A New Approach for Combining the Similarity Values in Ontology Alignment

Moussa Benaissa, Abderrahmane Khiat

A New Approach for Combining the Similarity Values in Ontology Alignment

Moussa Benaissa and Abderrahmane Khiat

LITIO Laboratory, University of Oran1 Ahmed Ben Bella,
B.P 1524 El M'Naouar 31000, Oran, Algeria
moussabenaissa@yahoo.fr
abderrahmane_khiat@yahoo.com

Abstract. Ontology Alignment is the process of identifying semantic correspondences between their entities. It is proposed to enable semantic interoperability between various knowledge sources that are distributed and heterogeneous. Most existing ontology alignment systems are based on the calculation of similarities and often proceed by their combination. The work presented in this paper consists of an approach denoted PBW (Precision Based Weighting) which estimates the weights to assign to matchers for aggregation. This approach proposes to measure the confidence accorded to a matcher by estimating its precision. The experimental study that we have carried out has been conducted on the Conference[1] track of the evaluation campaign OAEI[2] 2012. We have compared our approach with two methods considered as the most performed in recent years, namely those based on the concepts harmony and local confidence trust respectively. The results show the good performance of our approach. Indeed, it is better in terms of precision, than existing methods with which it has been compared.

Keyword: Ontologies, Ontology Alignment, ontology matching, Semantic Correspondences, Similarity, Aggregation of the Similarities, Combination of the Similarities.

1 Introduction

The Semantic Web Community, defined as a futuristic extension of the current web, has adopted ontologies as the cornerstone for its achieving in order to overcome the crucial problem of semantic heterogeneity that is inherent to its distributed and open nature. However, these ontologies are themselves heterogeneous. This heterogeneity may occur at syntactic, terminological, conceptual or semiotic levels [5].

[1] http://oaei.ontologymatching.org/2012/conference.

[2] OAEI (Ontology Alignment Evaluation Initiative) organizes evaluation campaigns aiming at evaluating ontology matching technologies. http://oaei.ontologymatching.org/

Ontology alignment, defined as the process of identification of semantic correspondences between entities of different ontologies to be aligned [5], is proposed as a solution to the problem of semantic heterogeneity by enabling the semantic interoperability between various sources of information.

We globally distinguish two approaches to identify the alignment between ontologies: reasoning-based approaches and those based on the calculation of similarities [12].

Most of the existing ontology alignment systems are based on the calculation of similarities between entities to align. In this category, we distinguish two types of systems: (1) systems which implement one single technique and (2) systems which combine several techniques, in order to estimate the similarity between two entities. The latter systems have become more frequent due to their flexibility and their easy extension [7]. Moreover, with the increasing complexity of ontologies on the Web (number and volume), the alignment cannot be performed reasonably in a purely manually way. Therefore it is imperative to develop automatic or at least semi-automatic systems to identify the alignment [11]. This situation is dictated by the lack of human expert especially in dynamic systems and by the concern to accelerate the alignment process [1].

Precisely, we propose in this paper an ontology alignment approach based on the calculation of similarities and which fits into the category of methods that combine several matchers. It is a statistical approach based on two heuristics to aggregate similarity values calculated by different matchers. The first estimates the candidate final alignment from the alignments identified by matchers, considering their intersection. The second provides an estimate of the weight to be assigned to the matchers with a view of their combination using a weighted summation strategy.

The rest of the paper is organized as follows. In the Section 2, we present some preliminary notions on ontology alignment in order to facilitate the reading of the paper content. The Section 3 contains the description of some related work to our approach. In the Section 4, we present an example in order to illustrate our approach. The Section 5 is dedicated to the presentation of the proposed approach. The Section 6 contains the experimental results obtained during the evaluation of our approach. Finally we give a conclusion and some future perspectives.

2 Preliminaries

In this section we present some preliminary notions of ontology alignment in order to facilitate the reading of the paper content. We outline the notions of ontology, similarity calculation techniques and alignment, respectively. We refer the reader, for more details, to the following references [5] [4].

