Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **5**, September **2017**

Efficient Ontology Meta-Matching Using Alignment Prescreening Approach and Gaussian Random Field Model assisted NSGA-II

Xingsi Xue

Fujian Provincial Key Laboratory of Big Data Mining and Applications
College of Information Science and Engineering
Fujian University of Technology
No.3 Xueyuan Road, University Town, Minhou, Fuzhou, Fujian, 350118, China
jack8375@gmail.com

Received February, 2017; revised May, 2017

Abstract. *Multi-Objective Evolutionary Algorithm (MOEA) is emerging as a new methodology to tackle the ontology meta-matching problem. However, for dynamic applications, besides the alignment's quality, runtime and memory consumption in the matching process are also of great importance. In this paper, we propose an efficient NSGA-II based ontology meta-matching technology to improve the efficiency of NSGA-II based ontology meta-matching technology. In particular, our approach can automatically prescreen the less promising ontology alignments to be combined, which can reduce the search space of NSGA-II and improve its runtime, and reduce the number of exact individual evaluations by using Gaussian Random Field Model (GRFM), which can decrease the memory consumption of NSGA-II. The experimental results show that the utilization of alignment prescreening approach and GRFM is able to significantly improve the efficiency without sacrificing the alignment's quality.*
Keywords: Ontology meta-matching, GRFM, NSGA-II

1. **Introduction.** Multi-Objective Evolutionary Algorithms (MOEA) is emerging as a new methodology to tackle the ontology meta-matching problem [2]. However, for dynamic applications, besides the alignment's quality, runtime and memory consumption in the matching process are also of great importance. In this paper, we propose an improved NSGA-II [1] to optimize the meta-matching process. Particularly, an alignment prescreening approach is first proposed to prescreen the less promising ontology alignments and reduce the search space, and then the Gaussian Random Field Model (GRFM) is used to speed up NSGA-II and decrease the memory consumption during the ontology meta-matching process.

The rest of this paper is organized as follows: section 2 introduces the basic definitions and the multi-objective optimal model of ontology meta-matching problem; section 3 describes the alignment prescreening approach; section 4 formulates the GRFM assisted NSGA-II; section 5 presents the experimental studies and analysis; finally, section 6 draws conclusions.

2. **Multi-Objective Ontology Meta-matching.** In this work, an ontology is defined as $O = (C, P, I, \Lambda, \Gamma)$ [3], where C, P, I, Λ, Γ are respectively referred to the set of classes, properties, instances, axioms and annotations. In addition, an ontology alignment A

between two ontologies is a correspondence set and each correspondence inside is defined as $(e_{O_1}, e_{O_2}, conf, =)$, where e_{O_1} and e_{O_2} are the entities of two ontology O_1 and O_2, respectively, $conf \in [0, 1]$ is a confidence value holding for the correspondence between e_{O_1} and e_{O_2}, $=$ is the relation of equivalence.

Since in the golden alignment, one entity in source ontology is matched with only one entity in target ontology and vice versa, based on the observations that the more correspondences found and the higher mean similarity values of the correspondences are, the better the alignment quality is [4], we propose the following ontology alignment quality measure:

$$\begin{cases} max\ (MF(X), avgSim(X)) \\ s.t.\ X = (x_1, x_2, \ldots, x_n)^T, x_i \in [0, 1] \\ \sum_{i=1}^{n-1} x_i = 1 \end{cases} \tag{1}$$

where the decision variable X is a n-dimension vector where $x_i, i \in [1, n-1]$ represents the i-th alignment's weight to be aggregated and x_n the threshold for filtering the aggregated alignment, and MF and $avgSim$ are the functions that respectively calculating X's corresponding alignment's MatchFeasure [5] and the mean similarity value of all the correspondences inside.

