Journal of Information Hiding and Multimedia Signal Processing ©2016 ISSN 2073-4212
Ubiquitous International Volume **7**, Number **2**, March **2016**

A Novel Ontology Matching Technology Based on NSGA-II

Li Jiang

Computer Department
Fuzhou Polytechnic
No 8 Lianrong Road, University Town, Minhou, Fuzhou, Fujian, Chian, 350108
86068374@qq.com

Xingsi Xue*, Pei-Wei Tsai, and Jeng-Shyang Pan

School of Information Science and Engineering
Fujian University of Technology
No 3 Xueyuan Road, University Town, Minhou, Fuzhou, Fujian, China, 350118
*Corresponding author: xxs@fjut.edu.cn; pwtsai@foxmail.com; jengshyangpan@fjut.edu.cn

Received September, 2015; revised November, 2015

Abstract. *Ontology is constructed or researchers to overcome the heterogeneous problem in a domain, but merely using ontology may raise the heterogeneous problem to a higher level. To solve the heterogeneous problem between two ontologies, it is necessary to determine the relationships that hold between the entities in them. The process of finding these correspondences is called ontology matching and the matching results are called ontology alignment. Various ontology matching approaches have been proposed so far, and the Evolutionary Algorithm (EA) based ontology matching technologies have been attracting more and more attentions, although the quality of the alignments obtained and the efficiency of the algorithm are both barely satisfactory. To address these issues in EA based ontology matching technologies, in this paper, an novel ontology matching technology based on NSGA-II is presented. In particular, in our work, a novel similarity measure based on Information Theory and a special mapping extraction approach based on the Naive Descending Extraction (NDE) algorithm are respectively proposed, a Multi-objective optimal model for ontology matching problem is presented and the problem-specific NSGA-II is designed. Experimental results show that our proposal is efficient and can find the best solution so far.*
Keywords: Ontology alignment, NSGA-II, Similarity measure, Mapping extraction approach.

1. **Introduction.** Ontologies are generally regarded as the solution to enable the interoperability between heterogeneous semantic data sources. However, because of human subjectivity, one entity (such as class, property or individual) in different ontologies can be defined with different names or in different ways. Therefore, merely using ontology may raise the heterogeneity problem to a higher level [1]. In order to address this issue, it is necessary to identify the correspondences between the entities of separate ontologies through so called ontology matching process.

It is highly impractical to match two ontologies manually especially when the size of the ontologies is considerably large, therefore, several ontology matching technologies have arisen over the years. Through various similarity measures, each of them could provide a numerical value of similarity between elements from separate ontologies that can be used

to decide whether those elements are semantically similar or not. In general, similarity measures could be categorized in syntactic, linguistic and taxonomy-based measures. Syntactic measures compute a string distance or edit distance between the ontology entities, and two widely used syntactic measures are Levenstein distance [9] and Jaro distance [10]. Linguistic measures calculate the similarity between ontology entities by considering linguistic relations such as synonymy, hypernym, and so on. Linguistic measure based on WordNet [3], which is an electronic lexical database where various senses of words are put together into sets of synonyms, has become very popular in ontology matching domain in recent years. Taxonomy-based measures consider only the specialization relation. The intuition behind taxonomic measures is that subsumption relation connect terms that are already similar, therefore, their neighbors may be also somehow similar. For instance, if super-concepts are the same, the actual concepts are similar to each other; if sub-concepts are the same, the compared concepts are also similar. The most notable ontology matching technology based on taxonomy-based measure is similarity flooding algorithm [11]. However, none of these similarity measures could performance better than others in all scenarios, which highly affects the quality of the alignments obtained by these ontology matching systems. Therefore, how to design a highly semantic recognizable similarity measure has become one of the critical problems for the success of ontology matching technology.

Moreover, since modeling two ontologies under alignment is a complex (nonlinear problem with many local optimal solutions) and time consuming task (large scale problem), particularly when the considered ontologies are characterized by a significant number of entities (resulting in large scale problem), approximate methods are usually used for computing the correspondence. From this point of view, evolutionary optimization methods could represent an effective approach for addressing this problem. Among ontology matching systems that make use of a evolutionary algorithm, the most notable one is GOAL (Genetics for Ontology ALignments) [12]. GOAL does not directly compute the alignment between two ontologies, but it determines, through a genetic algorithm, the optimal weight configuration for a weighted average aggregation of several similarity measures by considering a reference alignment. The same idea of implementing a meta-matching system to combine multiple similarity measures into a single aggregated metric is also developed in two more recent papers [13][14]. However, all of these systems work with only one of several common measures that used to evaluate the quality of an alignment, and these measures could simply evaluate the aligning results in one aspect, which could lead to the bias improvement of the alignment and decrease the final alignment obtained. Moreover, the time consumption of existing evolutionary algorithm based approach is high in general. Therefore, how to design an efficient evolutionary algorithm based ontology matching technology to automatically determine the correspondences between the entities in two ontologies is another challenge in ontology matching domain [8].