2.1 Notion of Ontology

Definition: Ontology is a six tuple [2]: $O = <C, R, I, H^C, H^R, X>$ where:

- C: set of concepts.
- R: set of relations.
- I: set of instances of C and R.
- H^C: denotes a partial order relation on C, called hierarchy or taxonomy of concepts. It associates to each concept its super or sub-concepts.
- H^R: denotes a partial order relation on R, called hierarchy or taxonomy of relations. It associates to each relation its super or sub-relations.
- X: set of axioms.

2.2 Techniques of the Similarities Calculation

There are basically five types of methods to calculate similarities [1]:

1. *Terminological Methods*. These methods are based on string matching and can be applied to the names, labels and descriptions of the entities. We cite as an example of matcher of this category: the edit distance.
2. *Linguistic Methods*. These methods are based on external resources as dictionary and thesaurus in order to calculate the similarities between the names, labels and descriptions of the entities. We cite as an example of a matcher of this category: similarity based on WordNet (Wu-Palmer).
3. *Structure-based Methods*. These methods exploit the internal structure (domain, range, properties and cardinality, etc.) and the external structure (hierarchy and the relation-ship between other entities) of the entities in order to calculate their similarities. We cite as an example of a matcher of this category: Resnik similarity.
4. *Semantic-based Methods*. These methods are essentially deductive and inferential and are based on formal semantic of generic or specific domains. We cite as an example of a matcher of this category: SAT solvers.
5. *Instance-based Methods*. These methods exploit the instances associated to the concepts (extensions) to calculate the similarities between them. We cite as an example of a matcher of this category: Jaccard similarity.

2.3 Notion of Ontology Alignment

The alignment of two ontologies is the process of identification of semantic correspondences between their entities. In this section, we briefly introduce the basic necessary concepts on the alignment in order to facilitate the reading of the paper content.

4.3.1. Notion of Correspondence

Let O and O' two ontologies. A Correspondence M between O and O' is quintuple < Id, e, é, r, n > where:

- Id: is a unique identifier of the correspondence M;
- e and e‘ are the entities of O and O' respectively (concepts, relations or instances);
- r: is the semantic relation between e and e‘ (equivalence (≡), more specific (⊆),more general (⊇), disjunction (⊥));
- n: is a measure of confidence, typically a value within [0, 1].

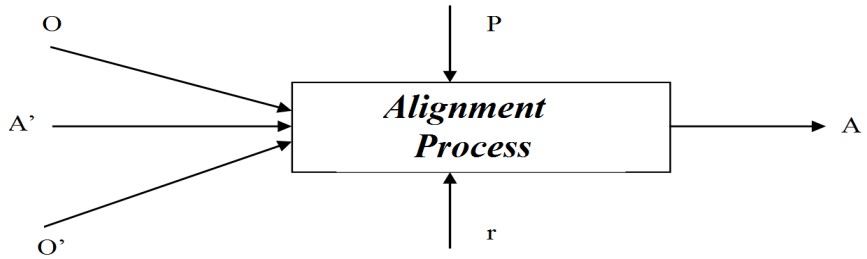

Fig. 1. Alignment Process

4.3.2. Notion of Alignment

The alignment can be defined as a set of correspondences. The alignment process (Fig. 1) receives as input two ontologies O and O' and produces as output an alignment A between entities of O and O'. Other elements complete this definition, namely:

- An initial alignment A' to be completed or refined by the process.
- The external resources r such as a thesaurus or a dictionary.
- The parameters P such as thresholds or weights.

The alignment process consists generally of the following steps:

1. *Analysis*: This step consists of extracting both the entities (concepts, relations, instances) of the two ontologies O and O ' and their characteristics which will be used to identify the alignment.
2. *Calculation of Similarities*: this step consists to execute the different matchers in order to calculate the similarities between entities to align.
3. *Similarity Values Aggregation*: This step consists to combine the similarity values calculated by the matchers in the previous step, into one value.
4. *Selection*: This step consists of applying a strategy, for example a threshold strategy in order to filter the alignment defined in the previous step. Other optimization techniques can also be applied at this level to optimize the extraction of the final alignment.

5. *Improvement of the Alignment*: descriptive logic techniques can be applied at this level to improve the final alignment by diagnosing and repairing any inconsistencies identified in the final alignment.

