in $\mathcal{T}$. Consider a simple example.

Source ontology $\mathcal{S} = \{a, b, c, d\}$, target ontology $\mathcal{T} = \{e, f\}$, and their alignment result: $\mathcal{S} \rightarrow \mathcal{T} = \{a \rightarrow e, b \rightarrow e, c \rightarrow f, d \rightarrow f\}$. In this case, ontology $\mathcal{S}$ is partitioned into 2 clusters: $\mathcal{P}_s = \{\{a, b\}, \{c, d\}\}$.

It naturally follows that a source ontology could be partitioned in different ways based on its alignment results with different target ontologies. Our clustering-based ontology alignment approach relies on the following fundamental hypothesis:

Hypothesis (H): *The partitions of a source ontology (based on alignment results with different target ontologies) are stable across ontology alignments.*

If this hypothesis is valid, it is feasible to leverage the alignment result of ontology $\mathcal{S}$ to ontology $\mathcal{T}_1$ to help a new alignment of $\mathcal{S}$ to ontology $\mathcal{T}_2$ as follows:

- Generate a partition (*i.e.,* a set of clusters) of $\mathcal{S}$, denoted as $\mathcal{P}_s$, from the alignment result of $\mathcal{S} \rightarrow \mathcal{T}_1$;
- To perform the alignment task of $\mathcal{S} \rightarrow \mathcal{T}_2$, instead of matching individual entities in $\mathcal{S}$ independently with the entities in $\mathcal{T}_2$, it may be more efficient and more accurate to match a cluster of entities in $\mathcal{P}_s$ to the entities in $\mathcal{T}_2$. The intuition is that the entities in one cluster are expected to match to the same entity in $\mathcal{T}_2$.

This approach would be particularly valuable to maintain alignments as ontologies evolve. For example, if a high-quality alignment from ontology $\mathcal{S}$ to ontology $\mathcal{T}_1$ has been produced through a manual or semi-automated process and ontology $\mathcal{T}_1$ then evolves to a new version $\mathcal{T}_1$', this approach would significantly reduce the amount of pairwise mappings to consider in order to build an alignment from $\mathcal{S}$ to $\mathcal{T}_1$'.

IFW	**CBM**
Provide FMO Transaction Reconciliation	Account Reconciliation
Request Amended Counterparty Confirmation	Account Reconciliation
Accumulate Futures Transaction Values	Account Reconciliation
Analyze FMO Transaction Details	Account Reconciliation
Compare FMO Transaction Details	Account Reconciliation
Verify FMO Transaction Details	Account Reconciliation

Table 1. Example of an IFW cluster based on manual alignment to CBM

Tables 1 and 2 show two examples of clusters obtained respectively through manual alignment and through automated alignment.

In Table 1, most entities in the IFW cluster (*i.e.,* 'Provide FMO Transaction Reconciliation', 'Request Amended Counterparty Confirmation', 'Accumulate Futures Transaction Values', and 'Analyze FMO Transaction Details') show little to no lexical or structural similarity between themselves or with the target

Mouse Anatomy	**Brenda Tissue**
intestine (no synonym)	intestine (synonyms: bowel, gut)
bowel (no synonym)	intestine (synonyms: bowel, gut)
gut (no synonym)	intestine (synonyms: bowel, gut)

Table 2. Example of a Mouse Anatomy cluster based on lexical alignment to Brenda Tissue Ontology

CBM entity, 'Account Reconciliation'. In fact, applying standard similarity functions to directly map IFW to CBM produce extremely poor results because, as mentioned in Section 1, the two models are very different from almost all perspectives: different vocabularies (IT vocabulary for IFW vs. business vocabulary for CBM), very different structures (deep nested structure for IFW vs. flat structure for CBM), modeling at different levels of abstraction (modeling at the IT process level for IFW vs. modeling at the business functions level for CBM). The semantic similarity between IFW entities in the cluster, which could not be computed from information present in both models, was indirectly identified by the domain experts (IBM consultants) when they map these IFW entities to the same CBM entity.

Table 2 shows partial results of aligning the adult Mouse Anatomy Ontology (MA) and Brenda Tissue Ontology (BTO) using the automated process [4] described in [9]. Like the IFW-CBM case, entities in the cluster of MA ontology do not exhibit any meaningful similarity that could be computed based only on information in MA ontology. However, as opposed to the previous IFW case, entities in MA are lexically similar to the mapped entity (*i.e., intestine* which has as explicit synonyms *bowel* and *gut*) in the target ontology. In this case, the alignment to BTO serves as a dictionary look up that uncovers the semantic similarity between *intestine*, *bowel*, and *gut*. This uncovered semantic similarity could then be used in the next alignment involving MA ontology.