Aiming at these two problems mentioned above, in this paper, we propose a novel similarity measure based on information theory [15] and a special mapping extraction approach to respectively improve the quality of ontology alignments and decrease the time consumption of the ontology matching process. On this basis, a multi-objective optimal model for the ontology matching problem is constructed, and a problem-specific NSGA-II [6] is designed to solve the ontology matching problem. The rest of the paper is organized as follows: Section 2 present a novel similarity measure based on Information Theory and a mapping extraction approach; Section 3 construct a multi-objective optimal model for the ontology matching problem, and gives the details of problem-specific NSGA-II for the ontology matching problem; Section 4 shows the experimental results and finally, in Section 5, we draw the conclusions.

2. Similarity Measure and Mapping Extraction Strategy.

2.1. Similarity Measure based on Information Theory.

The foundation of ontology matching technology is the similarity of entities [2]. In this paper, by referring to Shannon's information theory, we propose a semantic measure that is designed for relationship between different entities. Our measure is designed to combine a comprehensive set of syntax-level, lexical-level and structure-level measures of similarity to calculate the similarity value between entities.

To be specific, our approach assess the similarity of concepts according to the amount of information they provide, i.e. their Information Content (IC). In order to accurately estimate the IC of concepts and avoid depending on annotated corpora, whose creation is time consuming and sometimes difficult to obtain, we propose to estimate the IC of concepts by considering structural information extracted from the ontology. In our work, for each concept, we construct a profile by collecting the information, e.g. label, comment and property, from all its descendants. Then, the similarity of two entities e_1 and e_2, which respectively come from two ontologies O_1 and O_2, can be calculated by the following two asymmetric measures:

$$sim_1(e_1, e_2) = \frac{|prof(e_1) \bigcap prof(e_2)|}{|prof(e_1)|} \tag{1}$$

$$sim_2(e_1, e_2) = \frac{|prof(e_1) \bigcap prof(e_2)|}{|prof(e_2)|} \tag{2}$$

where $|prof(e_1)|$ and $|prof(e_2)|$ refers to the cardinality of the profiles on e_1 and e_2 respectively, $|prof(e_1) \bigcap prof(e_2)|$ refers to the number of similar elements shared by $prof(e_1)$ and $prof(e_1)$. If $0 \leq |sim_1(e_1, e_2) - sim_2(e_1, e_2)| \leq \delta$, e_1 and e_2 are semantically similar. In this study, δ is a threshold to measure the semantic equivalence between $sim_1(e_1, e_2)$ and $sim_2(e_1, e_2)$. Generally, δ should be set relatively small to reflect $sim_1(e_1, e_2)$ and $sim_2(e_1, e_2)$ has little difference when the entity e_1 and e_2 are semantically equivalent. However, if δ is set too small, we will miss many semantically equivalent terms. To obtain a suitable δ, we conduct experiments and find the semantic equivalence performs well when δ is assigned 0.1.

In particular, in our work, the similarity value of two profile elements is computed by calculated by aggregating SMOA distance [1], which is the most performing syntax measure for the ontology matching problem, and a linguistic measure, which calculate a synonymy-based distance through an electronic lexical database WordNet [3]. The aggregated similarity value equals:

- 1, if the SMOA distance or linguistic measure equals 1;
- the average of SMOA distance and linguistic measure, otherwise.

In our work, the threshold of the aggregated similarity value is empirically set as 0.9, i.e. two elements are considered to be similar if their aggregated similarity value is above 0.9.

2.2. Mapping Extraction Strategy.

When an alignment is obtained, a corresponding similarity matrix M can be generated as follows: each cell in M can be regarded as a candidate correspondence (its position in ith row and jth column represent entities e_{Si} and e_{Tj} from the source ontology and target ontology respectively, and the value of the cell represents the similarity between e_{Si} and e_{Tj}). Next, in order to filter the less promising mappings, a Naive Descending Extraction (NDE) algorithm [4] is utilized to extract the mappings from M. Specifically, there exists three steps as follows: (1) all cells are sorted in descending order according to their similarity values; (2) all the cells with the maximum value are recorded in a list L; (3) all the other cells values in M whose row or column is

the same as those in L are set to zero. The above steps iterate until no cell with value greater than 0 remains.