3. **Alignment Prescreening Approach.** It's obvious that the poorly performed ontology alignments are those having large distances from the aggregated alignment. In order to distinguish the poorly performed ontology alignments, given a set of similarity matrices $\{S_j\}$, we define the bias ratio BR of multiple similarity matrices as follows:

$$BR(\{S_j\}) = \frac{\sum_j (\sum_{e_{O_1} \to e_{O_2}} p(Map_j(e_{O_1}, e_{O_2}) | S_j, Map_{\{j\}}(e_{O_1}, e_{O_2})))}{total\_number} \tag{2}$$

where:

- $p(Map_j(e_{O_1}, e_{O_2}) | S_j, Map_{\{j\}}(e_{O_1}, e_{O_2}))$ is the difference probability of mapping (e_{O_1}, e_{O_2}) between aggregated mapping $Map_{\{j\}}$ and S_j's mapping Map_j, which can be calculated by the following formula:

$$p(Map_j(e_{O_1}, e_{O_2}) | S_j, Map_{\{j\}}(e_{O_1}, e_{O_2})) = \frac{|sim_{Map_j}(e_{O_1}, e_{O_2}) - sim_{Map_{\{j\}}}(e_{O_1}, e_{O_2})|}{max(sim_{Map_j}(e_{O_1}, e_{O_2}), sim_{Map_{\{j\}}}(e_{O_1}, e_{O_2}))}$$

  where $sim_{Map_j}(e_{O_1}, e_{O_2})$ and $sim_{Map_{\{j\}}}(e_{O_1}, e_{O_2})$ refer to the similarity value of e_{O_1} and e_{O_1} in Map_j and $Map_{\{j\}}$ respectively;
- $total\_number$ is the number of (e_{O_1}, e_{O_2})whose similarities in the aggregated matrix and each S_j do not both equal 0.

In this work, the threshold is set as 0.25 and the similarity matrix with $BR > 0.25$ will be discarded. In this way, if the average biases of all the similarity matrices are larger than the threshold, then merely one similarity matrix with the lowest BR will be selected as the final similarity matrix.

4. **Gaussian Random Field Model Assisted NSGA-II.** GRFM can be integrated into evolutionary optimization procedures in two different ways: (1) some generations are evaluated by the true objective function and some other generations are evaluated solely by the metamodel; (2) in each generation (apart from the very first one), metamodels and exact evaluation function are used in a cooperative manner. The second approach, which is adopted in our work, turns out to be quite robust and proved to be successful in many applications[6]. In order to filter individuals which are not promising, the offspring population's individuals need to be ranked based on $\hat{y}(x)$, which is predicted through

Table 1. Outline of Gaussian Random Field Model assisted NSGA-II.

The outline of Gaussian Random Field Model assisted NSGA-II
$generation = 0$;
Initialize the Population P_t;
Evaluate P_t and insert results into D;
while $generation < maxGeneration$ do
P'_t = generateByGeneticOperators(P_t);
Evaluate P'_t with GRFM derived from D;
Choose individual set $Q_t \subseteq P'_t$ according to CR;
$P_{(t+1)}$ = rankAndSelect($Q_t \cup P_t$);
$generation = generation + 1$;
end while

Table 2. Brief Description of OAEI 2016's Bibliographic Track.

ID	Brief description
1XX	Two ontologies have same structure, lexical and linguistic features
2XX	Two ontologies have different structure, lexical or linguistic features
3XX	Two ontologies are from real world's applications

metamodel and its corresponding standard deviation $\hat{s}(x)$. Moreover, we use the following formula to calculate the predicted value $\hat{f}(x)$ instead of using $\hat{y}$ directly [7]:

$$\hat{f}(x) = \hat{y}(x) + \hat{s}(x); \tag{3}$$

Once a non-dominated set is found by NSGA-II, a Constant Ratio (CR) selecting strategy[8] is applied to choose the most promising offspring for precise evaluation. The CR strategy makes extensive use of the GRFM information and thus it has the potential to improve the convergence significantly. In our work, we set the selecting ratio of CR to 0.25 and the outline of the GRFM assisted NSGA-II is presented in Table 1.

5. **Experimental Studies and Analysis.** In the experiment, we utilize the bibliographic track of OAEI 2016 [9] to test our approach's performance, whose brief description is shown in Table 2.

5.1. **Experimental Setup.** The similarity measures used in this work are as follows:

- Levenshtein distance based Syntactic Measure [10],
- Wordnet based Linguistic Measure [11],
- Similarity Flooding algorithm based Taxonomy Measure [12].

In our work, 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. Through the configuration of parameters chosen in this way, it has been justified by the experiments in this paper that parameters chosen are robust for all the heterogeneous problems presented in the testing cases, and it is hopeful to be robust for the common heterogeneous situations in the real world.

- Numerical accuracy = 0.01,
- Population scale = 200,
- Crossover probability = 0.6,
- Mutation probability = 0.02,
- Maximum generation = 3000.

