Matching biomedical ontologies through compact differential evolution algorithm

Xingsi Xue\textsuperscript{a,b,c} and Junfeng Chen\textsuperscript{d}

\textsuperscript{a}College of Information Science and Engineering, Fujian University of Technology, Fuzhou, People’s Republic of China; \textsuperscript{b}Intelligent Information Processing Research Center, Fujian University of Technology, Fuzhou, People’s Republic of China; \textsuperscript{c}Fujian Key Lab for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou, People’s Republic of China; \textsuperscript{d}College of IOT Engineering, Hohai University, Changzhou, People’s Republic of China

ABSTRACT

Although biomedical ontologies have been widely used in the life science domain, the heterogeneous problem among biomedical ontologies hampers their inter-operability. Thus, the establishment of meaningful links between heterogeneous biomedical ontologies, so-called biomedical ontology matching, is critical to the success of biomedical ontology engineering. To determine the biomedical ontology alignment with high quality, in this work, a Hybrid Compact Differential Evolution (HCDE) algorithm-based biomedical ontology matching technique is proposed. In particular, we propose a similarity metric on biomedical concepts, construct an optimal model for the biomedical ontology matching problem, and introduce a binomial crossover into CDE’s evolving the process to enhance its performance. The experiments are carried out on the Disease and Phenotype track and Biodiversity and Ecology track from the Ontology Alignment Evaluation Initiative (OAEI 2018). The experimental results show that HCDE can significantly improve the CDE in terms of the alignment’s quality, and the alignments obtained by HCDE are also better than OAEI 2018’s participants.

1. Introduction

In recent years, various biomedical ontologies, such as FMA (Detwiler, Mejino, & Brinkley, 2016) and SNOMED-CT (Filice & Kahn, 2019), have been widely used in the life science domain (Faria et al., 2018). However, existing biomedical ontologies that cover overlapping domains are mostly developed independently, and the different ways of defining the same biomedical concept yield heterogeneous problems among biomedical ontologies, which hampers their inter-operability. Thus, the establishment of meaningful links between biomedical concepts, so-called biomedical ontology matching, is critical to the success of biomedical ontology engineering (Oliveira & Pesquita, 2018).

Usually, a biomedical ontology possesses tens of thousands of biomedical concepts which are semantically ambiguous, and these biomedical concepts own complex relationships between each other. Thus, it is a big challenge to match them effectively. Being inspired by the success of Evolutionary Algorithm (Acampora, Loia, & Vitiello, 2013; Ginsca & Iftene, 2010; Martinez-Gil & Montes, 2011; Naya, Romero, & Loureiro, 2010; Wang, Ding, & Jiang, 2006; Xue & Pan, 2018; Xue & Wang, 2015) in ontology matching domain, in this work, we propose to use Differential Evolutionary algorithm (DE) (Storn & Price, 1997) to determine the biomedical ontology alignment with high quality. Since matching biomedical ontologies is a large-scale problem, we utilize a compact DE (CDE) (Mininno, Neri, Cupertino, & Naso, 2011) to avoid the memory overflow and introduce a local search strategy into CDE’s evolving process to improve its converging speed. In particular, our contributions made in this paper are as follows:

• a similarity metric on biomedical concept pair is proposed to calculate the similarity value of two biomedical concepts,
• an optimal model for biomedical ontology matching is constructed,
• a problem-specific Hybrid CDE (HCDE) is presented to effectively solve the biomedical ontology matching problem.

The rest of the paper is organized as follows: Section 2 presents the biomedical ontology matching problem and biomedical concept similarity metric; Section 3 presents in detail the HCDE-based biomedical ontology matching technique; Section 4 shows the experimental results; and finally, Section 5 draws the conclusion.
2. Preliminaries

2.1. Biomedical ontology matching

Biomedical ontology matching dedicates to determine an ontology alignment which consists of biomedical concept correspondences. Usually, some external resources, such as the electronic dictionaries and knowledge bases, are required in matching process. Since the quality of biomedical ontology alignment can be measured by the number of correspondences found and the mean similarity value of the biomedical concept correspondences, in this work, we utilize the following equation to measure a biomedical alignment’s quality:

\[
f(A) = \frac{2 \times |A| \times \sum \text{simValue}_i}{(|C_1| \times |C_2|) \times (|A| + \sum \text{simValue}_i)} \in [0, 1]
\]  

(1)

where \(|C_1|\), \(|C_2|\), and \(|A|\) are, respectively, the cardinalities of two biomedical concepts sets \(C_1\) and \(C_2\), and an alignment \(A\) between them, \(\text{simValue}_i\) is the similarity value of the \(i\)th correspondence of \(A\).

