Journal of Information Hiding and Multimedia Signal Processing ©2019 ISSN 2073-4212
Ubiquitous International Volume **10**, Number **1**, January **2019**

Matching Biomedical Ontologies Through Compact Hybrid Evolutionary Algorithm

Xingsi Xue*

College of Information Science and Engineering
Intelligent Information Processing Research Center
Fujian Provincial Key Laboratory of Big Data Mining and Applications
Fujian Key Lab for Automotive Electronics and Electric Drive
Fujian University of Technology
No.3 Xueyuan Road, University Town, Minhou, Fuzhou City, Fujian Province, 350118, China
*Corresponding author: jack8375@gmail.com

Junfeng Chen

College of IOT Engineering
Hohai University
No.200 North Jinling Road, Changzhou, Jiangsu, 213022, China
chen-1997@163.com

Dongxu Chen

Fujian Medical University Union Hospital
No.29, Xinquan Road, Fuzhou, Fujian, 350001, China
starryflyer2008@hotmail.com

Received April 2018; revised August 2018

Abstract. *Over the recent years, the biomedical ontologies have been developed to support a variety of applications. However, the subjectivity of different biomedical ontology designers leads to the generation of heterogeneous biomedical ontologies. In order to support the cooperations among the heterogeneous biomedical ontologies, it's necessary to identify the correspondences out of semantically identical entities of them, so-called biomedical ontology matching. In this paper, we propose a compact hybrid Evolutionary Algorithm (chEA), which utilizes a probabilistic representation of the population to perform the optimization process, and introduces a local search strategy to improve the efficiency. The Anatomy track and Large Biomed track, which are provided by the Ontology Alignment Evaluation Initiative (OAEI 2017), are utilized to test the performance of chEA. The experimental results show the effectiveness of our approach.*
Keywords: Biomedical ontology matching, Compact hybrid Evolutionary Algorithm, OAEI

1. **Introduction.** Over the recent years, the biomedical ontologies have been developed to support a variety of applications, such as the annotation of medical records [1], standardization of medical data formats [2], medical knowledge representation and sharing [3] and medical decision-making [4]. These vast usages of ontologies in the biomedical field have compelled researchers to develop more biomedical ontologies. However, the subjectivity of different biomedical ontology designers leads to the generation of heterogeneous biomedical ontologies. For example, the National Cancer Institute's thesaurus and ontology (NCI) [5] defines the entity "Myocardium", whereas the Foundation Model

of Anatomy (FMA) [6] uses the entity "Cardiac Muscle Tissue" to describe the muscles that surround and power the human heart. In order to support the semantic among the heterogeneous biomedical ontologies, it's necessary to identify the correspondences out of semantically identical entities of them, so-called biomedical ontology matching.

We can describe an biomedical ontology through its architecture graph (the nodes represent concepts and instances, while the edges stand for the relationship between them), and the problem of biomedical ontology matching is the determination of the largest isomorphic sub graph out of the two architecture graphs of two ontologies to be matched. Since the problem of modeling ontology matching is a complex (nonlinear with many local optimal solutions) and time-consuming task (large scale), particularly when the number of ontology entities is significantly large, Evolutionary Algorithm (EA) could be an efficient approach to address this problem [7]. However, existing EA based ontology matching techniques fail to match biomedical ontologies due to huge memory consumption and long runtime. Therefore, besides the quality of alignments, main memory consumption and runtime needed by the ontology matcher is of prime importance when matching the biomedical ontologies. In this paper, we propose a compact hybrid Evolutionary Algorithm (chEA), which utilizes a probabilistic representation of the population to perform the optimization process, and introduces a local search strategy to improve the efficiency. The contributions of this paper are listed as follows:

- An optimal model is constructed for biomedical ontology matching problem,
- A biomedical concept similarity measure is presented to calculate the similarity value of two biomedical concepts,
- A compact hybrid Evolutionary Algorithm is proposed to efficiently solve the biomedical ontology matching problem, and determine the high-quality biomedical ontology alignment.

The rest of the paper is organized as follows: Section 2 describes definition of ontology, ontology alignment, and the concept similarity measure; Section 3 presents the optimal model of biomedical ontology matching problem and the details of the compact hybrid Evolutionary Algorithm; Section 4 gives the experimental results and relevant analysis; finally, Section 5 draws the conclusions.