Due to the fact that the smaller the value of a cell is, the smaller the possibility that the candidate correspondence the cell represents is actually a real one is. So, we propose a mapping extraction strategy to improve the efficiency of naive descending extraction by modifying the condition of termination. In our proposal, the algorithm iterates until no cells value is greater than the threshold 0.5, which is set in empirical way to achieve the highest average alignment quality on all test cases of exploited dataset.

3. NSGA-II for Optimizing Ontology Alignments.

3.1. Multi-Objective Optimal Model for Optimizing Ontology Alignments.

In this section, the optimal model for the ontology matching problem is presented as follows:

$$\begin{cases} max f(X) = max(Recall(X), Precision(X)) \\ s.t. \quad X = (x_1, x_2, \cdots, x_n)^T, \\ \qquad x_i \in [1, |entitySetO_2|], i = 1, \cdots, n, \end{cases} \tag{3}$$

where $n = |entitySetO_1|$, $n = |entitySetO_1|$ and $|entitySetO_2|$ represents the cardinalities of the entity set of ontology O_1 and O_2, respectively. In this model, $x_i, i = 1, \cdots, n$, represents the i-th pair of the entity mapping, and the objective of this model is to maximize both recall and precision [5] of the alignment. Since the two objectives are contradictory, this problem is regarded as a multi-objective optimizing problem that could be solved by NSGA-II.

3.2. Chromosome encoding.

Let n_1 be the number of entities in ontology O_1, and n_2 be the number of entities in ontology O_2. Each chromosome in the population would be a one-dimensional array with n_1 integer elements. The first n_1 elements take values between 1 and n_2, denoted as $N_1 N_2, \cdots, N_{n_1}$, where $N_i = M(i) \in \{1, 2, \cdots, n_2\}$, and this means that the i-th instance in O_1 is mapped to the N_i-th entity in O_2.

3.3. Fitness Functions.

In our work, two evaluation criterions, namely recall and precision (f-measure is only the harmonic mean of recall and precision), are used to evaluate which individual can obtain a relatively better alignment.

3.4. Genetic operators.

3.4.1. *Selection.* As in nature, a selection process provides the mechanism for selecting better solutions to survive. First, we use Euclidean distance to work out the T (T is the number of neighbors of each weight vector) closest weight vectors to each weight vector. For each $i = 1, 2, ..., N$, set $C(i) = \{i_1, i_2, ..., i_T\}$, where $\lambda_{i_1}, \lambda_{i_2}, ..., \lambda_{i_T}$ are the T closest weight vectors to λ_i. Then, we randomly generate a number in the range $[0, 1]$, and if the number is smaller than the selection probability ps, we will randomly select two index k, l from $C(i)$, otherwise, we will select two index k, l from $\{1, 2, ..., N\}$.

3.4.2. *Crossover.* We check if the crossover could be applied according to the crossover probability pc, and if it is, the new individual y' is then generated from x^k, x^l, otherwise, the new individual y' is generated randomly. For crossover, we use the common one-cut-point method to carry out the crossover operation. First, a cut position in x^k, x^l is randomly determined and this position is a cut point which cuts x^k and x^l into two parts: the left part and the right part. Then, we simply combine the left part of x^k and the right part of x^l to form the new individual y'.

3.4.3. *Mutation.* Mutation operator assures diversity in the population and prevents premature convergence. In our work, for each bit in the individual, we check if the mutation could be applied according to the mutation probability and if it is, the value of that bit is then flipped.

3.5. **Generation of New Population.** First, we combine the parent population and the current population together and remove the redundancy of the chromosomes. Then, the new population is selected by the non-dominated-sorting and the crowd-distance calculation [6].

3.6. **Elite Strategy.** In our algorithm, a best individual is saved as a best solution found for the problem so far. Specifically, we first initialize the best individual as the one which has the largest f-measure in the initial population. Then, in the following generations, if the f-measure value of the elite of the current population is larger than that of the best individual, the best individual will be replaced by the elite of the current population. Finally, when the algorithm terminates, the best individual is recommended to the user as the best solution found for the problem of optimizing ontology alignments.