3 Measures of Cluster Similarity

To test our stability hypothesis (H), we need to evaluate the similarity between the partitions of the same ontology, which requires a similarity measure on a pair of partitions (*i.e.,* sets of clusters). To ease presentation, consider two alignments, $\mathcal{S}$ to $\mathcal{T}_1$ and $\mathcal{S}$ to $\mathcal{T}_2$. Based on their alignment results, we can generate two partitions of $\mathcal{S}$: $\mathcal{P}_{s,1} = \{C_1, C_2, \ldots, C_m\}$ and $\mathcal{P}_{s,2} = \{C'_1, C'_2, \ldots, C'_n\}$, where each cluster C_i or C'_j is a collection of entities in the source ontology $\mathcal{S}$. So the real challenge is to define an appropriate measure to evaluate the similarity of $\mathcal{P}_{s,1}$ and $\mathcal{P}_{s,2}$. A good measure needs to be symmetric and have a fixed range of values, preferably $[0, 1]$, such that similarity values computed for different pairs of partitions are comparable. Here we consider two similarity measures which are conceptually similar to similarity metrics for strings.

[4] The ontologies and the alignments are available at http://bioportal.bioontology.org/

3.1 Measure I: Jaccard Similarity on Entity Pairs

For each cluster C in the partition $\mathcal{P}_s$ of ontology $\mathcal{S}$, we can generate all pairs of entities in cluster C. Thus, the partition $\mathcal{P}_s$ can be represented as the union of all the sets of entity pairs (one set per cluster in $\mathcal{P}_s$). The generated set of entity pairs is *equivalent* to the original partition in the sense that we can re-generate the partition from the set of entity pairs. For instance, consider a partition $\mathcal{P}_1 = \{\{a, b\}, \{c, d, e\}\}$. The corresponding set of entity pairs is $\mathcal{P}'_1 = \{\{a, b\}, \{c, d\}, \{c, e\}, \{d, e\}\}$. Note that given $\mathcal{P}'_1$, we can re-generate the original partition $\mathcal{P}_1$. For another partition $\mathcal{P}_2$ (say, $\mathcal{P}_2 = \{\{a, b, c\}, \{d, e\}\}$), we can also generate a set of entity pairs as $\mathcal{P}'_2 = \{\{a, b\}, \{a, c\}, \{b, c\}, \{d, e\}\}$. The similarity of the two sets P'_1 and P'_2 can then be computed with the standard Jaccard similarity [1] by treating each entity pair (without considering the sequence of entities) as the basic element of a set. Therefore, the similarity of $\mathcal{P}_1$ and $\mathcal{P}_2$ can be computed as follows:

$$PSim_1(\mathcal{P}_1, \mathcal{P}_2) = \frac{|\mathcal{P}'_1 \cap \mathcal{P}'_2|}{|\mathcal{P}'_1 \cup \mathcal{P}'_2|} \tag{1}$$

where the numerator is the size of set intersection, and the denominator is the size of set union, with each entity pair as a basic unit in the set. The similarity measure $PSim_1$ has the desired property that it is symmetric (*i.e.,* $Sim_1(P_1, P_2) = Sim_2(P_2, P_1)$) and the range of the similarity value is $[0, 1]$. Furthermore, $PSim_1$ captures the effect of big clusters in a partition, because big clusters will generate entity pairs that are exponential in size to cluster size; thus reflecting the natural preference for big clusters.

3.2 Measure II: Partition Edit Distance

One measure that is closely related to similarity is distance. The distance between two partitions can be intuitively characterized by the minimum amount of work to transform one partition into the other, which is conceptually similar to edit distance (*i.e.,* the minimum number of edits, including insertion, deletion, and substitution) between two strings. The basic operations for partitions we consider include *Split* and *Merge*. A Split operation on a cluster C_1 creates two non-overlapping clusters C_2 and C_3, with the union of C_2 and C_3 including all the elements in C_1. Merge is an inverse operation of Split. To continue with the previous example, to transform partition $\mathcal{P}_1$ into partition $\mathcal{P}_2$, we need 2 operations: a Split operation on the cluster $\{c, d, e\}$ generates two clusters $\{c\}$ and $\{d, e\}$; and a Merge operation of the two clusters $\{a, b\}$ and $\{c\}$ creates a new cluster $\{a, b, c\}$, thus resulting in partition $\mathcal{P}_2$. So the edit distance between $\mathcal{P}_1$ and $\mathcal{P}_2$ is 2.