TABLE 3. Friedman's test on the execution time taken per generation. Each value represents the execution time (second), the number in round parentheses is the corresponding computed rank, and approach A, B and C respectively refer to the approach using NSGA-II, prescreening approach and NSGA-II, GRFM assisted NSGA-II.

ID	Approach A	Approach B	Approach C	Our Approach
101	1.766 (4)	0.809 (2)	1.341 (3)	0.791 (1)
103	1.941 (4)	0.895 (2)	1.298 (3)	0.813 (1)
104	1.936 (4)	0.882 (2)	1.368 (3)	0.841 (1)
201	26.237 (4)	15.145 (2)	23.667 (3)	14.527 (1)
203	23.129 (4)	20.339 (3)	20.010 (2)	16.760 (1)
204	23.137 (4)	14.123 (2)	18.344 (3)	11.757 (1)
205	22.538 (4)	14.861 (2)	16.904 (3)	11.636 (1)
206	22.593 (4)	15.905 (2)	16.954 (3)	13.676 (1)
221	23.208 (4)	15.637 (1)	17.839 (3)	15.909 (2)
222	22.472 (4)	15.743 (2)	18.481 (3)	13.501 (1)
223	28.851 (4)	19.707 (3)	17.207 (2)	13.419 (1)
224	22.796 (4)	15.126 (2)	17.109 (3)	14.736 (1)
225	23.220 (4)	13.194 (2)	19.431 (3)	10.852 (1)
228	5.622 (4)	2.520 (2)	4.243 (3)	2.486 (1)
230	19.158 (4)	15.480 (1)	16.435 (3)	15.921 (2)
231	22.996 (4)	16.855 (2)	20.010 (3)	15.801 (1)
301	11.337 (4)	8.009 (2)	9.525 (3)	7.900 (1)
302	7.734 (4)	6.530 (3)	5.776 (2)	5.124 (1)
304	17.247 (4)	14.525 (2)	16.926 (3)	10.565 (1)
Average	17.259 (4)	11.871 (2.05)	13.362 (2.84)	10.618 (1.11)

TABLE 4. Holm's test on the execution time taken per generation. Approach A, B and C respectively refer to the approach using NSGA-II, prescreening approach and NSGA-II and GRFM assisted NSGA-II.

i	Approach	z value	unadjusted $p-$value	$\frac{\alpha}{k-i}, \alpha = 0.05$
3	approach B	2.2442	0.0248	0.05
2	approach C	4.1303	$3.6229 \times e^{-5}$	0.025
1	approach A	6.900	$5.2003 \times e^{-12}$	0.0166

5.2. **Results and Analysis.** All the experimental results in the tables are the average values over ten independent runs. Specifically, Tables 3 to 6 show the statistical comparisons on their execution time and memory consumption per generation, respectively. Finally, Tables 7 and 8 show the statistical comparison among three single objective EA based ontology meta-matching approaches and our approach.

As can be seen from Table 3, in the Friedmans test, $\mathcal{X}_r^2 = 51.30 > \mathcal{X}_{0.05}^2 = 7.815$, which means there exists a significant difference between these approach and Holm's test is needed to further determine the concrete difference among them. In this work, the significance level of Holm's test $\alpha = 0.05$ and the results of Holm's test are shown in Table 4. As can be seen from Table 4, it is obvious that our proposal statistically outperforms other approaches on execution time at 0.05 significance level.

TABLE 5. Friedman's test on the memory consumption per generation by evaluation function. Each value represents the memory consumption (gigabyte) per generation, the number in round parentheses is the corresponding computed rank, and approach A, B and C respectively refer to the approach using NSGA-II, prescreening approach and NSGA-II and GRFM assisted NSGA-II.