3 Related Works

The aggregation of similarity values calculated by different matchers consists to combine them into one single value. There are basically three types of approaches to achieve this aggregation: the **weighting**, the **vote** and the **argumentation** [5]. In the vote strategy, the matchers are considered as independent sources of information and the decision to include a correspondence in the alignment is taken on the basis of a simple majority vote by the matchers for this correspondence. The argument strategy allows negotiating an alignment by exchanging arguments between agents. In the weighting strategy several techniques are proposed to combine the similarity values.

In [3], the authors quote the following strategies to combine similarity values calculated by different matchers: (1) Max: this strategy selects the maximum similarity value among the values calculated by different matchers); (2) Min: this strategy selects the minimum similarity value among the values calculated by different matchers); (3) Average (this strategy calculates the average value of the similarities calculated by different matchers); and (4) Weighting (this strategy calculates the weighted sum of the similarities calculated by different matchers). The latter, which is more frequent in ontology alignment systems [7], requires an estimate of the weights that reflect the importance of each matcher. In some systems this weights approximation is done manually by a human expert. This approach is difficult to implement given the enormous number of possible configurations [11] and has the major drawback to run correctly on a specific alignment task and not on another. It is therefore suitable that the weights estimation be specific to the current alignment task [8].

Several studies have addressed the problem of the weights estimation of different matchers. In [14], the authors propose an approach based on information theory and estimate the weight of each matcher based on the calculation of entropy (uncertainty of information) from the similarity values calculated by this matcher.

The works described in [9] and [13] present an approach based on genetic algorithms to give an estimate of weights assigned to different strategies used.

In [8] the authors propose the harmony concept for weighting the different matchers. The harmony **h** of a similarity matrix **sim** of **n** rows and **m** columns is defined by: « *the number of pairs of entities* (e_i, e'_j) *for which the similarity* ***sim*** (e_i, e'_j) *is the maximum at the same time on the row i and column j, divided by the maximum number of concepts of ontologies to align O and O'* ". This value **h** is assigned as weight to the matcher associated to the matrix **sim**. In [2] the authors propose a local confidence measure for a pair of entities unlike that proposed in [8], which is global to the entire similarity values matrix. This measure, denoted m, is defined for an entity e of the ontology O, by: $m = m_r - m_{nr}$ where m_r is the average of similarity values of enti-

ties that are associated to e and m_{nr} is the average of similarity values of entities that are not associated to e.

Other works such as [6] and [10] use machine learning techniques for automatic configuration of weights to be assigned to the matchers.

* The approach proposed in this paper is situated in the category of weighting techniques that combine the similarity values calculated by different matchers. It consists of a heuristic that estimates the weights to assign to the matchers. Contrary to the techniques mentioned above, this approach is of statistical nature and estimates the weights by an estimation of the precision standard metric.

4 Illustrative Example of the Approach

Let two ontologies O and O' which contain the concepts O: {Product, Provider, Creator} and O': {Book, Translator, Publisher, Writer} respectively.

The application of the edit distance metric and that based on WordNet between concepts of O and O' has generated the following two matrices of similarities.

1) If we filter out the matrix of similarities (Table 1) calculated with the edit distance, with a threshold s = 0.15 we obtain the following alignment:

   A1= {(Product, Translator), (Provider, Translator), (Provider, Publisher), (Provider, Writer), (Creator, Translator), (Creator, Writer)}.

Table 1. The Similarity Values Calculated by Edit Distance.

O /O'	Book	Translator	Publisher	Writer
Product	0.14	0.20	0.11	0.14
Provider	0.12	0.20	0.44	0.50
Creator	0.14	0.50	0.11	0.43

2) If we filter out the second matrix of similarities (Table 2) calculated using WordNet, with a threshold s = 0.15 we obtain the following alignment:

   A2= {(Product, Book), (Product, Writer), (Provider, Writer), (Provider, Book), (Creator, Book), (creator, Translator), (Creator, Writer)}.

Table 2. The Similarity Values Calculated Using WordNet.

