

Effective Ontology Matching in High-Performance Computing Environments

Axel Tenschert¹, Alexey Cheptsov¹

¹ HLRS – High-Performance Computing Center Stuttgart, University of Stuttgart,
Nobelstraße 19,
70569 Stuttgart, Germany
[tenschert, cheptsov}@hlrs.de](mailto:{tenschert, cheptsov}@hlrs.de)

Abstract. Extending complex information structures by means of ontology matching is of high interest for a number of tasks solved in the semantic web. The main motivation behind this work is that the procedure of ontology matching requires a robust and scalable solution that ensures the maximal efficiency of matching operations. That is especially important when thinking of matching large scale data among several ontologies, where the performance and scalability of performing the matching algorithms is settled to the point. In this paper, we propose an approach for distributed ontology matching, improving the matching's efficiency and scalability due to the distribution and parallelization of implemented algorithms. This enables applications performing ontology matching to get benefit of running in high-performance computing environments and ensures that the full potential of computing resources is enabled for the matching process.

Keywords: Ontology Matching, Semantic Content, High Performance Computing, Parallelization, Distribution, Grid Computing

1 Introduction

The progress of information and communication technologies has made available a huge amount of disparate information. The number of resources collecting the information is growing, accordingly, and therefore, the problem of managing heterogeneity among those resources is increasing. As a consequence, various solutions have been proposed to facilitate dealing with this situation, and specifically, for automating integration of distributed information sources. Among others, ontology matching from the field of semantic technologies has attracted significant attention.

An ontology typically provides a semantic vocabulary that describes a domain of interest and a specification of the meaning of terms used in the vocabulary. Depending on the precision of this specification, the notion of ontology encompasses several data and conceptual models, for example, sets of terms, classifications, database schemas, or fully axiomatized theories. However, when several competing ontologies are in use in different applications, most often they cannot interoperate as is, though the fact of using ontologies rises heterogeneity problems to a higher level.

Ontology matching is a solution to the semantic heterogeneity problem. It finds correspondences between semantically related entities of ontologies. These correspondences can be used for various tasks, such as ontology merging, query answering, data translation, etc. Thus, matching ontologies enables the knowledge and data expressed in the matched ontologies to interoperate.

The Goal of this work is to improve existing ontology matching methods by using grid infrastructures within a cluster. For this, ontologies are portioned into several parts. Each part of the portioned ontology is matched with the concepts of another ontology in parallel. However, this approach requires parallelization techniques which are applicable in a grid environment within a cluster. Through the distributed execution of a grid infrastructure it is possible to distribute the matching process on several resources. Furthermore, the cluster provides the matching procedure with the required compute resources.

When thinking about learning and teaching in higher education provided by the usage of semantic contents which are closely related to semantic web technologies, we also have to think about the utilization of ontologies. Through this the matching of several ontologies in an effective way is still a challenge. Therefore, this work presents an approach how to solve ontology matching for large scale datasets by adapting parallelization and distribution techniques.

2 Use Case

For this work one objective is to provide a user with one priority ontology which includes the knowledge structures of a given set of ontologies. Hereby, the user is enabled to use the knowledge of several ontologies by using only one priority ontology.

When thinking of merging ontologies to one priority ontology the research field of bioinformatics is of interest because of the need to investigate large scale data within a high number of ontologies. There is already a large number of different ontologies

available. Through this, the aim of this work is to provide a doctor or a scientist in the field of medicine to obtain required medical information quick in a user friendly way by the usage of one priority ontology. The benefit for such an ontology is the possibility to receive required information from the research field of bioinformatics very fast by considering only one data source, the extended priority ontology. Through this, learning and teaching in the field of semantic web are provided by this work. Whether the user is a doctor or a scientist, the presented approach supports receiving required datasets in order to extend already available knowledge structures.

