Reducing the search space in ontology alignment using clustering techniques and topic identification (Extended version)

Agnese Chiatti[1], Zlatan Dragisic[2], Tania Cerquitelli[1], and Patrick Lambrix[2]

[1] Department of Control and Computer Engineering, Politecnico di Torino, Italy
[2] Computer Science / Swedish e-Science Research Center, Linköping University, Sweden

Abstract. One of the current challenges in ontology alignment is scalability and one technique to deal with this issue is to reduce the search space for the generation of mapping suggestions. In this paper we develop a method to prune that search space by using clustering techniques and topic identification. Further, we provide experiments showing that we are able to generate partitions that allow for high quality alignments with a highly reduced effort for computation and validation of mapping suggestions for the parts of the ontologies in the partition. Other techniques will still be needed for finding mappings that are not in the partition.

Keywords: Knowledge representation, data mining, ontology alignment

This is an extended version of a paper published at K-CAP 2015, the 8th International Conference on Knowledge Capture. The K-CAP paper is published by ACM with doi: 10.1145/2815833.2816959.

1 Introduction

In recent years many ontologies have been developed and many of those contain overlapping information. Often we want to use multiple ontologies. For instance, companies may want to use community standard ontologies in conjunction with company-specific ontologies. Applications may need to use ontologies from different areas or from different views on one area. In each of these cases it is important to know the relationships between the concepts (and relations) in the different ontologies. Further, the data in different sources within the same domain may have been annotated with different but similar ontologies. Knowledge of the inter-ontology relationships would in this case lead to improvements in search, integration and analysis of data. To address this major issue, much research has been recently devoted to ontology alignment (OA), i.e., to finding mappings between concepts and relations in different ontologies (e.g., [4]). The research field of OA is very active with its own yearly workshop as well as a yearly event, the Ontology Alignment Evaluation Initiative (OAEI, e.g., [5]), that focuses on evaluating systems that automatically generate mapping suggestions. Many systems have been built and overviews can be found in e.g., [4, 15] and at the ontology matching web site http://www.ontologymatching.org.

Some recent work (e.g., [15, 9]) has defined challenges that need to be addressed when dealing with *large* ontologies. Among those, scalibility is an important challenge as shown by the fact that many participants in the OAEI have performance problems when dealing with large ontologies. One technique to deal with scalabiblity is to reduce the search space for the generation of mapping suggestions. Instead of dealing with all concept pairs (C_1, C_2), with C_1 belonging to the first ontology and C_2 to the second ontology, only a subset of these pairs is considered. Each such subset is called a mappable part. Reducing the search space leads to reduced computation during the generation of mapping suggestions. Further, the expectation is that fewer mapping suggestions are computed and therefore the validation effort for the domain experts, who decide whether a mapping suggestion is correct or not, should be reduced.

Most of the previous approaches for reducing the search space focused on segmenting or partitioning the ontologies using an initial alignment. For instance, in [6] subgraphs of concepts with good coverage of the initial mappings are selected. In [11] the initial alignment is used to partition the ontologies based on their structure. In [8] the ontologies are partitioned using a variant of ROCK [7] exploiting a link measure based on structural and linguistic similarities between concepts. In all these approaches the selection of the mappable parts of the different ontologies is based on the existence of mappings in these parts. Some approaches use the locality of anchors (concepts in mappings), i.e., descendants, ancestors and neighbors, to reduce the search space [14, 16]. In [1] the selection of concepts is based on (sub)schemas. The selection of the mappable (sub)schemas is based on similarity of concepts (e.g., their names) in the (sub)schemas.

In this paper we describe a method to prune the search space for OA that does not require an initial alignment. As in other approaches, we focus on aligning concepts. We do this by using clustering techniques and topic identification before the actual generation of mapping suggestions (Section 4). Further, the method was demonstrated in a set of experiments (Section 5). We use different clustering techniques and OA systems on the ontologies of the Anatomy track of the OAEI, and show that we are able to generate partitions that allow for high quality alignments while highly reducing the effort for computation and validation of mapping suggestions for the parts of the ontologies in the partition. Other techniques will still be needed for finding mappings that are not in the partition. Finally, a summary and future work are given in Section 6.