O /O'	Book	Translator	Publisher	Writer
Product	0.18	0.12	0.12	0.15
Provider	0.17	0.11	0.14	0.29
Creator	0.18	0.47	0.12	0.15

The alignment A which consists of the semantic correspondences identified by the two matchers simultaneously is as follows: A = A1 ∩ A2 = {(Creator, Translator), (Provider, Writer), (Creator, Writer)}. A represents the estimator of final candidate alignment.

The estimator of the precision of the matcher edit distance is: P_1=3/6=0.50.

The estimator of the precision of the matcher based on WordNet is: P_2=3/7=0.43.

The weights to be assigned to matchers are: w_1 = 0.50 and w_2 = 0.43.

The matrix of the combined similarities is as follows:

Table 3. The Combined Similarity Values.

O /O'	Book	Translator	Publisher	Writer
Product	0.16	0.16	0.11	0.14
Provider	0.14	0.16	0.30	0.40
Creator	0.16	0.49	0.11	0.30

If we filter out the matrix of combined similarities (Table 3), with a threshold s = 0.30 we obtain the following alignment, (Provider, Publisher), (Provider, Writer), (Creator, Translator), (Creator, Writer)}. For more details about the approach see section 5.

5 The Proposed Approach

The architecture of our approach denoted PBW (Precision Based Weighting) is illustrated in Fig. 2. We have in input two ontologies O_1 and O_2 to be aligned. The ***Analysis Module*** performs the entities extraction from O_1 and O_2 using API Jena. Then, the ***Similarities Generation Module*** calculates for each pair of concepts (C, C') ∈ $O_1 \times O_2$ three similarity values using three techniques namely: the edit distance [5], the Jaro metric [5] and the similarity metric based on WordNet (Wu-Palmer algorithm) [5]. It should be noted at this level that the parameter object of the comparison is primarily the *aggregation method of similarity values* (the estimation of the weights to be assigned to matchers).

For that reason, we have set the same matchers for all three compared methods H, LCD (see the section 3 for the definition of these methods) and PBW in order to not have the results skewed by the choice of matchers. Therefore, the selection of the matchers has not been the subject of special attention. We have limited to the linguistic-based and string-based matchers. These similarity values are used by the ***Weights Estimation Module*** in order to calculate the confidence to be associated to the matchers mentioned above. The ***Similarities Combination Module*** generates then the combined similarity values using a weighting summation strategy. Finally, the ***Alignment Extraction Module*** selects the final alignment. This selection is simply performed by filtering the combined similarity values on a given threshold.

The contribution of the paper lies in the combination of similarity values. We detail below the principle of the proposed approach.

The approach proposed in this paper is an aggregation approach of similarity values calculated by several matchers. It fits into the category of automatic techniques for assigning weights to matchers which estimates their importance. We give in this section its principle.

Let O and O' be two ontologies to be aligned and let $M_1,\ldots, M_k$ k matchers which execute in parallel and calculate the similarity values between entities $e_1,\ldots, e_n$ for O and $e'_1,\ldots, e'_m$ for O' respectively. Let us note $S_1,\ldots, S_k$ the similarities matrices generated by matchers $M_1,\ldots, M_k$ respectively. The problem here is to assign to each matcher M_i a weight w_i which expresses its importance in a given alignment task.