Further, an optimal model for the biomedical ontology matching problem is constructed as follows:

\[
\text{max } F(X) \quad \text{s.t. } X = (x_1, x_2, \ldots, x_{|C_1|})^T
\]  

(2)

\[x_i \in \{0, 1, 2, \ldots, |C_2|\}, \quad i = 1, 2, \ldots, |C_1|\]

where \(|C_1|\) and \(|C_2|\), respectively, represent the cardinalities of two biomedical concepts sets \(C_1\) and \(C_2\), \(x_i, i = 1, 2, \ldots, |C_1|\) represents the \(i\)th pair of correspondence, and in particular, \(x_i = 0\) means the \(i\)th source concept is mapped to none. Supposing \(A\) is \(X\)’s corresponding alignment, the objective function \(F(X)\) is equal to \(f(A)\).

2.2. Similarity metric

Biomedical similarity metric is the foundation of the biomedical ontology matching technique (Cross, 2018). In this work, we utilize a profile-based similarity metric to measure to what extent two biomedical concepts are similar to each other. Given the concept hierarchies of two biomedical ontologies, for each biomedical concept, we construct a profile for it by collecting the name, property name, and method name from itself and all its direct ascendant and descendants. Then, the similarity value of two biomedical concepts \(c_1\) and \(c_2\) is calculated through their profiles \(p^1\) and \(p^2\):

\[
\sum_{j=1}^{|p_1|} \max_{i=1, \ldots, |p_2|} (\text{sim}'(p^1_i, p^2_j)) + \sum_{j=1}^{|p_2|} \max_{i=1, \ldots, |p_1|} (\text{sim}'(p^2_j, p^1_i))
\]

\[
2 \times \min(|p^1|, |p^2|)
\]  

(3)

where

- \(|p^1|\) and \(|p^2|\) are the cardinalities of \(p^1\) and \(p^2\), respectively,
- \(p^1_i\) and \(p^2_j\) are, respectively, the \(i\)th element of \(p^1\) and \(j\)th element of \(p^2\),
- \(\text{sim}'()\) computes the similarity value between \(p^1_i\) and \(p^2_j\)

3. Hybrid compact differential evolution algorithm

To save the memory consumption and improve the converging speed, in this work, we utilize the compact version of DE, i.e. CDE, and combine CDE (global search) with a binomial crossover (local search) to address the biomedical ontology matching problem. In the following, three kernel components of HCDE are presented, i.e. the encoding mechanism, mutation operator and the local search, and the pseudocode of HCDE is presented at last.

3.1. Encoding mechanism

In our proposal, a Probability Vector (PV) (Xue & Pan, 2018) is utilized to characterize the entire population, and each element inside stands for the probability that holds true for a biomedical concept correspondence. We utilize the grey encoding mechanism to encode each biomedical concept mapping. When decoding, the number obtained represents the index of a target biomedical concept, and in particular, the value 0 means a source concept is not mapped to any target concept.

3.2. Mutation operator

In HCDE, three solutions, namely \(\text{ind}_r\), \(\text{ind}_s\), and \(\text{ind}_l\), are sampled from the PV, and an offspring \(\text{ind}_{\text{off}}\) is generated as follows:

\[
\text{ind}_{\text{off}} = \text{ind}_r + F(\text{ind}_l - \text{ind}_s)
\]  

(4)

where \(F = 1.5\) is a scale factor that determines how far the generated offspring is from \(\text{ind}_l\). For the biomedical ontology matching problem, which is a discrete problem, we introduce the edit distance to measure two individuals’ distance. In the following, we present the equation about the calculation of two individuals \(\text{ind}_r\) and \(\text{ind}_s\)’s edit distance:

\[
\text{editDistance}(\text{ind}_r, \text{ind}_s) = \sum_{i=1}^{\text{ind}_r} |\text{ind}_{r,i} - \text{ind}_{s,i}|
\]  

(5)
where \( |ind_i| \) is the cardinality of \( ind_i, ind_{i,j} \) and \( ind_{i,j} \) are, respectively, the \( i \)th element of \( ind_i \) and \( ind_s \). Next, the offspring \( ind_{off} \) are generated by partly flipping the elements of \( ind_i \) and the number of altered elements is determined by a random number in \([0, 1]\) and \( editDistance(ind_i, ind_s) \). For the sake of clarity, the pseudocode of the mutation operator is given in Algorithm 1:

**Algorithm 1 Mutation Operator**

```plaintext
ind_{off} = ind_i.copy();
for i = 0; i < |ind_i|; i++ do
    if ind_i[i] then
        append i to an array index;
    end if
end for

totalNum = round(rand(0, 1) \times editDistance(ind_i, ind_s));
j = 0;
k = 0;
while j < totalNum do
    if rand(0, 1) > 0.5 then
        ind_{off}[index[k]] = (ind_{off}[index[k]] + 1) mod 2;
        remove the kth element from index;
        j++;
    end if
    k = (k + 1) mod index.length();
end while
```

### 3.3. Local search

To improve CDE’s converging speed, a local search strategy in introduced, which searches for the optimal solution in the neighbourhood of the elite solution. In particular, we execute the local search in each generation which is implemented with the binomial crossover. For the sake of clarity, the pseudocode of the binomial crossover is given in Algorithm 2.