2 Data Mining

Within data mining, i.e. the process of extracting significant patterns from a raw data set, cluster analysis aims at defining reliable subgroups from the starting data set, given some notion of similarity between data items in the groups. **Clustering Techniques.** K-means is a partitional clustering technique attempting to find a user-specified number of clusters (K), represented by their centroids. The algorithm guesses the initial centroids and populates the K clusters by assigning each data item to its closest centroid. Then, each centroid is updated based on the items assigned to each cluster. The assignment and update steps are repeated until there is no change in the cluster configuration. Predicting a good value for K is not easy, but a parameter sensitivity analysis can help. A possible measure of accuracy of the final clustering is the silhouette index [13], based on intra-cluster cohesion and inter-cluster separation. For algorithms using Euclidean distance, accuracy can be derived from the sum of the squared errors (SSE), which is based on the distance between data items and their centroids. In the SSE-min approach, the value for K that minimizes the SSE value is selected, while the SSE-max approach maximizes the marginal decrease in the SSE curve.

On the contrary, DBSCAN [3] is a density-based algorithm where density is defined by counting the number of data items within a specified radius of a data item. The algorithm gathers data items that are density connected in the same cluster. Essentially, two data items are density connected if they can both be reached from the same data item by following chains of data items in dense regions.

3 Ontology Alignment

From a knowledge representation point of view, ontologies may contain concepts, relations, axioms and instances. Concepts and relations are often organized in hierarchies using the is-a relation. The task of OA is to create an alignment between ontologies, i.e., a set of mappings between entities from the different ontologies. The most common kinds of mappings are equivalence mappings and mappings using is-a and its inverse.

A large number of OA systems have been developed: many are based on the computation of similarity values between entities in two input ontologies and can be described as instantiations of the framework in Figure 1. Part I (computation of mapping suggestions) contains different components. A preprocessing component can be used to modify the original ontologies, e.g., to extract specific features of the concepts in the ontologies, or to reduce search space for finding mapping suggestions. This specific component is the focus of this paper. The matching component includes one or several several matchers that calculate similarities between the entities from the different ontologies. Mapping suggestions are then determined by combining and filtering the results generated by one or more matchers. By using different preprocessing, matching, combining and filtering techniques, we obtain different alignment strategies. The result of part I is a set of mapping suggestions. In part II, the mapping suggestions are presented to a domain expert, who accepts or rejects them. The accepted mapping suggestions are part of the final alignment. Any sub-set of the final alignment is a partial alignment (PA). The acceptance and rejection of suggestions may also influence further suggestions. Further, a conflict checker or a mapping debugging system could be used to avoid conflicts introduced by the mapping suggestions. There can be several iterations of parts I and II.

4 Workflow

We present a general workflow for the generation of mappable parts in two ontologies. The detailed algorithms are described in Section 5.

In the *data collection and preprocessing phase*, the ontologies are represented in a suitable format for clustering. In the *cluster analysis phase* a clustering algorithm is

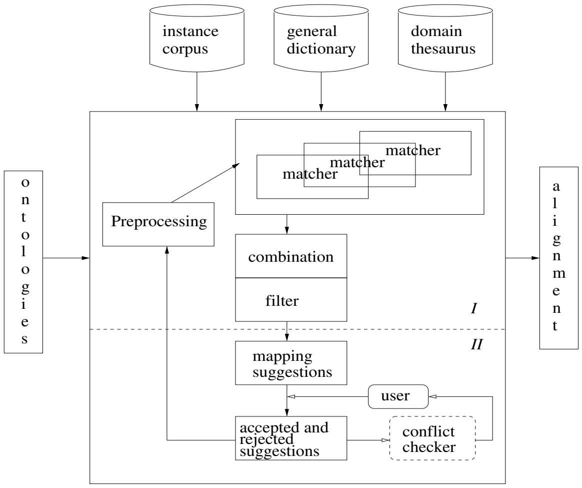

Fig. 1. An existing framework (extension of [12]).

run on the output of the previous phase. A sensitivity analysis is used to select the best parameter values for the clustering algorithms. In the *text mining preprocessing phase* the clusters are treated as documents (using the labels of the concepts in the clusters) and processed to obtain a suitable format for the next phase. In the *topics identification phase* we identify the main topics for the cluster documents. Clusters from the same ontology are merged into larger clusters based on their topics. Finally, in the *mappable parts generation phase* clusters from different ontologies are connected into mappable parts using the topics.