Definition: The edit distance between two partitions $\mathcal{P}_1$ and $\mathcal{P}_2$, denoted as $ED(\mathcal{P}_1, \mathcal{P}_2)$, is the length of the shortest edit path composed of Splits and Merges from $\mathcal{P}_1$ to $\mathcal{P}_2$. A nice property of the partition edit distance is that it is symmetric, *i.e.,* $ED(\mathcal{P}_1, \mathcal{P}_2) = ED(\mathcal{P}_2, \mathcal{P}_1)$. Although the edit path from $\mathcal{P}_1$ to $\mathcal{P}_2$ is different from the path of transforming $\mathcal{P}_2$ to $\mathcal{P}_1$, these two paths have the same length, since the two basic operations of Merge and Split are inverse of each other.

Because the edit distance between two partitions of the same ontology is dependent on ontology size, we need a normalization factor to transform edit distance into a similarity measure. The normalization factor we consider here is ontology size, *i.e.,* the number of entities in a source ontology. Thus, the similarity measure derived from edit distance of two partitions $\mathcal{P}_1$ and $\mathcal{P}_2$ is:

$$PSim_2(\mathcal{P}_1, \mathcal{P}_2) = 1 - \frac{1}{|\mathcal{S}|} ED(\mathcal{P}_1, \mathcal{P}_2) \tag{2}$$

where $|\mathcal{S}|$ is the size of the source ontology. The similarity measure $PSim_2$ is also symmetric.

3.3 Measure III: Mapping Quality

The above two measures reflects the stability of partitions from the similarity perspective. We also propose another measure to evaluate the actual quality of mappings which are generated based on the clustering information. To this end, we simulate the procedure of generating partitions of a source ontology and applying the clustering information for a new alignment that involves the same source ontology:

- Generate a partition $\mathcal{P}_1$ of the source ontology $\mathcal{S}$ based on the mapping result from $\mathcal{S}$ to a target ontology $\mathcal{T}_1$;
- For a new alignment task from $\mathcal{S}$ to another target ontology $\mathcal{T}_2$, generate the mappings as follows:
  - For each cluster C in the partition $\mathcal{P}_1$, randomly pick one entity s from C and find the mapped entity t in $\mathcal{T}_2$;
  - Generalize the mapping to other entities in the same cluster, with the mappings being $\{\langle s', t\rangle | s' \in C\}$.

Since in this paper we focus on many-to-one mappings, we exclude the one-to-one mappings from the estimation of `precision` and `recall`, the two classical metrics for measuring mapping quality.

$$\texttt{precision} = \frac{|\texttt{M} \cap M_{GS}|}{|\texttt{M}|}, \texttt{recall} = \frac{|\texttt{M} \cap M_{GS}|}{|M_{GS}|}$$

where `M` is the mappings generated using the strategy described above, and M_{GS} is the gold-standard (*i.e.,* reference) mappings. Note that $|\texttt{M}|$ and $|M_{GS}|$ are equal in this scenario, so `precision` and `recall` are equal, and we will only report results in `precision` in the experiment section. In addition to the mapping quality, we also measure the amount of saving of human effort to generate the mappings, compared to the baseline approach of independently generating mappings for each entity in the source ontology from scratch. The human effort is estimated as the number of mappings that require human input. We thus define mapping efficiency with utilization of clustering information as:

$$\texttt{efficiency} = 1 - |\mathcal{P}_{m2o}|/|\mathcal{S}_{m2o}|$$

where $|\mathcal{P}_{m2o}|$ is the number of *non-singleton* clusters (*i.e.,* clusters with more than one entity) in the partition $\mathcal{P}$ of the source ontology $\mathcal{S}$, and $|\mathcal{S}_{m2o}|$ is the total number of entities in the non-singleton clusters. Intuitively, the bigger the clusters, the more efficient the approach based on clustering. At the same time, however, bigger clusters tend to be more *impure* (*i.e.,* meaning entities in the same cluster are mapped to different entities in the target ontology). Therefore, clusters of size exceeding the optimal value will adversely affect mapping quality.

