RiMOM2: A Flexible Ontology Matching Framework

Zhichun Wang, Xiao Zhang, Lei Hou, Juanzi Li

Department of Computer Science and Technology, Tsinghua University

Haidian District, Beijing, China, 10084

{zcwang, zhangxiao, houlei, ljz}@keg.cs.tsinghua.edu.cn

ABSTRACT

With the development of the Linking Open Data (LOD) project, large amount of semantic datasets have been published on the Web. Due to the open and distributed nature of the Web, the published data may be heterogeneous both in the schema level and instance level. Matching the entities of different datasets is very important for integrating information from different data sources. Recently, much work has been done in the domain of ontology matching to resolve the schema heterogeneity problem. However, there is lack of a unified framework for matching entities both in the schema level and instance level. This paper presents a tool RiMOM2 which provides a framework for both ontology schema matching and instance matching.

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic networks

General Terms

Algorithms, Experimentation, Performance

Keywords

Ontology Matching, Instance Matching, Heterogeneity

1. INTRODUCTION

In recent years the Web has evolved from a global information space of linked documents to one where data is linked as well. A large amount of linked data has been published on the Web leading to the creation of the Web of Data, a global data space containing billions of assertions[1]. Although there are several rules guiding the publication of linked data, two important issues are still needed further investigation. The first issue is the schema heterogeneity problem. Although the use of common schemas/ontologies is encouraged in order to make it easier for client applications to process linked data, existing datasets often employ their own schemas. When schemas in the same domain are defined by different organizations, they will be quite different from each other. The heterogeneity in schemas hinders the data sharing and data integration. Second, the number of established RDF links connecting data is much less than the number of real links between data. There is lack of tools that can build semantic correspondences between entities in both schema and instance level of different datasets. Therefore, LOD provides a new environment for investigating the ontology matching techniques, and also raises some challenges for ontology matching.

This paper presents a tool RiMOM2, which is a flexible framework for both ontology schema matching and instance matching. RiMOM2 is a new version of our previous system RiMOM[2]. RiMOM can automatically select and combine several different matchers, but the matchers and aggregation methods are fixed and user can not define their own matchers or aggregation function. Therefore we extend our previous work and proposed RiMOM2, a more flexible framework for ontology matching and instance matching. In RiMOM2, all of the original methods are organized as functional components that provide users more convenience to add their own strategies and define their own process through a user interface. Meanwhile, for the new users, we provide a default process by using the original method.

2. The ARCHITECTURE OF RiMOM2

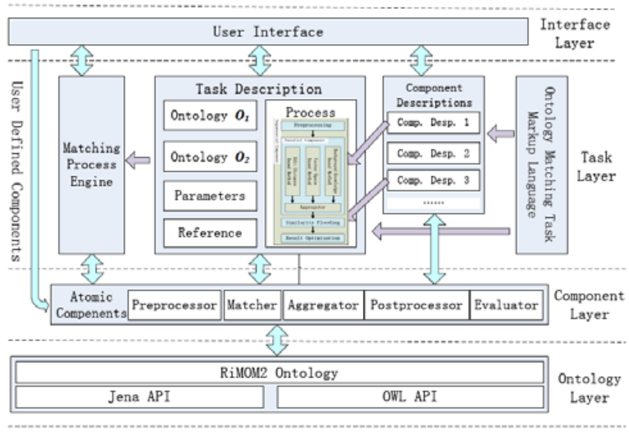

Figure 1 RiMOM2 Framework

RiMOM2 is a flexible framework for ontology matching which is in four-layer architecture, as shown in Figure 1. The bottom Ontology Layer provides a set of uniform operations for ontology matching tasks with different underlying ontology APIs. We analyze the matching process and abstract five kinds of atomic components from the process: Preprocessor, Matcher, Aggregator, Postprocessor and Evaluator. The Matching Process can be considered as a composition of selected components. RiMOM2 has different implementations for each kind component, and registers them in the component store with their descriptions. The description file defines the input, output and parameters settings for the atomic component and describes their functionality and usage. Then RiMOM2 employs the Ontology Matching Task Markup Language to integrate the atomic components into a complete matching process. The Ontology Matching Task Markup Language is a scheme for describing the matching process stored in Task Description File. In the scheme we define two upper-level components: Sequence and Parallel. The Sequence component is composed of a list of Sequence, Parallel and atomic components. The Parallel is composed of a set of components in parallel and an Aggregator. The whole matching process is a Sequence component. Consequently the schema allows users to describe very complicated matching process. The

Engine loads the information from the Task Description File and executes the matching task. The Interface Layer deals with all user-involved operations, including user input, customization for matching process, results display so on. With such design, RiMOM2 achieves two levels of flexibility for ontology matching tasks. Firstly, users can design their own matching process according to the characteristics of ontologies. Secondly, advanced users can develop new atomic components for functionality they need and utilize it in their matching process. RiMOM2 also provides a default matching process for new users.