2. **Preliminaries.**

2.1. **Ontology, Ontology Alignment and Ontology Matching Process.**

In this work, an ontology is defined as a 3-tuples (C, P, A), where C is the set of classes, i.e. the set of concepts that populate the domain of interest; P is the set of properties, i.e. the set of relations existing between the concepts of domain; A is a set of axioms, i.e. the statements that say what is true about the modeled domain, such as subclass, equivalent classes and disjoint classes.

An alignment A between two ontologies is defined as a set of correspondences, and each correspondence is a 4-tuples $(e, e', n, =)$, where e and e' are the entities of two ontology respectively, $n \in [0, 1]$ is a confidence value holding for the correspondence between the entities e and e', $=$ means the equivalence relationship between two entities.

The ontology matching process can be defined as a function θ which, from a pair of ontologies O and O' to align, an input alignment A_I, a set of parameters p, a set of resources r, returns a new alignment A_N between these ontologies: $A_N = \theta(O, O', A_I, p, r)$ [8]. The output alignment A_N is a set of semantic matchings, and each one of them is used for linking an entity belonging to the first ontology with a similar entity belonging to the second ontology.

2.2. **Biomedical Concept Similarity Measure.** Biomedical concept similarity measure is the foundation of biomedical ontology matching [9]. In this work, we utilize an asymmetrical concept similarity measure to calculate similarity value between two biomedical concepts. First, for each biomedical concept, we construct a profile for it by collecting the label, comment, and property information such as label, domain and range, from itself and all its direct descendants. Then, the similarity of two biomedical concepts c_1 and c_2 is measured based on the similarity of their profiles p_1 and p_2, which can be calculated by the following two asymmetrical measures:

$$sim_1(p_1, p_2) = \frac{|p_1 \bigcap p_2|}{|p_1|} \tag{1}$$

$$sim_2(p_1, p_2) = \frac{|p_1 \bigcap p_2|}{|p_2|} \tag{2}$$

where $|p_1|$ and $|p_2|$ are the cardinality of the profile p_1 and p_2 respectively, $|p_1 \bigcap p_2|$ is the number of identical elements in p_1 and p_2, and the similarity of e_1 and e_2 is calculated through the following formulas:

$$sim(e_1, e_2) = \begin{cases} \frac{sim_1(p_1,p_2)+sim_2(p_1,p_2)}{2}, & \text{if } |sim_1(p_1, p_2) - sim_2(p_1, p_2)| \leq \delta \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

In this work, $\delta = 0.06$ is the threshold to measure the extent of the semantic equivalence between $sim_1(p_1, p_2)$ and $sim_2(p_1, p_2)$. When the similarity value between two profile elements is above the threshold, they're identified as semantically similar. Moreover, the similarity value of two profile elements is calculated by N-gram distance [10], which is the most performing string-based similarity measure for the biological ontology matching problem, and a linguistic measure, which calculate a synonymy-based distance through Unified Medical Language System (UMLS) [11]. Given two words w_1 and w_2, their similarity $sim(w_1, w_2)$ is calculated according to the following formula:

$$sim(w_1, w_2) = \begin{cases} 1, & \text{if two words are synonymous} \\ N - gram(w_1, w_2), & \text{otherwise} \end{cases} \tag{4}$$

3. Compact Hybrid Evolutionary Algorithm.

3.1. **The Optimal Model for Biomedical Ontology Matching Problem.** 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, we utilize the following metric to measure the quality of a biomedical ontology alignment:

$$I(A) = 2 \times \frac{\phi(A) \times \frac{\sum_{i=1}^{|A|} \delta_i}{|A|}}{\phi(A) + \frac{\sum_{i=1}^{|A|} \delta_i}{|A|}} \tag{5}$$

where $|A|$ is the number of correspondences in A, ϕ is a function of normalization in [0,1], δ_i is the similarity value of the ith correspondence in A, and α is a parameter used to tradeoff the ontology alignments characterized by high recall (with the decreasing of α) or high precision (with the increase of α). On this basis, the optimal model of biomedical ontology matching problem is defined as follows:

$$\begin{cases} max \quad I(X) \\ s.t. \quad X = (x_1, x_2, ..., x_{|O_1|}, x_{|O_1|+1})^T \\ \qquad x_i = 1, 2, \cdots, |O_2| \\ \qquad x_{|O_1|+1} \in [0, 1] \end{cases} \tag{6}$$

where the decision variable X represents an alignment between ontologies O_1 and O_2, x_i represent the ith correspondence between ith concept in O_1 and x_ith concept in O_2, $|O_1|$ and $|O_2|$ are the cardinalities of the concept set in O_1 and O_2 respectively, and $x_{|O_1|+1} \in [0, 1]$ is the threshold to filter the final alignment.