3.4 Discussion

The three measures described above reflect different information aspects for the hypothesis testing about partition stability. The Jaccard similarity indicates whether the partitions generated based on mappings to different target ontologies are at the same granularity. For example, if the target ontology $\mathcal{T}_1$ is more fine-grained than another target ontology $\mathcal{T}_2$, we expect that the Jaccard similarity of the two partitions of the source ontology is relatively low. A simple example will illustrate this fact. Suppose we have one partition containing just one cluster $\{a, b, c, d\}$, and the other partition is $\{\{a, b\}, \{c, d\}\}$. It is easy to see that the target ontologies in the two alignments are at different granularity. The Jaccard similarity of the two partitions is 1/3, which is relatively low. The partition edit distance, on the other hand, is insensitive to such partition granularity. Continue that simple example. We can see that the edit distance between the two partitions is 1, so the normalized similarity based on the edit distance is 1 - 1/4 = 0.75. The advantage of edit distance is that it can capture both the cases where entities mapped to the same entity are mapped to different entities in another target ontology, and the cases where entities mapped to different entities in one target ontology are mapped to the same entities in another ontology. The third measure, mapping quality, provides information about whether the partition information is reliable for end use. That is, how much the users can trust the partition information provided by one alignment task, when they perform a related alignment task in the same domain with the same source ontology. In some sense, mapping quality is a hybrid measure of Jaccard similarity and partition edit distance, and can provide an estimate of usefulness of the clustering information for end users.

4 Evaluating Partitioning Stability

In this section, we evaluate the stability of partitioning, using the three measures defined in Section 2, on one dataset from the financial domain and one from the life sciences domain that is publicly available on the BioPortal website.

4.1 IFW - CBM: Ontology Evolution Scenario

As discussed earlier in Section 1, we first studied the case where we had two high-quality manual alignments: IFW-CBM$_1$ and IFW-CBM$_2$, where CBM$_2$ reflects an evolution of CBM$_1$. CBM$_1$ has 65 entities, and CBM$_2$ has 120 entities; they overlap in 37 entities. There are 2165 entities in IFW that are involved in many-to-one mappings. The partition of IFW based on the mappings from IFW to CBM$_1$ consists of 62 clusters, and the partition based on the mappings from IFW to CBM$_2$ consists of 111 clusters. The average cluster size in both partitions is 25. Recall that the average cluster size determines the mapping efficiency, *i.e.,* the amount of human effort that can be saved by leveraging the clustering information. Therefore, the mapping efficiency in this case is expected to be high; the actual efficiency value is 0.95. We also calculated the similarity of the two partitions: (1) The similarity based on partition edit distance is 0.89; and (2) the Jaccard similarity is 0.53. The low Jaccard similarity is likely due to the fact that the number of clusters in two partitions is quite different (62 vs. 111), as is the cluster size. As a consequence, the number of entity pairs generated from the clusters of IFW entities changes significantly. Because Jaccard similarity is quite sensitive to the size of the sets of entity pairs, the two partitions have a low Jaccard similarity. Jaccard similarity clearly reflects the actual change in granularity of the two versions of CBM.

The mapping precision metric is not symmetric, which means using the clustering information based on the mappings from IFW to CBM$_1$ to generate mappings for IFW to CBM$_2$ may have a precision quite different from that in the other direction. Therefore, we estimated mapping precision in both directions, and the average precision is 0.78. To determine the improvement in precision due to the use of clustering information, we measured the overlap between the two alignment results (*i.e.,* IFW-CBM$_1$ and IFW-CBM$_2$) as the baseline. The intuition is that, if we directly use one alignment result to generate mappings for the other alignment, only the overlap of the two alignments can generate correct mappings; the precision for this approach is 0.38. So through the utilization of clustering information from one alignment for the other alignment, we improve the mapping precision by 0.4; which is statistically significant. As mentioned in Section 1, the lexical and structural similarity between IFW and CBM is extremely low; we actually ran our alignment algorithm [5] for the two alignments IFW-CBM$_1$ and IFW-CBM$_2$ and got a precision around 0.01. In this scenario, manual mapping is therefore a must, and improving precision by 0.4 by alignment re-use is a significant saving.