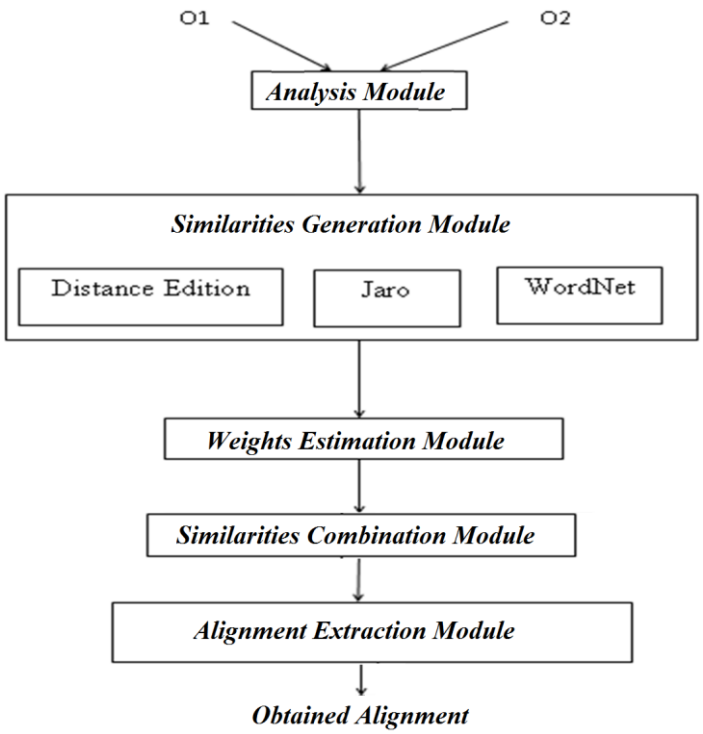

Fig. 2. The Architecture of the Application.

The intuition behind this approach consists to assign to the matcher M_i the weight w_i which is equal to an estimation of the precision of M_i. Indeed, as the precision metric is a good estimation of the matcher quality; we propose to use it as an estimator of the weight that will be assigned to the matcher.

We distinguish for the matcher M_i two subsets among the set of semantic correspondences between entities of ontologies O and O' to be aligned. On one hand we have the set P_i of the correspondences qualified positively and which belong to candidate alignment and On the other hand, we have the set N_i of those, negatively qualified and which do not belong to candidate alignment. The set P_i is defined as follows:

$P_i = \{(e_i, e'_j) \in O \times O' / (S_i(i, j) \geq s$ where s is a given threshold$\}$.

Since to estimate the precision of a matcher, we need a reference alignment and in the absence of such alignment we propose to estimate it (the reference alignment) by the set P which denotes the set of positive correspondences identified simultaneously by all matchers. In other words: $P = \cap_{i=1,\ldots,k} P_i$. We therefore propose for the matcher M_i the following estimator for the precision: $w_i = |P_i \cap P| / |P_i|$

Where |E| denotes the number of all elements of the set E. This estimator represents the weight to be assigned to M_i.

The approach can be made operational by the following process:

- Calculate, for each matcher M_i, the set P_i defined above.
- Calculate the set P. In some alignment tasks, the case where P is empty can occur. The weights assigned to the matchers are therefore is null. To overcome this situation, we have estimated P, for each matcher M_i, as follows: $P = \{(e_i, e'_j) \in O \times O' / (S_i(i, j) \geq s$ where s is a threshold relatively high$\}$. We have retained the following formula to specify the threshold: s = the highest similarity value calculated by the matcher M_i from which a particular constant value n is subtracted. For example, if the maximum similarity value is equal to 0.80 and n=0.25 then s = 0.8*0.25=0.40.
- Assign to each matcher M_i the weight w_i.
- Calculate the matrix of combined similarity values M. the matrix M is calculated by the following formula: $M(i, j) = (\sum_k w_k * S_k(i, j)) / (\sum_k w_k)$.
- Filter M according to the threshold s.

6 The Experimental Study

In order to evaluate our approach, we have used the conference track of OAEI 2012 evaluation campaign. This track consists of a collection of 16 ontologies describing the field of the conferences organization. It is constituted of 21 tests for which reference alignments are available, from a total of 120 possible tests resulting from the pairwise combination of 16 ontologies. Each test consists of two ontologies and a reference alignment.

The tests have been carried as follows: we have implemented the three methods (H, PBW and LCD), then we have executed these methods on ontologies tests of the conference track.

As evaluation criteria we have used the standard metrics that are precision, recall and F-measure to evaluate our approach. These metrics are defined as follows:

$$\text{Precision} = P(A,R) = \frac{|R \cap A|}{|A|} \quad \text{Recall} = R(A,R) = \frac{|R \cap A|}{|R|} \quad \text{F-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

Where |R| denotes the number of the reference alignment mappings and |A| denotes the number of matches found by our approach.