3.7. **Efficiency Improvement Strategy.** In the algorithm, the most time consuming process is the fitness computation, namely the process of reading alignments, aggregating alignments and evaluating the final alignment obtained. It is particularly worth noting that alignments to be aggregated may need to be read into memory for several times during this process. Therefore, it is inevitable to consume a lot of time. To avoid repeatly reading the alignments, we propose to read the n alignments to n similarity matrices all at once before the algorithm starts, where the ith row and the jth column of the similarity matrix represent the entities e_{Si} and e_{Tj} of the ontologies O_1 and O_2 respectively and the corresponding value in the matrix is the confidence measure of the mapping between e_{Si} and e_{Tj}. In this way, the efficiency of the algorithm can be improved dramatically.

4. **Experimental Results and Analysis.** In our work, we use the well-known benchmark provided by the Ontology Alignment Evaluation Initiative (OAEI) 2012[7]. Table 1 shows a brief description about the benchmarks.

Table 1. Brief Description of Benchmarks

ID	Brief description
101-104	The ontologies under alignment are the same or the first one is the OWL Lite restriction of the second one
201-210	The ontologies under alignment have the same structure, but different lexical and linguistic features
221-231	The ontologies under alignment have the same lexical and linguistic features, but different structure
301-304	The ontologies under alignment are real world cases

4.1. **Experiments Configuration.** The NSGA-II uses the following parameters which represent a trade-off setting obtained in empirical way to achieve the highest average alignment quality on all test cases of exploited dataset, which is robust against the heterogeneous situations in our experiment.

- Population size = 20 individuals,
- Crossover probability = 0.60,
- Mutation probability = 0.01,

- Max evaluation time = 250.

It should be noted that, with respect to the values of the above parameters, such as population size and max evaluation time, they could be higher if the scale of the researching domain is large, and the data scale in our work is medium.

4.2. **Results and Analysis.** Table 2 shows the comparison of the evaluation metrics obtained through NSGA-II using the original approach and improved approach. In Table 2, Symbol R and P represent recall and precision [7] value respectively. As it can be

TABLE 2. Comparison of the alignments quality obtained through original approach with those obtained through improved approach

ID	$F-measure(R, P)$ Original Approach	$F-measure(R, P)$ Improved Approach
101	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
103	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
104	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
201	0.94 (0.90, 0.98)	0.94 (0.90, 0.98)
203	0.99 (0.98, 1.00)	1.00 (1.00, 1.00)
204	0.98 (0.98, 0.99)	1.00 (1.00, 1.00)
205	0.93 (0.89, 0.99)	0.95 (0.94, 0.97)
206	0.70 (0.67, 0.73)	0.82 (0.70, 0.99)
221	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
222	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
223	0.99 (0.98, 1.00)	1.00 (1.00, 1.00)
224	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
225	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
228	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
230	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
231	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
301	0.76 (0.76, 0.76)	0.86 (0.81, 0.91)
302	0.73 (0.63, 0.88)	0.77 (0.63, 1.00)
304	0.93 (0.88, 0.99)	0.97 (0.95, 0.99)

seen from Table 2, the recall, precision and f-measure value obtained through NSGA-II using our improved approach are much better than that using the original approach. Specifically, in benchmarks 101, 103, 104, 201, 221, 222, 224, 225, 228, 230, and 231, the solutions found by NSGA-II using our improved approach are equal to that using the original approach. In benchmarks 203, 204, 205, 206, 223, 301, 302, and 304, NSGA-II using our improved approach found much better results than that using the original approach. Therefore, we may draw the conclusion that the performance of our improved approach is much better than the original approach. Table 3 shows the running time taken by NSGA-II using the original approach and the improved approach. It can be seen clearly from the table that the improvement of the running time taken by NSGA-II using our improved approach over that using the original approach is greater than 95% in all benchmarks except 101, 103, and 104 whose improvement is about 80%.

Table 4 shows the comparison of our approach with the participants in OAEI 2012 where the numbers inside are the average of all testing cases. It can be seen clearly from the table that our approach outperforms all of the participants in OAEI 2012 in terms of recall, precision and f-measure. Thus we may draw the conclusion that the efficiency

Table 3. Comparison of the running time consumption of the original approach and that of the improved approach

ID	Time (ms) Original Approach	Time (ms) Improved Approach	Improvement (%)
101	45594	9297	80%
103	53500	10391	81%
104	51797	10296	80%
201	1294515	17266	99%
203	1023609	13032	99%
204	953234	13547	99%
205	923047	12516	99%
206	938703	11375	99%
221	997375	13250	99%
222	982781	12516	99%
223	1179938	16766	99%
224	945500	9984	99%
225	997062	13031	99%
228	204906	7672	96%
230	803625	16172	98%
231	1036609	13516	99%
301	469375	12469	97%
302	336625	12000	96%
304	820906	13391	98%