4.2 Large Scale Evaluation on BioPortal Ontologies

The BioPortal website contains 149 ontologies, 9.3K ontology comparisons, and 1.75 million matchings of elements in various ontologies that were largely lexically generated.

Recall that we can create one partition of the source ontology from one ontology alignment result. For a given source ontology $\mathcal{S}$, there could be multiple

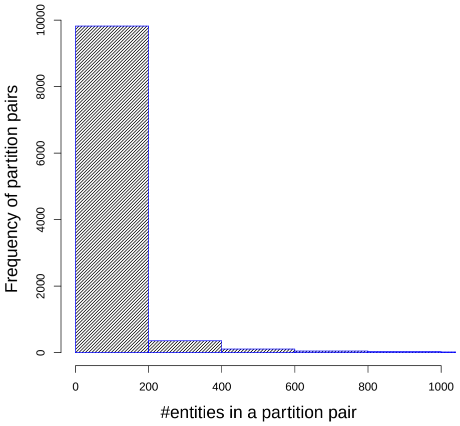

Fig. 1. Histogram of #entities in partition pairs

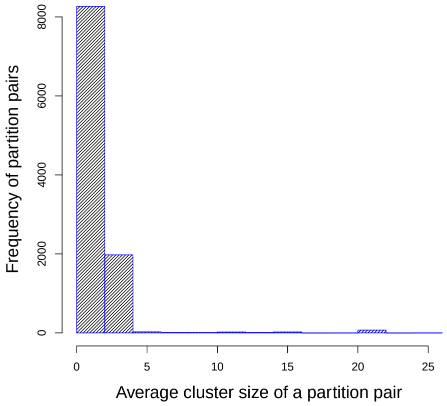

Fig. 2. Histogram of average cluster size per partition

partitions of $\mathcal{S}$, based on the alignment results with respect to different target ontologies. We used the two similarity measures (described in Section 3) to estimate pairwise similarity of the partitions on the same source ontology. If an ontology $\mathcal{S}$ is aligned with k ontologies, we will generate k partitions of $\mathcal{S}$, and there will be $\binom{k}{2}$ similarity computations of the pairs of partitions. Therefore, the total number of pairwise comparison of partitions is $\sum_{i=1}^{K} \binom{k_i}{2}$, where K is the number of ontologies, and k_i is the number of times an ontology $\mathcal{S}_i$ is aligned with other ontologies. In this setting, we have altogether 24K similarity computations between generated partitions.

Since we were focused on many-to-one matching scenarios, we needed to preprocess the expected matchings from the BioPortal website before analyzing the similarity of partitions on the same source ontology, using the following steps:

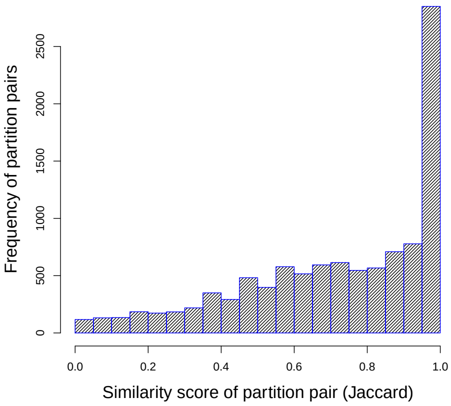

Fig. 3. Histogram of Jaccard similarity values

1) For a pair of partitions on the same source ontology, we identified the entities that appeared in both partitions. We removed from further analysis those entities that only appeared in one of the partitions. The rationale for this pruning was that these entities were really analogous to missing observations. That is, if an entity is missing from the partition it could be either due to incomplete alignment by domain experts, or because it is a singleton in this alignment, or because it should have been mapped to a different cluster. Since we had no way of knowing which of the three cases these entities fell into, we basically eliminated the entities from the analysis.
2) For any entity that is a singleton cluster in both partitions, we also removed them from further analysis; although the singleton clusters common in two partitions do not actually affect the similarity values, due to the robustness of our similarity measures.
3) To make the analysis meaningful, we also removed ontology comparisons that contained less than 10 entities involved in many-to-one matchings.

