Effective Ontology Matching in High-Performance Computing Environments

Axel Tenschert[1], Alexey Cheptsov[1]

[1] HLRS – High-Performance Computing Center Stuttgart, University of Stuttgart,
Nobelstraße 19,
70569 Stuttgart, Germany
{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.

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. Thus, this work presents an approach for matching ontologies in an effective and robust way.

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 large scale ontologies from the field of bioinformatics. The benefit for such an ontology is the possibility to receive required information (e.g. medical datasets) 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. The presented approach supports receiving required datasets in order to extend already available knowledge structures.

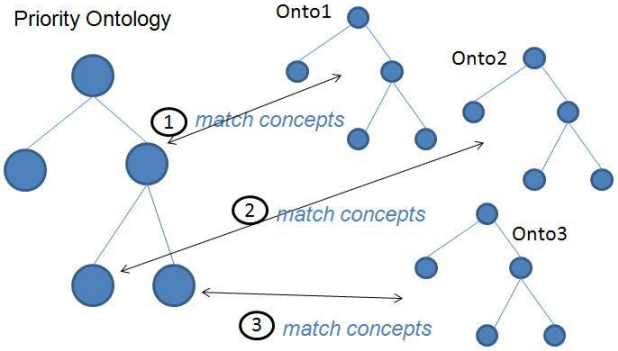

Figure 1: matching concepts of ontologies

One priority ontology is used instead of a large set (see figure 1). For example, a scientist or a doctor needs to receive knowledge about human diseases. Therefore, the possibility to match the concepts of a selection of ontologies and merge concepts which are similar is beneficial. However, it is important to define a sequence for matching the set of ontologies. In the case that three ontologies about human diseases, called Onto1, Onto2 and Onto3 are considered the next step is to match Onto2 or Onto3 with the priority ontology. However, the priority ontology is selected arbitrary by the user.

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 and how to improve a sequential matching.

3 Approaches for Ontology Matching

Currently 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 computing resources in order to meet the requirements of the matching and merging methods. Hence, there are several issues which have to be solved for ensuring a scalable matching solution, e.g. identification of the most beneficial matching approach, ensure scalability and robustness, sequence for matching the ontologies, identification of beneficial ontology repositories.

At present approaches consider a division of selected ontologies with the aim to execute the matching algorithms independently from other parts of the ontology. At this one ontology is divided into several parts. Falcon-AO, an automatic ontology matching system and part of the Falcon[1] infrastructure, is related to this issue. Falcon-AO supports the division of ontologies into several parts by the PBM[2] with the aim to match selected ontologies together. However, it is still a challenge to provide the matching with the required computing resources.

For the selection of ontologies from the field of bioinformatics there are several ontology 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 ontolgies but to find ontolgies 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 computing resources and the grid infrastructure allows a distributed execution of the ontology matching.

4 Distributed Ontology Matching

The complexity of matching ontologies entails the problem of matching them in a scalable way. For this, distribution and parallelization techniques are used to increase scalability by executing it at same time in parallel. Furthermore, the required computing resources are provided by executing the ontology matching in a cluster environment.

---

[1] Falcon infrastructure: http://iws.seu.edu.cn/projects/matching/

[2] PBM = Partition-based matcher

For this it is important to consider existing approaches for ontology matching in parallel on distributed resources with the aim to adapt those techniques and improve them for this work. Hence, we will consider the LarKC project[3] 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.

The new approach is to set one ontology from the given set as the priority ontology which 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 in a cluster environment to provide the required computing 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 the concepts of several selected ontologies. After the matching is executed the parts of the priority ontology are merged together again and the new extended priority ontology is generated (see Figure 2).

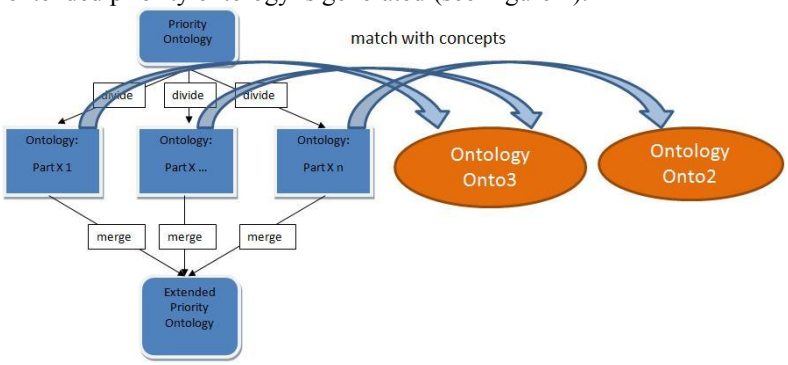

Figure 2: matching with several ontologies

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. The allocated computing resources are set individually by the user considering his specific requirements. Further on, the user selects a number of nodes so that the jobs are executed on the selected nodes in the cluster. However, 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.

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

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 for the semantic web and for learning and teaching which requires semantic content. 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.

Thus, 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