Table 1. Topics and mappable parts.

Topic	AMA clusters	NCI-A clusters	Keywords
Arteries	1, 12, 22, 48, 50	12, 19, 20, 26, 58, 80	AMA: Arteries, Blood, Vessel, Trunk, Dorsal, Digit, Common, Superior NCI-A: Arteries types, Arteries, Anterior, Circumflex, Internal, External, Superior, Articular, Branch, Inferior
Bones	8, 36, 60, 63	55, 66, 70, 81, 82	AMA: Joint, Bone, Vertebra, Digit, Foot, Cervical, Phalanx, Lumbar, Metacarpe, Metatars, Carpal, Sacral, Rib, Thoracic, Hand NCI-A: Vertebra, cartilage, Rib, Bone, Head, Phalanx, Foot, Digit, Hand
Hair parts	–	14, 61	NCI-A: Follicles, Hair, Stratum

5 Experiments

Setup. For the experiments we used the following setup. We used the **ontologies** of the Anatomy track of the OAEI, which contains the ontologies Adult Mouse Anatomy

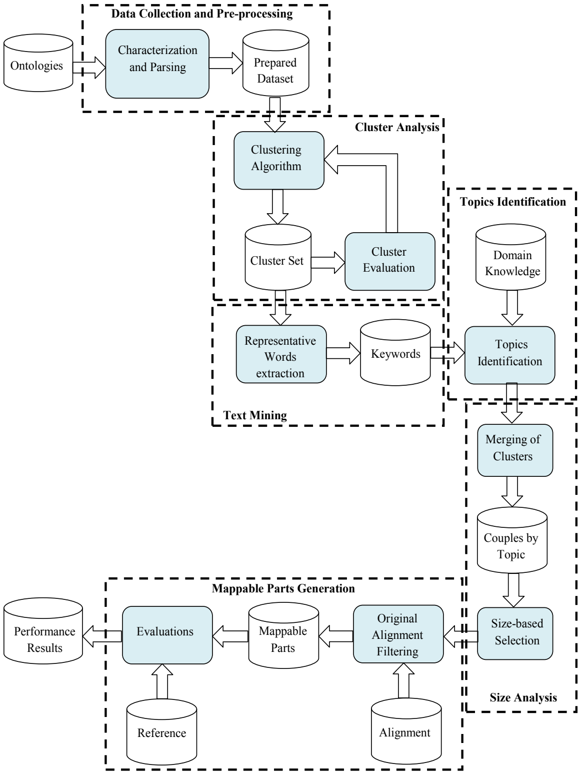

Fig. 2. The proposed framework for search space reduction in OA.

(AMA) and the anatomy part of the NCI Thesaurus (NCI-A). AMA contains 2737 concepts and NCI-A contains 3298 concepts, giving 9,026,626 potential concept pairings. Further, a reference alignment with 1516 equivalence mappings is available.

We **represented** the ontologies as binary symmetric matrices where $M_{i,j} = 1$ if concept i and concept j are related via is-a (regardless of level), and 0 otherwise. For **clustering** we used both K-means and DBSCAN. As similarity measures we used the Jaccard similarity for K-means and the cosine similarity for DBSCAN. For the K-means sensitivity analysis, we used several techniques (silhouette index, SSE-min and SSE-max). In the **text mining preprocessing** phase we tokenized the concept labels in each cluster, converted them into lower case, removed stop words, performed stemming and truncated tokens longer than 25 characters. Finally, based on the resulting tokens the cluster was represented using a tf-idf vector, each containing a list of tokens which we consider keywords for the clusters. In the **topics identification** phase we define, based on the keywords, a set of topics of appropriate granularity per ontology. In the current experiment, this phase was performed manually. Clusters in the same ontology with the same topic are merged. Then we relate the topics of the two ontologies to obtain **mappable parts**.