After preprocessing the expected matchings, we had 10.4K pairs of partitions for the similarity comparison. Figure 1 shows the distribution of the number of entities in each pair of partitions. The average number of entities involved in many-to-one matching scenarios is 64, which ensures that our analysis of partitioning stability is based on a reasonable number of data points and is reliable. Figure 2 shows the distribution of average cluster size per partition. It is clear that a majority of the partitions have small clusters, with a size of 2 or 3; note that we have excluded singleton clusters generated from one-to-one mappings. Since the mapping result is incomplete and often only covers a small part of the ontology, we made only considered the entities mentioned in both matchings, which partially explains small clusters.

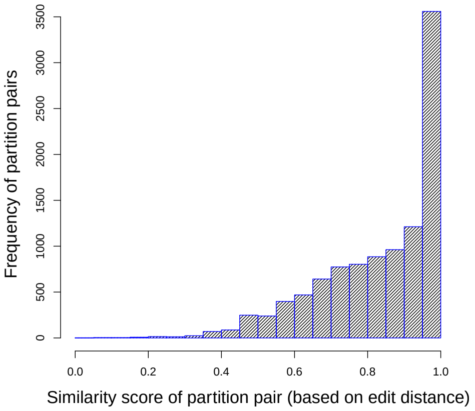

Fig. 4. Histogram of similarity values based on partition edit distance

Figure 3 shows the distribution of the similarity values of all pairs of partitions using Jaccard similarity on entity pairs (see Section 3.1). The mean of the similarity values is 0.72, and the standard deviation is 0.26. Figure 4 shows the distribution of the similarity values of all pairs of partitions based on partition edit distance (see Section 3.2). The mean of the similarity values is 0.84, and the standard deviation is 0.16.

Both Figure 3 and Figure 4 show that the partition of an ontology $\mathcal{S}$ is reasonably stable based on the results of aligning $\mathcal{S}$ with different ontologies. This observation indicates that we can leverage the partition of ontology $\mathcal{S}$ constructed from an existing alignment result to help new ontology alignments, which can be done in the following way: (1) Given the result of aligning $\mathcal{S}$ to $\mathcal{T}_1$, we can generate a partition (*i.e.,* clusters) of $\mathcal{S}$, denoted as $\mathcal{P}_s$; (2) For a new alignment from $\mathcal{S}$ to $\mathcal{T}_2$, we match each cluster of entities in $\mathcal{P}_s$ to the same entity in $\mathcal{T}_2$. This alignment strategy has two benefits: (i) it improves alignment quality, since the alignment tool can aggregate the information from all entities in a cluster to make alignment decisions rather than make decisions based on individual entities independently; and (ii) it improves alignment efficiency, because the alignment of one entity in a cluster can be easily generalized to the other entities in the same cluster. Figure 5 shows the distribution of precision when we apply the mapping strategy to the 10.4K ontology pairs. The average precision is 0.92, with a standard deviation of 0.11. This result verifies that it is viable to utilize the clustering information from one ontology pair for the alignment of another pair in the same domain, certainly with the same source model. Figure 6 shows the distribution of mapping efficiency in terms of the percentage of mappings that can be automatically generated by leveraging the partition information. The average efficiency is 0.37, with a standard deviation of 0.19. As explained in the previous section, the efficiency is highly dependent

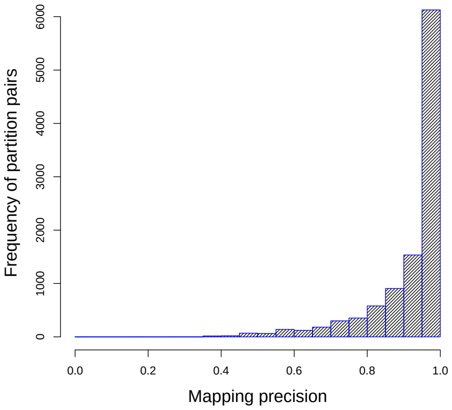

Fig. 5. Histogram of mapping precision values

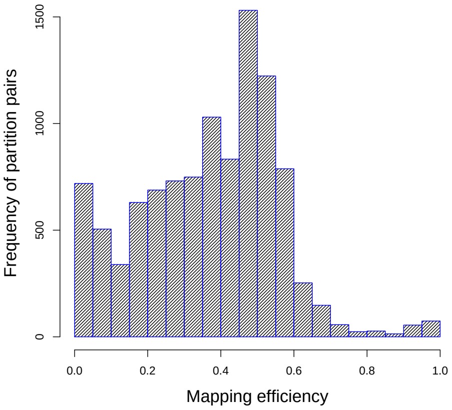

Fig. 6. Histogram of mapping efficiency values

on the average cluster size; the bigger the average cluster size, the higher the efficiency. Since the average cluster size of the partitions is small (see Figure 2), the efficiency is thus modest.