ID	Approach A	Approach B	Approach C	Our Approach
101	68.485 (4)	20.396 (3)	18.219 (2)	18.130 (1)
103	33.020 (4)	20.058 (3)	18.356 (2)	18.135 (1)
104	35.142 (4)	20.724 (3)	22.952 (2)	20.688 (1)
201	224.281 (4)	174.471 (3)	158.794 (2)	152.580 (1)
203	219.355 (4)	166.168 (3)	164.440 (2)	158.940 (1)
204	135.096 (4)	108.664 (2)	119.379 (3)	105.807 (1)
205	210.275 (4)	164.365 (3)	130.698 (2)	127.554 (1)
206	206.870 (4)	161.680 (3)	132.568 (2)	126.956 (1)
221	185.348 (4)	145.674 (3)	124.082 (2)	108.327 (1)
222	187.774 (4)	155.582 (3)	129.341 (2)	113.199 (1)
223	222.302 (4)	185.502 (3)	142.441 (2)	139.903 (1)
224	203.368 (4)	160.380 (1)	173.225 (3)	171.045 (2)
225	203.790 (4)	162.194 (1)	173.223 (3)	171.847 (2)
228	174.386 (4)	154.386 (2)	184.379 (3)	134.376 (1)
230	176.487 (4)	165.909 (2)	188.328 (3)	152.754 (1)
231	184.253 (4)	154.878 (3)	130.659 (2)	121.127 (1)
301	84.511 (4)	84.484 (3)	79.187 (2)	74.429 (1)
302	229.535 (4)	193.194 (3)	152.161 (2)	148.311 (1)
304	219.960 (4)	167.646 (3)	155.451 (2)	152.017 (1)
Average	168.644 (4)	135.071 (2.63)	126.204 (2.26)	116.638 (1.11)

TABLE 6. Holm's test on the memory consumption per generation. Approach A, B and C respectively refer to the approach using NSGA-II, prescreening approach and NSGA-II and GRFM assisted NSGA-II.

i	Approach	z value	unadjusted $p-$value	$\frac{\alpha}{k-i}, \alpha = 0.05$
3	approach C	2.7456	0.0060	0.05
2	approach B	3.6290	0.0003	0.025
1	approach A	6.8998	$5.2076 \times e^{-12}$	0.0166

As can be seen from Table 5, in the Friedmans test, the computed $\mathcal{X}_r^2 = 48.52 > \mathcal{X}_{0.05}^2 = 7.815$ and in the Holm's test, our proposal statistically outperforms other approaches on memory consumption at 0.05 significance level.

Finally, we carry out the statistical comparison on the alignment's quality in terms of f-measure among Genetic Algorithm (GA) based [13], Memetic Algorithm (MA) based [14], Particle Swarm Optimization (PSO) based [4] ontology meta-matching approaches and our approach. Since all these approaches only optimize one objective function, in order to compare the alignment's quality, we select the solution with the highest harmonic mean of two objectives from the Pareto front as the output solution.

TABLE 7. Friedman's test on the alignment's quality obtained by three single objective EA based ontology meta-matching approaches and our approach. Each value represents the f-measure, and the number in round parentheses is the corresponding computed rank.

ID	GA	MA	PSO	Our Approach
101	1.00 (2.5)	1.00 (2.5)	1.00 (2.5)	1.00 (2.5)
103	0.99 (4)	1.00 (2)	1.00 (2)	1.00 (2)
104	0.99 (4)	1.00 (2)	1.00 (2)	1.00 (2)
201	0.50 (3)	0.62 (2)	0.42 (4)	0.94 (1)
203	0.97 (3)	0.96 (4)	1.00 (1)	0.99 (2)
204	0.94 (4)	0.97 (3)	0.98 (1.5)	0.98 (1.5)
205	0.83 (2)	0.79 (3)	0.73 (4)	0.93 (1)
206	0.84 (4)	0.88 (2)	0.85 (3)	0.92 (1)
221	0.99 (3.5)	0.99 (3.5)	1.00 (1.5)	1.00 (1.5)
222	0.99 (3)	0.99 (3)	0.99 (3)	1.00 (1)
223	0.99 (2.5)	0.99 (2.5)	0.99 (2.5)	0.99 (2.5)
224	1.00 (2.5)	1.00 (2.5)	1.00 (2.5)	1.00 (2.5)
225	0.99 (4)	1.00 (2)	1.00 (2)	1.00 (2)
228	0.99 (3)	0.99 (3)	0.99 (3)	1.00 (1)
230	0.93 (3.5)	0.93 (3.5)	0.98 (2)	1.00 (1)
231	0.99 (3)	0.99 (3)	0.99 (3)	1.00 (1)
301	0.70 (2.5)	0.70 (2.5)	0.64 (4)	0.81 (1)
302	0.61 (3)	0.63 (2)	0.04 (4)	0.85 (1)
304	0.83 (3)	0.87 (2)	0.72 (4)	0.93 (1)
Average	0.90 (3.16)	0.91 (2.63)	0.86 (2.71)	0.97 (1.5)