For each approach we investigated its influence on the performance of existing **OA systems**: SAMBO [12], a system that traditionally performed well in the OAEI Anatomy track as well as all participating systems in the OAEI Anatomy track in 2014 [2]. We used the standard measures from the OAEI. Precision is the ratio of the number of found correct mappings to the total number of mapping suggestions. Recall is the ratio of the number of found correct mappings to the total number of correct mappings. The F-measure is the harmonic mean of precision and recall. We also computed how much the search space was pruned.

Results. Results are reported separately for 4 clustering techniques: DBSCAN and K-means, where K was determined based on SSE-min (KM-SSE-min), SSE-max (KM-SSE-max) and silhouette index (KM-sil). For each of these we generated clusters, keywords, topics and mappable parts. Table 1 shows some topics and mappable parts generated using KM-sil. The topic 'Arteries' consists of 5 original clusters from AMA and 6 from NCI-A. The most common keywords in these parts are also shown. The topic 'Hair parts' only appears in NCI-A and thus is not contained in a mappable part. For KM-sil 13 mappable parts are generated (Table 2), for SSE-min 7, for SSE-max 12 and for DBSCAN 13.

Regarding the quality of the alignments, Table 2 shows the results for KM-sil combined with SAMBO per topic-related mappable part with as reference the projection of the full reference alignment on each mappable part. The projection contains only the concept pairs in the full reference alignment for which both concepts are included in the mappable part. Table 3 shows the results for all clustering algorithms combined with all variants of each system when considering all topics. In the table, for each measure a pair of values is given, where the first is computed with respect to the full reference alignment, and the second with respect to the projection of the full reference alignment on the mappable parts.

Table 4 shows in the first column how many concept comparisons need to be performed by the OA systems for each clustering algorithm. The number of comparisons is reduced to a small fraction (1.8% to 2.9%) of the potential pairings. The number of mapping suggestions to be validated is reduced to a fraction which ranges from 11% to 43%, depending on the OA system and the specific clustering technique.

Discussion. The approaches allowed us to partition the ontologies and use mappable parts of the ontologies for which the OA systems would perfom well. We found similar behavior for the OA systems. In general, the precision and recall of the OA systems is often better in the mappable parts than on whole ontology (see Table 3). An expected disadvantage is that, as fewer suggestions are generated, the recall compared to the full reference alignment is low. However, when compared to the computation and validation effort, the proposed approaches found regions where the amount of effort leads to a much improved return (Table 4). The computation effort when using the mappable parts is reduced by at least 97% for all systems, while the validation effort is reduced by at least 60%.

In our experiments KM-SSE-max always required the fewest comparisons and generated the fewest mapping suggestions. It also had the lowest recall compared to the full reference alignment (between 16% and 18% of the recall for the systems). KM-sil always required the most computation (264,536 pairs) and generated the most map-

Table 2. Results per topic-related mappable part for K-sil combined with SAMBO.

	Precision	F-Measure	Recall
SAMBO	0.89	0.875	0.861
Arteries	0.925	0.953	0.982
Bones	0.988	0.847	0.741
Digestive	0.8	0.889	1
Epithelium	1.0	1.0	1.0
Head	-	-	0
Heart	1.0	1.0	1.0
Lymphatic	1.0	1.0	1.0
Muscles	0.949	0.923	0.897
Nervous	0.9	0.947	1.0
Reproductive	0.973	0.9	0.837
Respiratory	1.0	0.8	0.667
Tissues	0.727	0.842	1.0
Veins	0.886	0.929	0.975

ping suggestions for each system (between 28% and 43% of the suggestions originally generated by the system). It also had the highest recall compared to the full reference alignment (between 30% and 43% of the recall for the systems), although not always with respect to the mappable parts only.