A matching strategy for ontologies will be useful in order to manage large scale ontologies. One priority ontology is used instead of a large set. For example, if there is a user who is in the need of receiving knowledge about human diseases there is the possibility to match the concepts of these ontologies and merge concepts which are similar. However, when matching these ontologies about human diseases one by one and merge them after the matching it is important to define which ontologies should be merged first. In case that three ontologies about human diseases, called Onto1, Onto2 and Onto3 and a user who defines that Onto1 is the priority ontology, are considered, the next step is to match Onto2 or Onto3 with the priority ontology.

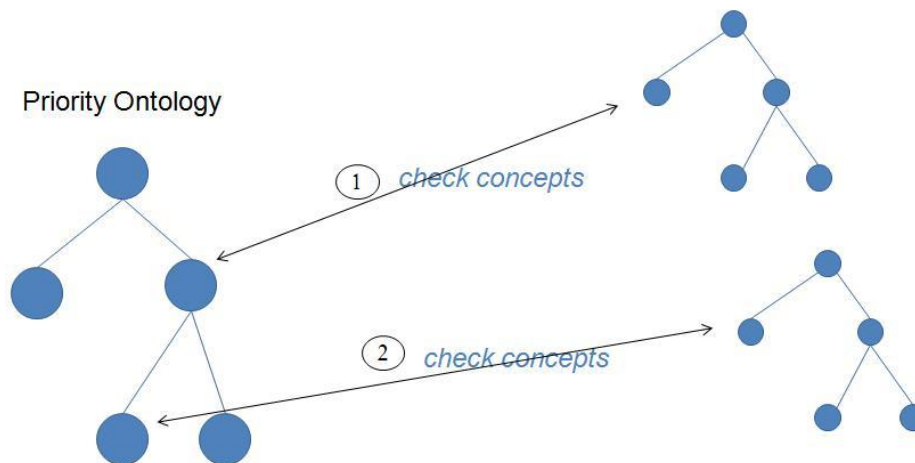


Figure 1: matching concepts of ontologies

As presented in figure 1 the concepts of the ontologies are matched with the aim to merge them. In the following sections we will take a closer look to the matching strategies. For this, we will clarify in which way distribution and parallelization of matching ontologies is executed.

3 Approaches for Ontology Matching

Current approaches for ontology matching are used in order to merge ontologies together. Therefore the selected ontologies are matched by an adequate algorithm in order to ensure a proper merging. However, these approaches require a high amount

of compute resources in order to meet the requirements of the matching and merging methods.

There are several issues which have to be solved for ensuring a scalable matching solution.

1. How to match the ontologies / which approach is most beneficial
2. How to match in a scalable and robust manner
3. Which ontologies and concepts of the ontologies have to be matched first / is there a priority setting
4. Ontologies are selected from a repository / which repositories are available and beneficial to meet the specific requirements of a user

In order to solve the mentioned issues the matching strategy is important. Current approaches consider a division of selected ontologies with the aim to execute the matching algorithms independently from other parts of the ontology. In detail this means that one ontology is divided into several parts which contain a certain set of concepts of the divided ontology. Each part is matched with another selected ontology. This method is increasing the performance of ontology matching. Currently Falcon-AO, an automatic ontology matching system, is available which is part of the Falcon¹ infrastructure. Falcon-AO supports the division of ontologies into several parts by the PBM² with the aim to match selected ontologies together. However, it is still a challenge to provide the matching with the required compute resources.

For the selection of ontologies from the field of bioinformatics there are several ontologies repositories available in the web, some of them are listed below.

- BioPortal: <http://bioportal.bioontology.org/ontologies>
- Clinical Bioinformatics Ontology:
<https://www.clinbioinformatics.org/cbopublic/>
- The University of Manchester:
<http://www.cs.man.ac.uk/~stevensr/menupages/ontologies.php>

When searching for adequate ontologies in the research field of bioinformatics it is not a problem to find ontologies but to find ontologies which are most beneficial for a specific task and to keep the effort for matching them low in a temporal solvable manner. Therefore, the idea is to match a selection of ontologies with the aim to merge them together to one extended priority ontology. This task is provided by the usage of a grid infrastructure in a cluster. The cluster provides the required compute resources and the grid infrastructure allows a distributed execution of the ontology matching.