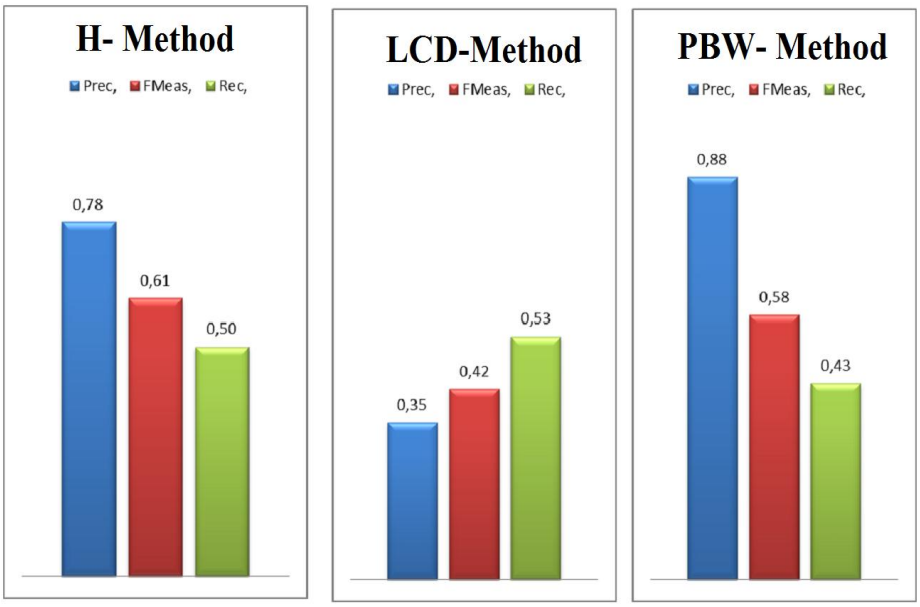

Fig. 3. The Global Results (All Tests of Conference Track) of the Three Methods

We envision in this experimental analysis to compare our precision based weighting approach of the similarity values aggregation (Noted PBW-method in the graphs) with the two most efficient aggregation methods [10] [15] namely the method based on the harmony concept [8] (Noted H-method in the graphs) and the method based on the concept of local confidence [2] (Noted LCD-method in the graphs).

It should be noted at this level that our approach as well as those with which it has been compared belong to the same category of methods based on the weighting.

We have adopted the following methodology in order to conceive the experimental protocol. For each of the 21 tests of the conference track, we have calculated three matrices of similarities by the matchers edit distance, Jaro and WordNet, respectively. Subsequently, from these matrices we calculated the weights to assign to the matchers by the three compared methods (harmony, local confidence and our method). Then we have calculated the matrices of the combined similarities and we have selected the alignments by filtering using a given threshold s for each of the three methods. Finally, for each test we have calculated precision, recall and F-measure for

each method. To conclude, we have calculated the average precision, recall and F-measure for all tests of the conference track.

The results are shown in Fig. 3 and Fig. 4.

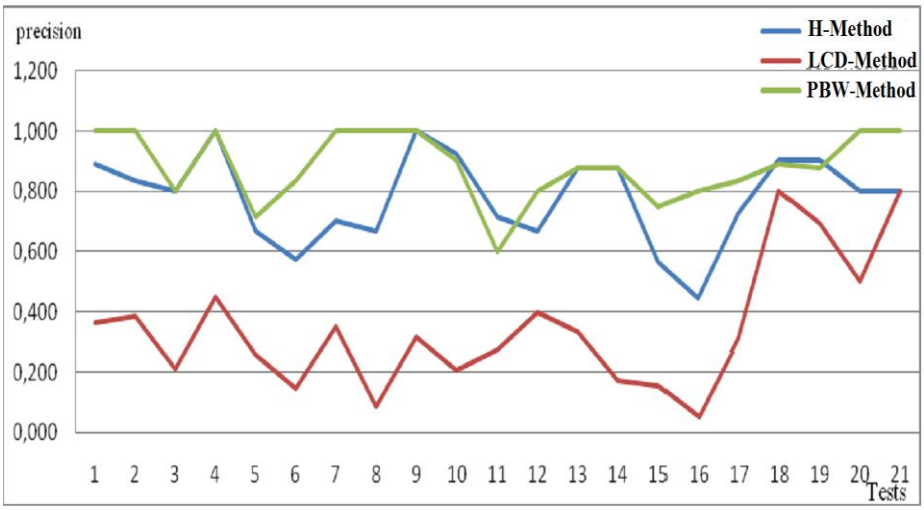

Fig. 4. The Detailed Results of the Three Methods in Terms of Precision

The experimental results obtained (Fig. 3) show that globally i.e. for all tests:

1. Our approach PBW is significantly more efficient than the H and LCD methods in terms of precision.
2. Our approach PBW is more efficient than the LCD method and slightly less efficient than the H method in terms of F-measure.
3. Our approach PBW is less efficient than the H and LCD methods in terms of recall.