3. CUSTOMIZING A MATCHING APPROACH IN RiMOM2

Here we customize a unified approach in RiMOM2 for both ontology schema and instance matching; the selected matchers, aggregator and postprocessor in this approach are introduced in the following subsections. We also evaluate the approach on OAEI tests, the results show that RiMOM2 is among the top systems for ontology matching.

3.1 Matchers

Name-based matcher: This matcher computes the edit-distance between the labels of two entities.

Description-based matcher: This matcher compares the descriptions of two entities.

Instance-based matcher: This matcher is only used for schema matching by making use of the information of instances.

Attribute-based matcher: This matcher is only for instance matching. Here all the data type property values of a instance are combined as its composite description. The matcher also uses the same method in Description-based matcher to compute the similarity of two entities' composite descriptions.

3.2 Aggregator

We define a *Voting-based Aggregator*. This aggregator treats each matcher as an independent decision maker, and obtains a set of predicted matches for each matcher. If a matcher predicts a candidate match as a true match, it is voting for that match. In order to combine the matchers' results, the aggregator calculates the union of different sets of predictions from multiple matchers, and records the votes for each match. In most cases, one entity in a dataset can only match with one entity in the other dataset, which is known as 1-1 matching rule. In the set of combined predictions, there may be some matches conflict with the 1-1 matching rule. We resolve the conflicts by keeping the match with the most votes and discarding other matches conflict with it.

3.3 Postprocessor

In both schema and instance levels of an ontology, there are rich structural information. The structural information is usually used to propagate the similarity between entities, such as in the similarity flooding algorithm. Here we define a *Constrained Similarity Propagation* method, which only propagates the similarities of matches generated by the voting-based aggregator. This new method can effectively avoid propagating unreliable similarities and ensure better results.

3.4 Evaluation

We evaluate the customized approach of RiMOM2 on the OAEI'2011 Benchmark and Person-Restaurant datasets. Benchmark is a task of ontology schema matching and Person-Restaurant is a task of ontology instance matching. Figure 2 and Figure 3 show the results on these two datasets respectively. We also compare the performance of our approach with other top systems participated in OAEI'2011 in the figures. According the F-measure, our approach ranks the second and first in Benchmark and Person-Restaurant datasets respectively.

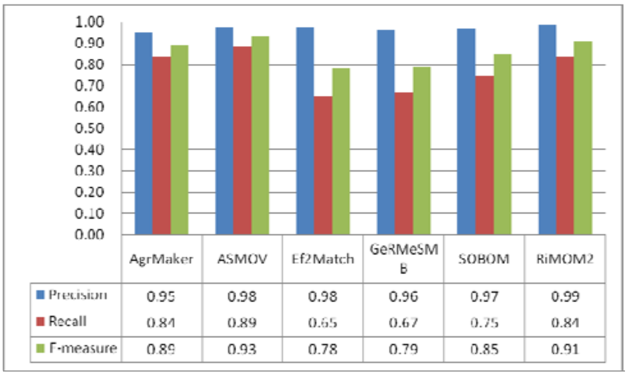

Figure 2 Results of OAEI'10 Benchmark Track

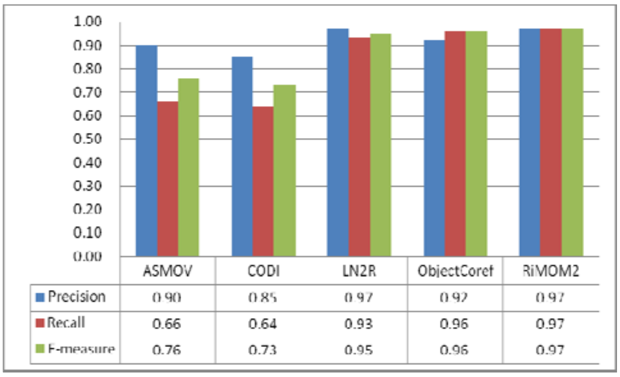

Figure 3 Results of OAEI'10 Person-Restaurant Track

4. ACKNOWLEDGMENTS

The work is supported by the National Natural Science Foundation of China (No. 60973102), the National Basic Research Program of China (973 Program) (No. 2007CB310803), and the State Key Program of National Natural Science of China (Grant No. 61035004), it is also supported by IBM SUR joint project.

5. REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked data -the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.

[2] J. Li, J. Tang, Y. Li, and Q. Luo. Rimom: A dynamic multistrategy ontology alignment framework. *IEEE Trans. on Knowl. and Data Eng.*, 21(8):1218–1232, 2009