4 Distributed Ontology Matching

There are several ontology matching strategies and merging tools available. Nevertheless, the complexity of matching ontologies entails the problem of matching

¹ Falcon infrastructure: <http://iws.seu.edu.cn/projects/matching/>

² PBM = Partition-based matcher

them in a scalable way. In order to solve this problem distribution and parallelization techniques are used to speed up the matching by executing it at same time in parallel. Furthermore, the required compute resources are provided by executing the ontology matching within a cluster environment.

For this it is important to consider existing approaches for ontology marching in parallel on distributed resources with the aim to adapt those techniques and improve them for this work. Hence, we will consider the LarKC project³ in which new techniques for processing large datasets in the research field of the semantic web are developed for the usage of concrete use cases. Within the European founded LarKC project ontologies are used as well and new techniques for the usage of large scale data sets are developed and used for real time applications. For this parallelization techniques are used to run several processes at same time. Furthermore, within the LarKC project parallelization techniques are considered to execute processes in a cluster environment.

In order to meet the requirements of matching ontologies in parallel one ontology from the given set becomes the priority ontology. This priority ontology is extended by matching the concepts of the priority ontology with the concepts of the other selected ontologies. However, the clue is to execute the matching procedure parallel at same time for many concepts within a cluster environment to provide the required compute resources. For this, methods for executing processes in parallel in a cluster are of interest.

When thinking of matching concepts of the priority ontology at same time, the first step is to divide the priority ontology into several parts. The concepts of each part are matched in parallel with a selected ontology from the set.

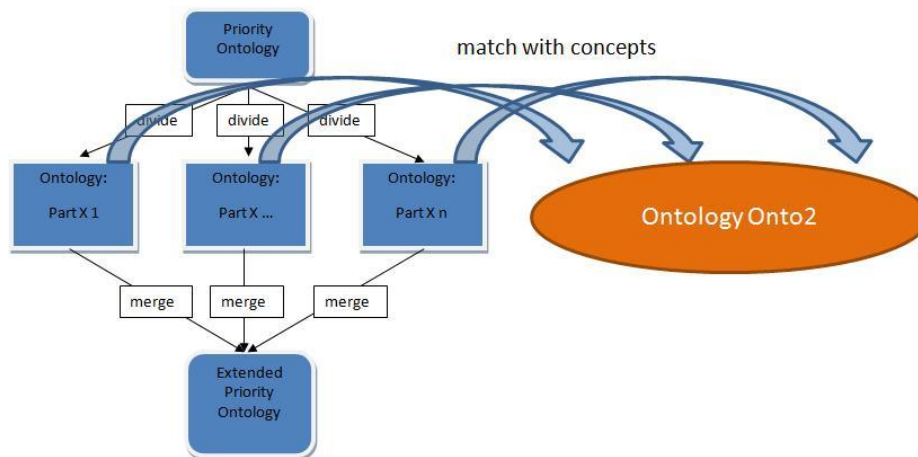


Figure 2: division of the priority ontology

As presented in figure 2 the idea of dividing the priority ontology and match the concepts of each part in parallel with the concepts of a selected ontology from the available set of ontologies will be main part of the idea to increase the scalability of

³ LarKC (abbr. The Large Knowledge Collider): <http://www.larkc.eu/>

ontology matching. After the matching is executed the parts of the priority ontology are merged together again and the new extended priority ontology is generated. However, it is possible to select each portioned part of the priority ontology with one selected ontology (e.g. in figure 2: Onto 2) but it is possible to match each part of the priority ontology with different ontologies from the available set.