The analysis of the detailed results on all tests of the conference track (Fig. 4) show that our approach is more efficient than H and LCD methods in terms of precision for all tests of the conference track of OAEI 2012 evaluation campaign, considered individually.

7 Conclusion and Perspectives

We have presented in this paper a dynamic approach to estimate automatically the weights to be assigned to different matchers in a given alignment task, in order to combine the similarity values calculated by the matchers in a context of ontology alignment.

The experimental results show the good performance of our proposed approach. Indeed, it is better in terms of precision than other methods, local and global, deemed among the most efficient ones in recent years. In addition, it shows a good F-measure relative compared to the local method.

As future perspective we envision to intensify the experiments by considering other tests and combining other similarities calculation techniques.

Bibliography

1. Bellahsene, Z., Duchateau, F.: Tuning for Schema Matching Schema Matching and Mapping. Bellahsene, Z., Bonifati, A., Rahm, E., Data-Centric Systems and Applications, Springer, (2011)
2. Cruz, I., Antonelli, F. P., Stroe, C.: Efficient selection of mappings and automatic quality-driven combination of matching methods. International Workshop on Ontology Matching, (2009)
3. Do, H., Rahm, E.: COMA - A system for flexible combination of schema matching approaches", Proceedings of the 28th VLDB Conference, Hong Kong, China, (2002)
4. Ehrig, M: Ontology Alignment: Bridging the Semantic Gap. Springer, (2007)
5. Euzénat, J., Shvaiko, P.: Ontology Matching. Springer, (2013)
6. Ichise, R.: Machine learning approach for ontology mapping using multiple concept similarity measures. ACIS-ICIS, IEEE Computer Society, (2008)
7. Li, J., Tang, J., Li, Y., Luo, Q.: RiMOM: A Dynamic Multistrategy Ontology Alignment Framework. IEEE, Transactions On Knowledge and Data Engineering, Vol. 21, (2009)
8. Mao, M., Peng, Y., Spring, M.: A harmony based adaptive ontology mapping approach. In Proceedings of international conference on semantic web and web services (SWWS), (2008)
9. Martinez-Gil, J., Alba, E., Aldana-Montes, J.: Optimizing Ontology Alignments by Using Genetic Algorithms. In Christophe Gueret, Pascal Hitzler, and Stefan Schlobach, editors, Nature inspired Reasoning for the Semantic Web, CEUR Workshop Proceedings, (2008)
10. Ngo, D.: Enhancing Ontology Matching by Using Machine Learning, Graph Matching and Information Retrieval Techniques. Thèse de doctorat de l'université de Grenoble, (2012)
11. Rahm, E.: Towards Large-Scale Schema and Ontology Matching. Schema Matching and Mapping, Zohra Bellahsene, Angela Bonifati, Erhard Rahm, (Eds.), Data-Centric Systems and Applications, Springer, (2011)
12. Silvana, C., Ferrara, A., Montannelli, S., Varese, G.: Ontology and Instance Matching. Lecture Notes in Computer Science Vol. 6050, (2011)
13. Wang, J., Ding, Z., Jiang, C.: GAOM: Genetic Algorithm based Ontology Matching. In Proceedings of IEEE Asia-Pacific Conference on Services Computing, (2006)
14. Wang, R., Wu, J., Liu, L.: Strategies Prediction and Combination of Multi-strategy Ontology Mapping. ICICA 2010, Part II, CCIS 106, Springer-Verlag, Heidelberg, (2010)
15. (Site 1). http://oaei.ontologymatching.org/2011/results/oaei2011.pdf (accessed January 2015)