In general, DBSCAN led to higher convergence times compared to K-Means, since according to the DBSCAN algorithm each point may be visited at least once. Even though it does not require the a priori knowledge about the explored domain that is otherwise necessary to set up the K parameter, it implies higher computational time and this evidence is particularly critical when scaling to large, high-dimensional data. Another element that significantly stretched the overall computational time, due to its manual implementation, was the merging of different clusters once the topic they relate to had been identified. However, we found evidence during the experimentation that this step can contribute to the quality of results in a considerable way. In fact, when evaluating the alignment performance, the analyzed sets have to be of enough relevant size and this assumption has been validated by two main evidences:

After the cluster analysis, some very focused and small clusters can be identified and treating those clusters as stand-alone sets (which we did during the preliminary phases of our experimentation and modelling of the framework) leads to worse results than the ones obtained when forming the mappable parts grouped by topic.

The NCI-A is a richer and more nested data set than the AMA, therefore the first ontology allows to extract knowledge at an higher level of detail that in many specific cases cannot be found when mining the mouse ontology as well. Hence the sets are merged in order to make the mappable parts comparable in terms of both size and granularity between the two input data sources.

In other words, a merging phase has been included in the framework for both quality and accuracy purposes, although it is a time-consuming process. The automation of this phase may be desirable in order to minimize the computational issues but it should be designed accordingly in order to maintain, on the other hand, the benefits deriving from

Table 3. Results for all clustering algorithms combined with all variants of all systems.

	SAMBO		AML		LogMap-Bio		XMap	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
	0.89	0.861	0.956	0.932	0.888	0.906	0.94	0.85
KM-sil	0.931/0.929	0.322/0.897	0.97/0.97	0.346/0.963	0.924/0.924	0.339/0.945	0.938/0.938	0.322/0.897
KM-SSE-min	0.942/0.94	0.248/0.862	0.977/0.977	0.276/0.961	0.938/0.938	0.271/0.943	0.949/0.949	0.248/0.862
KM-SSE-max	0.92/0.908	0.152/0.85	0.981/0.981	0.17/0.949	0.947/0.947	0.166/0.923	0.953/0.953	0.147/0.82
DBSCAN	0.936/0.936	0.239/0.899	0.971/0.971	0.265/0.948	0.947/0.947	0.261/0.968	0.921/0.921	0.238/0.885
	MaasMtch		RSDLWB		AOT		AOTL_2014	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
	0.914	0.716	0.978	0.607	0.436	0.775		
KM-sil	0.966/0.966	0.302/0.842	0.978/0.978	0.267/0.744	0.604/0.604	0.309/0.862	0.787/0.787	0.024/0.125
KM-SSE-min	0.975/0.975	0.232/0.807	0.975/0.975	0.206/0.718	0.596/0.596	0.238/0.828	0.800/0.800	0.018/0.064
KM-SSE-max	0.963/0.963	0.139/0.776	0.909/0.909	0.106/0.588	0.524/0.524	0.142/0.79	0.950/0.950	0.013/0.070
DBSCAN	0.968/0.968	0.218/0.809	0.944/0.944	0.179/0.667	0.662/0.662	0.226/0.841	0.905/0.905	0.013/0.047
	LogMap		LogMap-C		LogMapLite			
	Precision	Recall	Precision	Recall	Precision	Recall		
KM-sil	0.945/0.945	0.318/0.886	0.984/0.984	0.281/0.783	0.955/0.955	0.296/0.824		
KM-SSE-min	0.956/0.956	0.245/0.851	0.988/0.988	0.208/0.725	0.964/0.964	0.228/0.794		
KM-SSE-max	0.961/0.961	0.146/0.816	0.995/0.995	0.125/0.699	0.985/0.985	0.129/0.717		
DBSCAN	0.958/0.958	0.238/0.885	0.985/0.985	0.219/0.814	0.970/0.970	0.212/0.787		

the integration of the analyst's knowledge within the process. In fact, the application of a manual merging step showed the advantage of allowing non-trivial associations when identifying different concepts as belonging to the same topic. For instance, one of our main prerequisites in this sense was relying on some a priori anatomical knowledge while forming the mappable parts.