In the next, we utilize chEA to solve the biomedical onology matching problem, which can save memory consumption and runtime without sacrificing alignment's quality. In the following, we present two main components of chEA, i.e. chromosome encoding mechanism and local search strategy, as well as the algorithm's pseudo-code.

3.2. **Chromosome Encoding Mechanism.** In this work, we utilize the Probability Vector (PV), a binary vector with each gene's value in [0,1], to characterize the entire population in population-based EA. The information in PV can be divided into two parts: one stands for the correspondences in the alignment, and the other for a threshold. We represent both the correspondences and threshold through the binary coding mechanism in the field of computer science according to the number of target activities and the numerical accuracy of threshold. When decoding, we calculate the corresponding decimal numbers. In the first part, the numbers obtained represent the indexes of the target activities, and in particular, the value 0 means corresponding source activity does not map to any target activities. While in the second part, the decimal number should be plus the numerical accuracy. This is because given the number accuracy acc, the threshold value will be expressed by an integer in $[0, \frac{1}{acc}]$ with a binary code.

3.3. **Local Search Strategy.** Local search strategy dedicates to generate various individuals to search the vicinity range of the elite solution ind_{elite}. In this work, by referring to the work in [13], a $C \times D$ matrix M is constructed and we use it to generate neighbour individuals of the ind_{elite}, where C is the scale of neighbour population and D is the number of dimensions. With respect to C, a larger value of it may perform better exploitation and especially for the multi-modal problem, but increase the computation complexity. Here, we empirically set $C = 5$. For the sake of clarity, given a permutation possibility p_p, the pseudo-code of generating M is shown as follows:

//Initialize M
1. for(int $i = 0; i < C; i++$)
2. for(int $j = 0; j < D; j++$)
3. $M_{ij} = 0$;
4. end for
5. end for

//Permutate M
6. for(int $i = 0; i < C; i++$)
7. generate $j = round(rand(0, D))$;
8. while $(rand(0, 1) < p_p)$
9. $M_{ij} = 1$;
10. $j = j + 1$;
11. if $(j == D)$
12. $j = 0$;
13. end if
14. end while
15. end for

By flipping the value in M, i.e. converting the zero elements in M into one and non-zero elements into zero, we can obtain $\overline{M}$. On this basis, the neighbour population of the ind_{elite} can generate through the following formula:

$$\overrightarrow{ind_{neighbor}} = M \otimes \overrightarrow{ind_{elite}} + \overline{M} \otimes \overrightarrow{X} \tag{7}$$

where $\overrightarrow{ind_{elite}} = \begin{bmatrix} ind_{elite} \\ ind_{elite} \\ \cdots \\ ind_{elite} \end{bmatrix}_{C \times D}$, $\overrightarrow{X} = \begin{bmatrix} ind_1 \\ ind_2 \\ \cdots \\ ind_C \end{bmatrix}_{C \times D}$ and ind_i, $i = 1, 2, \cdots, C$, is generated by PV, and the operator $\otimes$ is multiplication of corresponding matrix elements.

Finally, we obtain the best individual in the neighbor population $ind_{localBest} = opti\{\overrightarrow{ind_{neighbor}}(i)\}$, $i = 1, 2, \cdots, C$.

3.4. The Pseudo-code of Compact Hybrid Evolutionary Algorithm. Input:

- num: the length of chromosome;
- $maxGen$: maximum number of generations;
- cro: exponential crossover probability;
- sl: step length when updating PV.

Output: ind_{elite}: the solution with best fitness value.

Step 1) Initialization:

1. generation=0;
2. for($i = 0$; $i < num$; i++)
3. $PV[i] = 0.5$;
4. end for
5. generate an individual ind_{elite} by means of PV;

Step 2) Update PV:

6. generate ind_s by means of PV;
7. $ind_{new} = localSearch(ind_{elite}, ind_s)$;
8. $[winner, loser] = compete(ind_{elite}, ind_{new})$;
9. if ($winner == ind_{new}$)
10. $ind_{elite} = ind_{new}$;
11. end if
12. for($i = 0$; $i < num$; i++)
13. if (winner[i]==1)
14. $PV[i] = PV[i] + sl$;
15. else
16. $PV[i] = PV[i] - sl$;
17. end if
18. end for

Step 3) Stopping Criteria:

19. if ($maxGen$ is reached or each bit of PV is either 1 or 0)
20. stop and output ind_{elite};
21. else

TABLE 1. Comparison of our approach with the participants in OAEI 2017 on Anatomy track

Systems	R	P	F	*runtime*
AML	0.93	0.95	0.94	47
YAM-BIO	0.92	0.94	0.93	70
POMap	0.90	0.94	0.93	808
LogMapBio	0.89	0.88	0.89	820
XMap	0.86	0.92	0.89	37
LogMap	0.84	0.91	0.88	22
KEPLER	0.74	0.95	0.83	234
LogMapLite	0.72	0.96	0.82	19
SANOM	0.77	0.89	0.82	295
Wiki2	0.73	0.88	0.80	2204
ALIN	0.33	0.99	0.50	836
QUATRE	0.93	0.96	0.94	293
chEA	0.93	0.98	0.95	34

22. generation=generation+1;
23. go to Step 2);
24. end if

4. **Experimental Results and Analysis.** In order to study the effectiveness of our approach, we exploit the Anatomy [1] and Large Biomed [2] track, which are provided by the Ontology Alignment Evaluation Initiative (OAEI 2017) [3]. The Anatomy track includes two ontologies (1 task), i.e. the Adult Mouse Anatomy (AMA) ontology (2,744 classes) and a part of NCI describing the human anatomy (3,304 classes). Large Biomed track (3 tasks) aims at finding alignments between FMA, SNOMED CT, and NCI, which respectively contains 78,989, 122,464 and 66,724 classes. Particularly, Large Biomedic track is split into three matching problems: FMA-NCI, FMA-SNOMED and SNOMED-NCI, and each matching problem in these tasks involving different fragments of the input ontologies.

In this work, step length $sl = 0.01$, scale of neighbour population $C = 10$, local search's permutation possibility $pp = 0.6$, and the algorithm terminates when it runs up to 3000 generations. These parameters are set in an empirical way to achieve the highest average alignment quality on all test cases of exploited datasets. In order to compare the quality of our proposal with other process model matchers, we evaluate the obtained alignments with the traditional recall, precision and f-measure [12], and the experimental results in the tables are the average values over thirty independent runs.

4.1. **Results and analysis.** In order to compare the quality of our proposal with the participants of OAEI 2017 [4] and a state-of-the-art EA, i.e. QUasi-Affine TRansformation Evolutionary (QUATRE) algorithm [14]. We evaluate the obtained alignments with traditional recall, precision and f-measure, and our approach's results in Table 1 and Table 2 are the mean values in thirty time independent executions. The symbols P, R and F in tables stand for precision, recall and f-measure, respectively.

[1] http://oaei.ontologymatching.org/2017/anatomy/index.html
[2] http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/2017/
[3] http://oaei.ontologymatching.org/2017
[4] http://oaei.ontologymatching.org/2017/results/index.html

TABLE 2. Comparison of our approach with the participants in OAEI 2017 on Large Biomed track

Systems	R	P	F	Runtime(s)
Task1: whole FMA and NCI ontologies				
XMap*	0.85	0.88	0.87	130
AML	0.87	0.84	0.86	77
YAM-BIO	0.89	0.82	0.85	279
LogMap	0.81	0.86	0.83	92
LogMapBio	0.83	0.82	0.83	1552
LogMapLite	0.82	0.67	0.74	10
Tool1	0.74	0.69	0.71	1650
QUATRE	0.85	0.91	0.87	393
chEA	0.86	0.91	0.88	68
Task2: whole FMA and SNOMED ontologies				
XMap*	0.84	0.77	0.81	625
YAM-BIO	0.73	0.89	0.80	468
AML	0.69	0.88	0.77	177
LogMap	0.65	0.84	0.73	477
LogMapBio	0.65	0.81	0.72	2951
LogMapLite	0.21	0.85	0.34	18
Tool1	0.13	0.87	0.23	2140
QUATRE	0.83	0.87	0.84	862
chEA	0.83	0.90	0.86	139
Task3: whole SNOMED and NCI ontologies				
AML	0.67	0.90	0.77	312
YAM-BIO	0.70	0.83	0.76	490
LogMapBio	0.64	0.84	0.73	4728
LogMap	0.60	0.87	0.71	652
LogMapLite	0.57	0.80	0.66	22
XMap*	0.55	0.82	0.66	563
Tool1	0.22	0.81	0.34	1150
QUATRE	0.72	0.88	0.79	862
chEA	0.75	0.92	0.82	286

As can be seen from Table 1, our approach's f-measure is the best among all the participants in OAEI 2017, and the runtime taken by our approach is is in the third place. In Table 2, in terms of f-measure, our approach's results are ranked the first in task1, task2 and task3. With respect to the running time, in task1, task2 and task3, our approach is in the second place. In two tracks, our approach outperforms AML, which is the top ontology matcher and developed primarily for the biomedical ontology matching, in all tasks in terms of f-measure and runtime. Comparing with QUATRE in all testing cases, although it can obtain quite similar results, the runtime needed is much longer than chEA. To conclude, chEA can efficiently match the biomedical ontologies.