Table 4. Comparison of our approach with the participants in OAEI 2012

System	Recall	Precision	F-Measure
MapSSS	0.77	0.99	0.87
YAM++	0.72	0.98	0.83
AROMA	0.64	0.98	0.77
AUTOMSv2	0.54	0.97	0.69
WeSeE	0.53	0.99	0.69
Hertuda	0.54	0.90	0.68
HotMatch	0.50	0.96	0.66
Optima	0.49	0.89	0.63
WikiMatch	0.54	0.74	0.62
ServOMap	0.43	0.88	0.58
LogMap	0.45	0.73	0.56
MaasMatch	0.57	0.54	0.56
MEDLEY	0.50	0.60	0.54
ServOMapLt	0.20	1.00	0.33
ASE	0.54	0.49	0.51
Our approach	0.94	0.99	0.96

of NSGA-II using our improved approach is much higher than that using the original approach.

5. **Conclusion.** Ontology matching plays an increasing important role in ontology engineering, and in this paper, ontology matching process is regarded as a optimization problem. Since the evaluation of the quality of an alignment has two conflicting aspects, recall and precision, the optimization of optimizing ontology alignments can be viewed as a two objectives optimization problem. Then, NSGA-II, a multi-objective evolutionary algorithm, is used to address it. However, the solution and the efficiency of the original approach based on NSGA-II are both barely satisfactory. To improve the quality of the solution and the efficiency of searching process, we first propose a similarity measure based on the information theory , and then present a mapping extracting strategy to improve the efficiency of the NSGA-II based ontology matching technology. Our experimental results show that our proposal can efficiently determine the ontology alignment in terms of both the quality of the matching results and the running time.

Acknowledgment. This work is supported by the Education Department of Fujian Province Science and Technology Projects (No. JB12314) and the National Natural Science Foundation of China (No. 61503082).

REFERENCES

[1] J. Euzenat and P. Shvaiko, *Ontology Matching*, Springer, Berlin-Heidelberg, Germany, pp.80, 2007.
[2] A. Maedche and S. Staab, Measuring Similarity between Ontologies, *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management*, pp.251-263, 2002.
[3] G. A. Miller, WordNet: A lexical database for English, *Communications of the ACM*, vol.38, no.11, pp.39-41, 1995.
[4] C. Meilicke and H. Stuckenschmidt. Analyzing mapping extraction approaches, *Proceedings of the 6th International Semantic Web Conference*, 2007.
[5] C. J. Van Rijsbergen, *Information Retrieval*, Springer, Butterworth, London, 1975.
[6] K. Deb, S. Agrawal and A. Pratap, A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II, *Proceedings of the Parallel Problem Solving from Nature VI Conference*, vol.1917, pp.849-858, 2000.
[7] Ontology Alignment Evaluation Initiative (OAEI), Available at http://oaei.ontologymatching.org/2012/, Accessed on April 15, 2015.
[8] P. Shvaiko and J. Euzenat, Ontology matching: state of the art and future challenges, *IEEE Transactions on Knowledge and Data Engineering*, vol.25, no.1, pp.158-176, 2013.
[9] V. Levenshtein, Binary Codes Capable of Correcting Deletions, Insertions and Reversals, *Soviet Physics-Doklady*, vol.10, pp. 707-710, 1966.
[10] A. Maedche and S. Staab, Measuring Similarity between Ontologies, *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management*, pp.251-263, 2002.
[11] S. Melnik, H. Garcia-Molina and E. Rahm, Similarity flooding: A versatile graph matching algorithm and its application to schema matching, *Proceedings of 18th International Conference on Data Engineering*, pp.117-128, 2002.
[12] J. Martinez-Gil, E. Alba and J. F. Aldana-Montes, Optimizing ontology alignments by using genetic algorithms, *Nature Inspired Reasoning for the Semantic Web*, vol.419, pp.31-45, 2008.
[13] J. M. V. Naya, M. M. Romero and J. P. Loureiro, Improving ontology alignment through genetic algorithms, *Soft Computing Methods for Practical Environment Solutions: Techniques and Studies*, pp.240-259, 2010.
[14] A-L. Ginsca and A. Iftene, Using a genetic algorithm for optimizing the similarity aggregation step in the process of ontology alignment, *Proceedings of 9th Roedunet International Conference (RoEduNet)*, pp.118-122, 2010.
[15] P. Resnik, Using Information content to evaluate semantic similarity in a taxonomy, *Proceedings of 14th International Joint Conference on Artificial Intelligence*, vol.1, pp.448C53, 1995.