6 Conclusion

We have introduced approaches for reducing the search space in OA using clustering and topic identification. By partitioning the ontologies, we generated mappable parts for which the computation and validation time is much reduced while maintaining good quality of the alignment for the parts. We do need other techniques for aligning the other parts of the ontologies. However, we can use the proposed techniques to obtain an initial alignment, which could be used as a basis for PA-based [11] and session-based [10] approaches to complete the alignment. An issue for future work is to address a current limitation, i.e., we want to investigate techniques to automate the topic identification step. Further, other clustering techniques should be investigated.

Acknowledgments We thank SeRC, CUGS and the EU FP7 project VALCRI (FP7-IP-608142) for financial support.

Table 4. Number and percentage (truncated) of comparisons, and for all variants of each system number and percentage (truncated) of the mapping suggestions.

	Comparisons	SAMBO	AML	LogMap-Bio	XMap	MaasMtch	RSDLWB	AOT	AOTL_2014	LogMap	LogMap-C	LogMapLite
full	9,026,626	1466	1478	1547	1370	1187	941	2698	167	1398	1061	1148
KM-sil	264,536 (2.9%)	524 (35%)	540 (36%)	556 (35%)	520 (37%)	474 (39%)	414 (43%)	777 (28%)	47 (28%)	510 (36%)	433 (40%)	469 (40%)
KM-SSE-min	206,834 (2.2%)	399 (27%)	429 (29%)	438 (28%)	396 (28%)	361 (30%)	321 (34%)	606 (22%)	35 (20%)	388 (27%)	320 (30%)	359 (31%)
KM-SSE-max	163,963 (1.8%)	250 (17%)	263 (17%)	265 (17%)	234 (17%)	219 (18%)	176 (18%)	410 (15%)	20 (11%)	231 (16%)	191 (18%)	198 (17%)
DBSCAN	258,308 (2.8%)	388 (26%)	414 (28%)	417 (26%)	392 (28%)	341 (28%)	288 (30%)	518 (19%)	21 (12%)	377 (26%)	337 (31%)	331 (28%)

References

1. H.-H. Do and E. Rahm. Matching large schemas: approaches and evaluation. *Information Systems*, 32:857–885, 2007.
2. Z. Dragisic et al. Results of the ontology alignment evaluation initiative 2014. In *International Workshop on Ontology Matching*, pages 61–104, 2014.
3. M. Ester et al. A density-based algorithm for discovering clusters an large spatial databases with noise. In *2nd KDD*, pages 226–231, 1996.
4. J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer, 2007.
5. J. Euzenat et al. Ontology alignment evaluation initiative: Six years of experience. *J Data Semantics*, XV:158–192, 2011.
6. A. Gross et al. GOMMA results for OAEI 2012. In *7th Int W on Ontology Matching*, pages 133–140, 2012.
7. S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5):345–366, 2000.
8. W. Hu, Y. Zhao, and Y. Qu. Partition-based block matching of large class hierarchies. In *1st Asian Semantic Web Conference*, pages 72–83, 2006.
9. E. Jimenez-Ruiz et al. Large-scale interactive ontology matching: Algorithms and implementation. In *20th ECAI*, pages 444–449, 2012.
10. P. Lambrix and R. Kaliyaperumal. A session-based approach for aligning large ontologies. In *10th ESWC*, pages 46–60, 2013.
11. P. Lambrix and Q. Liu. Using partial reference alignments to align ontologies. In *6th ESWC*, pages 188–202, 2009.
12. P. Lambrix and H. Tan. SAMBO - a system for aligning and merging biomedical ontologies. *J Web Semantics*, 4(3):196–206, 2006.
13. P. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comp and Appl Math*, 20:53–65, 1987.
14. M. H. Seddiqui and M. Aono. An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size. *J Web Semantics*, 7(4):344–356, 2009.
15. P. Shvaiko and J. Euzenat. Ontology matching: state of the art and future challenges. *IEEE TKDE*, 25(1):158–176, 2013.
16. P. Wang, Y. Zhou, and B. Xu. Matching large ontologies based on reduction anchors. In *22nd IJCAI*, pages 2243–2348, 2011.