5. **Conclusion.** In this work, in order to overcome the drawbacks in traditional EA based ontology matching techniques, we propose a compact hybrid Evolutionary Algorithm to

efficiently match the biomedical ontologies. In particular, we utilize one PV to characterize the entire population in population-based EA, which can significantly save the memory consumption; and then we introduce the local search strategy into the evolving process to reduce the runtime. Moreover, we construct an optimal model for biomedical ontology matching problem, and present a biomedical concept similarity measure to ensure the quality of ontology alignment. In the experiment, OAEI 2017's Anatomy track and Large Biomed track are utilized to test our approach's performance, and the results show that our approach can efficiently determine the biomedical ontology alignments with high quality.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (Nos. 61503082 and 61403121), Natural Science Foundation of Fujian Province (No. 2016J05145), Fundamental Research Funds for the Central Universities (No. 2015B20214), Scientific Research Foundation of Fujian University of Technology (Nos. GY-Z15007 and GY-Z17162) and Fujian Province Outstanding Young Scientific Researcher Training Project (No. GY-Z160149).

REFERENCES

[1] H. López-Fernández, M. Reboiro-Jato, D. Glez-Peña, F. Aparicio, D. Gachet, M. Buenaga and F. Fdez-Riverola, BioAnnote: A software platform for annotating biomedical documents with application in medical learning environments, *Computer methods and programs in biomedicine*, vol.111, no.1, pp.139-147, 2013.

[2] J. J. Cimino and X. Zhu, The practical impact of ontologies on biomedical informatics, *Year book of medical informatics*, vol.2006, pp.124-135, 2006.

[3] D. Isern, D. Sánchez, A. Moreno, The practical impact of ontologies on biomedical informatics, *Ontology-driven execution of clinical guidelines*, vol.107, no.2, pp.122-139, 2012.

[4] P. De Potter, H. Cools, K. Depraetere, G. Mels, P. Debevere, J. De Roo, C. Huszka, D. Colaert, E. Mannens and R. Van de Walle, Semantic patient information aggregation and medicinal decision support, *Computer methods and programs in biomedicine*, vol.108, no.2, pp.724-735, 2012.

[5] J. Golbeck, G. Fragoso, F. Hartel, J. Hendler, J. Oberthaler and B. Parsia, The National Cancer Institute's thesaurus and ontology, *Web Semantics: Science, Services and Agents on the World Wide Web*, vol.1, no.1, pp.1-5, 2012.

[6] C. Rosse and J. L. Mejino Jr, A reference ontology for biomedical informatics: the Foundational Model of Anatomy, *Journal of biomedical informatics*, vol.36, no.6, pp.478-500, 2003.

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

[8] J. Euzenat and P. Valtchev, Similarity-based ontology alignment in OWL-Lite, *16th European Conference on Artificial Intelligence Proceeding*, pp.333-337, 2004.

[9] A. Maedche and S. Staab, Measuring Similarity between Ontologies, *Proceedings of the 14th International Conference on Knowledge Engineering and Knowledge Management*, pp.251-263, 2002.

[10] G. Kondrak, N-gram similarity and distance, *International symposium on string processing and information retrieval*, pp.115-126, 2005.

[11] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, *Nucleic acids research*, vol.32, pp.267-270, 2004.

[12] C. J. Van Rijsberge, Information Retrieval, University of Glasgow, Butterworth, London, 1975.

[13] Z. Meng and J. S. Pan, Monkey king evolution: a new memetic evolutionary algorithm and its application in vehicle fuel consumption optimization, *Knowledge-Based Systems*, vol.97, pp.144-157, 2016.

[14] Z. Meng, J. S. Pan J S and H. Xu, QUasi-Affine TRansformation Evolutionary (QUATRE) algorithm: a cooperative swarm based algorithm for global optimization, *Knowledge-Based Systems*, vol.109, pp.104-121, 2016.