TABLE 8. Holm's test on the alignment's quality obtained by three single objective EA based ontology meta-matching approaches and our approach.

i	Approach	z value	unadjusted $p-$value	$\frac{\alpha}{k-i}, \alpha = 0.05$
3	MA	2.6978	0.0070	0.05
2	PSO	2.8888	0.0039	0.025
1	GA	3.9631	$7.3983 \times e^{-5}$	0.0166

As can be seen from Table 7, in the Friedmans test, the computed $\mathcal{X}_r^2 = 17.0612 > \mathcal{X}_{0.05}^2 = 7.815$. In the Holm's test, as shown in Table 8, our approach statistically outperforms other single objective EA based ontology meta-matching approaches on the alignment's quality at 0.05 significance level.

6. **Conclusion.** Ontology meta-matching is a challenge in ontology matching domain. In this paper, we propose an improved NSGA-II based ontology meta-matching technology, which can automatically filter the less promising alignments and reduce the exact evaluations during the evolving process of NSGA-II. The experimental results show that utilization of alignment prescreening approach and GRFM is able to significantly reduce the runtime and memory consumption without sacrificing the quality of the ontology alignment, and our results are also better than other EA based ontology matching approaches.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No. 61503082), Natural Science Foundation of Fujian Province (No.

2016J05145), Scientific Research Startup Foundation of Fujian University of Technology (No. GY-Z15007), Fujian Province outstanding Young Scientific Researcher Training Project (No. GY-Z160149) and China Scholarship Council.

REFERENCES

[1] K. Deb, S. Agrawal, A. Pratap, A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II, *International Conference on Parallel Problem Solving From Nature*, Paris, France, pp.849–858, 2000.

[2] G. Acampora, U. Kaymak, V. Loia, A. Vitiello, Applying NSGA-II for solving the Ontology Alignment Problem, *IEEE International Conference on Systems, man, and Cybernetics*, Manchester, United Kingdom, pp.1098–1103, 2013.

[3] G. Acampora, V. Loia, A. Vitiello, Enhancing ontology alignment through a memetic aggregation of similarity measures, *Information Sciences*, vol.250, pp.1–20, 2013.

[4] J. Bock, J. Hettenhausen, Discrete particle swarm optimisation for ontology alignment, *Information Sciences*, vol.192, pp.152–173, 2012.

[5] X. Xue, Y. Wang, Optimizing Ontology Alignments through a Memetic Algorithm Using both MatchFmeasure and Unanimous Improvement Ratio, *Artificial Intelligence*, vol.223, pp.65–81, 2015.

[6] A. Ratle, Accelerating the convergence of evolutionary algorithms by fitness landscape approximations, *Parallel Problem Solving by Nature*, Amsterdam, The Netherlands, pp.87–96, 1998.

[7] M. Fleischer, The measure of Pareto optima: Applications in multi-objective metaheuristics, *Evolutionary Multiobjective Optimisation*, Faro, Portugal, pp.519–533, 2003.

[8] F. Hamdi, B. Safar, C. Reynaud, H. Zargayouna, Alignment-based Partitioning of Large-scale Ontologies, vol. 292, Springer Berlin Heidelberg, Berlin, Germany, pp.251–269, 2010.

[9] OAEI, Ontology Alignment Evaluation Initiative (OAEI), http://oaei.ontologymatching.org/2016, 2016.

[10] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Physics Doklady*, vol.10, no.8, pp.707–710, 1966.

[11] G. A. Miller, WordNet: A lexical database for English, *Communications of the ACM*, vol.38, no.11, pp.39–41, 1995.

[12] S. Melnik, H. Garcia-Molina, E. Rahm, Similarity flooding: A versatile graph matching algorithm and its application to schema matching, *The 18th International Conference on Data Engineering*, San Jose, CA, pp.117–128, 2002.

[13] J. Martinez-Gil, E. Alba, J. F. A. Montes, Optimizing ontology alignments by using genetic algorithms, *Proceedings of the First International Conference on Nature Inspired Reasoning for the Semantic Web*, Vol.419, pp.1–15, 2008.

[14] G. Acampora, V. Loia, S. Salerno, A. Vitiello, A hybrid evolutionary approach for solving the ontology alignment problem, *International Journal of Intelligent Systems*, vol.27, no.3, pp.189–216, 2012.