We also note that the observed stability of clusters for the BioPortal ontologies is not simply an artifact of the fact that the mappings were computed using lexical matching. For instance, the concepts *COO:F0005386 hyaluronidase activity*, *CCO:F0004395 hyaluronate lyase activity*, and *CCO:F0000824 hyaluronoglucosaminidase activity* are all mapped to *PHI:0000199 hyaluronidase activity* based on their broad synonyms. Yet, each of the 3 concepts is mapped to different concepts in the Gene Ontology (GO). It is clear that stability is independent of whether or not lexical similarity drives the alignment process, which was also shown earlier with the IFW-CBM alignments.

Can the use of clustering information improve the alignment for BioPortal ontologies? Unfortunately, we do not have the luxury of having overlaps between two versions of the same model, like the IFW-CBM case, that can be used as a baseline. We do note, however, that there were a substantial number (48,261) of mappings, generated by our clustering-based approach, that are missing from the BioPortal website. Since the mappings provided by BioPortal are incomplete [9], it is unclear whether some entities in part of a cluster should not be mapped to any entity in the target ontology or the extra mappings we found are valid. Although we were unable to verify the validity of all the mappings due to lack of expertise, a number of them seemed correct based on their synonyms (see Table 3 for a few examples below). In the table, CLL is missed because it is an acronym for chronic lymphocytic leukemia, lung neoplasms is missed because it is a synonym of lung cancer, and similarly, RB1 is missed because it is an acronym for retinoblastoma. This observation indicates that our clustering-based alignment approach can improve the recall for the alignments of BioPortal ontologies; note that the average precision estimated with the existing mappings is 0.92.

Concept 1	**Concept 2**
estrogen receptor alpha (CDR0000322904)	estrogen receptor (PRO_000007204)
retinoblastoma (MPATH:378)	RB1 (CDR0000043571)
non-small cell lung cancer (CDR0000040862)	Lung neoplasms (D008175)
renal cell carcinoma (CDR0000038140)	carcinoma, renal cell (C1534)
B-cell chronic lymphocytic leukemia (CDR0000039824)	CLL (LP34550-1)

Table 3. Examples of missed matches as defined by clustering

5 Related Work

The alignment technique we proposed in this paper, which exploits internal structures of ontologies discovered through existing high-quality alignments, can be contrasted with previous work in terms of its singular focus on many-to-one and one-to-many alignments and in terms of the novelty of its approach to learning from existing alignments.

Although many approaches have been proposed to perform ontology alignment in the literature, there have been, to the best of our knowledge, no significant efforts to tailor the alignment process for alignments with cardinality different from one-to-one. After computing an aggregate similarity score for each candidate matching, most state-of-the art systems (*e.g.,* AgreementMaker [3] and BLOOMS [15]) simply return the matchings above a given threshold under a given alignment cardinality constraint (*e.g.,* one-to-one, one-to-many, many-to-one) without any consideration for the internal structures implied by one-to-many or many-to-one alignments. Other systems (*e.g.,* ASMOV [12]) have been optimized for one-to-one alignments to the point of considering multiple entity correspondences, where the same entity in one ontology is matched with multiple entities in the other ontology, as an inconsistency check in the final semantic verification step. This bias for one-to-one alignments also transpires from the relatively large collection of mostly one-to-one ontology alignments used to evaluate and systematically characterize the performance of state-of-the-art ontology alignment systems at the annual Ontology Alignment Evaluation Initiative [5] event.

Prior work on learning from existing high-quality alignments (*e.g.,* [5], [4], [7] and [6]) has typically taken a machine learning approach to customize the alignment process either for a given pair of ontologies, for which a partial reference alignment is available, or for a domain where multiple reference alignments are available. The outcome of this traditional learning approach is the specification of the optimal value for each parameter of the alignment process for a particular alignment or for alignments in a given domain. However, the learning approach does not work well when there is little or no lexical/structural similarity between the ontologies to align; in which cases the similarity functions can provide little signal for the learning process. Furthermore, no information is learned about the intrinsic structure of ontologies and then used to help new alignment. In contrast, in this paper, we describe how existing many-to-one (or one-to-many) alignments can be used to discover internal structures (*i.e.,* grouping entities within an ontology); such structures can then be leveraged in new ontology alignment as discussed in Section 2.