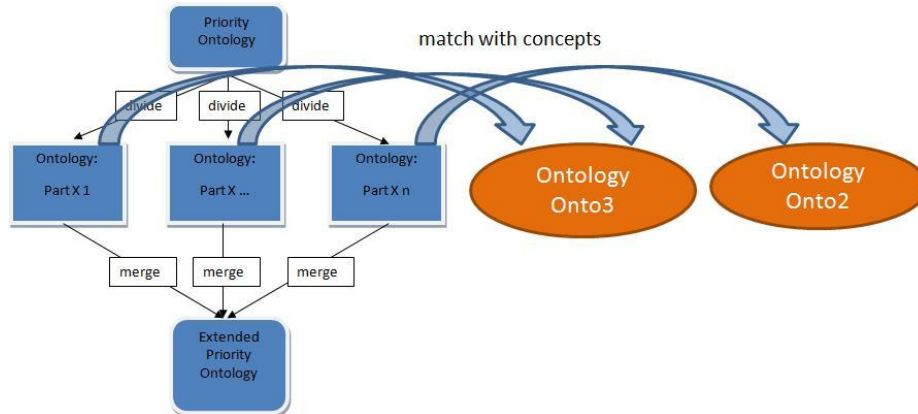


Figure 3: matching with several ontologies

Figure 3 presents how to match the concepts of the priority ontology with several ontologies from a known set in parallel. However, it is still an open issue how to ensure a scalable and robust matching. For this, each matching procedure is executed in a cluster. When executing a job in the cluster the number of required nodes and which job is executed on which node is defined. This proceeding ensures a user friendly way which meets the requirements of the user in terms of compute resources. In detail this means, the allocated compute resources are set individually by the user considering his specific requirements. The user selects a number of nodes. After this, the jobs are executed on the selected nodes in the cluster.

Table 1. Parallel ontology matching within a cluster.

Matching Part	Node
Ontology Part X 1	Node A
Ontology Part X ...	Node B
Ontology Part X n	Node C

However, the distribution of the matching of the partitioned parts of the priority ontology presented in table 1 is just an example. The distribution of the jobs depends on the size of data and size of the jobs. Therefore, it is possible to execute several jobs on one node. To summarize the mentioned issues for an effective ontology matching within a high-performance computing environment the described procedure is listed below.

1. Selection of ontologies from known repositories
2. Defining one priority ontology from the selection

3. Division of priority ontology into several parts
4. Defining required compute resources in a cluster
5. Matching of ontology parts in parallel in the cluster
6. Merging parts to extended priority ontology

5 Conclusions

The presented approach for matching ontologies in a high-performance computing environment is an effective method to solve the challenge of matching in a scalable, robust and timesaving way. Though this, it is of high interest to topics of the semantic web such as learning and teaching for which semantic content from ontologies is required. However, the presented work is an overview about ideas and methods which are analysed to solve the challenge of effective ontology matching. Furthermore, within the LarKC project parallelization and distribution techniques for executing semantic data structures are analysed and developed. These techniques are considered for this work as well. Through this work considers new technologies from the research field of the semantic web.

Parallelization and Distribution techniques are effective methods for ontology matching when thinking about large scale ontologies. Hence, these are valuable techniques for the semantic web applications related to learning and teaching.

Acknowledgments. This work has been supported by the LarKC project (<http://www.larkc.eu/>) and has been partly funded by the European Commission's IST activity of the 7th Framework Program under contract number 215535. This work expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this work.

References

1. The LarKC Project, Website <http://www.larkc.eu/>
2. Euzenat, J., Shvaiko, P., Ontology Matching. Springer. Berlin; Heidelberg (2007)
3. P. Shvaiko, J. Euzenat: Ten Challenges for Ontology Matching In Proceedings of ODBASE, 2008.
4. Oberle, D., Semantic Management of Middleware. Springer Science+Business Media, Inc. NewYork (2006)
5. Pellegrini, T., Blumauer, A., Semantic Web. Springer-Verlag. Berlin; Heidelberg [u.a.] (2006)
6. The National Center for Biomedical Ontology: <http://bioportal.bioontology.org/>
7. Clinical Bioinformatics Ontology: <https://www.clinbioinformatics.org/cbopublic/>
8. The University of Manchester: <http://www.cs.man.ac.uk/~stevensr/menupages/ontologies.php>