To the best of our knowledge, [9] and [10] are the only related work which attempts to learn structural characteristics of ontologies from matchings. However, our work is different from [9] in terms of its goals. The main goal of [9] is to uncover the network structure of the set of ontologies, and learn from their links (*i.e.,* entity matchings) the interesting properties of the ontologies in the particular domain; for example, which ones are the most relevant and most appropriate to serve as background knowledge for domain-specific tools. Our goal is to uncover internal ontological structures to enhance future alignments. Reference [10] proposed an alignment technique to generate mappings between source ontology and target ontology by composing previously determined mappings that involve intermediate ontologies. Our work differs from [10] in that we evaluated the soundness of the hypothesis that the partition (*i.e.,* clustering of entities) of the source ontology is stable across ontology alignments, which validates the underlying assumption made by [10]; so our work is complementary to [10].

6 Conclusions

In this paper we proposed the hypothesis that the internal structure of an ontology, *i.e.,* clusters of its entities discovered from the many-to-one alignment scenario, is stable across ontology alignments in the same domain. To evaluate this hypothesis, we defined two novel metrics to measure the similarity of

[5] http://oaei.ontologymatching.org/2010/

clusters generated for one ontology based on its alignments with different target ontologies. Experimental evaluation with datasets from the financial domain and the healthcare and life sciences domain demonstrated that the stability hypothesis is valid. In addition, we designed a mapping strategy that can leverage the clustering information for new alignment tasks, and characterized the effectiveness of this mapping strategy in terms of the impact on mapping quality and mapping efficiency. Experimental evaluation showed that clustering information discovered from one alignment can help improve, with a statistical significance, the mapping quality and mapping efficiency of a new alignment.

References

1. Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2007.
2. Namyoun Choi, Il-Yeol Song, and Hyoil Han. A survey on ontology mapping. *SIGMOD Rec.*, 2006.
3. Isabel F. Cruz, Flavio Palandri Antonelli, and Cosmin Stroe. Agreementmaker: efficient matching for large real-world schemas and ontologies. *Proc. VLDB Endow.*, 2:1586–1589, August 2009.
4. Anhai Doan, Jayant Madhavan, Pedro Domingos, and Alon Halevy. Ontology matching: A machine learning approach. In *Handbook on Ontologies in Information Systems.* Springer, 2003.
5. Songyun Duan, Achille Fokoue, and Kavitha Srinivas. One size does not fit all: Customizing ontology alignment using user feedback. In *ISWC*, 2010.
6. Kai Eckert, Christian Meilicke, and Heiner Stuckenschmidt. Improving ontology matching using meta-level learning. In *ESWC*, 2009.
7. Marc Ehrig, Steffen Staab, and York Sure. Bootstrapping ontology alignment methods with APFEL. In *ISWC*, 2005.
8. Yves R. Jean-Mary et al. ASMOV: Results for OAEI 2009. In *OM*, 2009.
9. Amir Ghazvinian, Natalya F. Noy, Clement Jonquet, Nigam Shah, and Mark A. Musen. What four million mappings can tell you about two hundred ontologies. In *ISWC*, 2009.
10. Anika Gross, Michael Hartung, Toralf Kirsten, and Erhard Rahm. Mapping composition for matching large life science ontologies. In *International Conference on Biomedical Ontology*, 2011.
11. Md. Seddiqui Hanif and Masaki Aono. Anchor-Flood: Results for OAEI 2009. In *OM*, 2009.
12. Yves R. Jean-Mary, E. Patrick Shironoshita, and Mansur R. Kabuka. Ontology matching with semantic verification. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):235 – 251, 2009. The Web of Data.
13. Juanzi Li, Jie Tang, Yi Li, and Qiong Luo. RiMOM: A dynamic multistrategy ontology alignment framework. *IEEE Trans. Knowl. Data Eng.*, 2009.
14. Natalya F. Noy. Semantic integration: a survey of ontology-based approaches. *SIGMOD Rec.*, 2004.
15. Catia Pesquita, Cosmin Stroe, Isabel Cruz, and Francisco M. Couto. Blooms on agreementmaker: Results for oaei 2010. In *OM*, 2010.
16. Peng Wang and Baowen Xu. Lily: Ontology alignment results for OAEI 2009. In *OM*, 2009.