**Algorithm 2 Binomial Crossover**

```plaintext
ind_{off} = elite;
i = round(elite.length() \times rand(0, 1));
ind_{off}[i] = ind_{off}[i];
while rand(0, 1) < crossover_probability do
    ind_{off}[i] = ind_{off}[i];
i = i + 1;
if i == n then
    i = 1;
end if
end while
```

The binary crossover randomly copies a sequential fragment of \( ind_{off} \)’s genes into the corresponding positions of \( ind_{off} \), to generate \( ind_{elite} \)’s neighbour solution. This procedure is similar to the two-point crossover where the first cut point is randomly selected from \([1, 2, \ldots, n]\) where \( n \) is the length of a solution, and the second point is determined such that \( L \) consecutive genes (counted in a circular manner) are taken from \( ind_{off} \). In this work, since \( ind_{off} \) and \( elite \) are both generated through \( PV \), they are similar in term of chromosome’s information, i.e. many gene bit values of them are the same. Therefore, even when \( crossover\_probability \) is large, the \( ind_{off} \) generated

**Algorithm 3 Hybrid Compact Differential Evolution Algorithm**

```plaintext
** Initialization **
generation = 0;
for i = 0; i < length; i++ do
    PV[i] = 0.5;
end for

generate an individual \( ind_{elite} \) through \( PV \);
while generation < maxGen do
    ** Mutation **
    generate three solutions \( ind_i, ind_s \) and \( ind_{i,j} \) through \( PV \);
    generate an offspring \( ind_{off} \) through Equation 4;
    ** Crossover **
    \( ind_{off} = ind_{off} \);
    for i = 0; i < length; i++ do
        if rand(0, 1) < crossover_probability then
            ind_{off}[i] = ind_{elite}[i]
        end if
    end for
    ** Local Search **
    \( ind_{new} = localSearch(ind_{elite}, ind_{off}) \);
    ** Elite Selection **
    [winner, loser] = compete(ind_{elite}, ind_{new});
    if winner == ind_{new} then
        ind_{elite} = ind_{new};
    end if
    ** PV Update **
    for i = 0; i < length; i++ do
        if winner[i] == 1 then
            PV[i] = PV[i] + step;
        else
            PV[i] = PV[i] - step;
        end if
    end for
    generation = generation + 1;
end while
output ind_{elite};
```
3.4. The pseudocode of hybrid compact differential evolution algorithm

Given the length of a solution (or PV) length, the maximum number of generations maxGen, the binomial crossover probability pc, the step length for updating PV step, the pseudo-code of HCDE is given in Algorithm 3.

4. Experiment

4.1. Experimental setup

In order to study the effectiveness of HCDE, we exploit the Disease and Phenotype track¹ and Biodiversity and Ecology track² from the Ontology Alignment Evaluation Initiative (OAEI 2018).³ The Disease and Phenotype track is composed of two tasks which, respectively, require matching Human Phenotype Ontology (HP)⁴ to Mammalian Phenotype Ontology (MP),⁵ and Human Disease Ontology (DOID)⁶ to the Orphanet Rare Disease Ontology (ORDO).⁷ The Biodiv track requires matching the Environment Ontology (ENVO)⁸ and the Semantic Web for Earth and Environment Technology Ontology (SWEET),⁹ and the Flora Phenotype Ontology (FLOPO)¹⁰ and the Plant Trait Ontology (PTO).¹¹ All these biomedical ontologies are particularly useful for biodiversity and ecology research and have been used in various projects, which are semantically rich and contain tens of thousands of classes.

In order to compare with CDE-based biomedical ontology matcher, which is the nonlocal search version of HCDE, and the participants of OAEI 2018,¹² we evaluate the obtained alignments with traditional recall, precision and f-measure (Van Rijsbergen, 1986). HCDE and CDE’s results in the tables are the mean values in 30 independent executions, and the symbols P, R and F in the tables stand for precision, recall and f-measure, respectively. HCDE uses the following parameters which represent a trade-off setting obtained in an empirical way to achieve the highest average alignment quality on all exploited dataset.

- Step length for updating PV = 0.1;
- Crossover probability = 0.6;
- Maximum generation = 3000.

| Systems | R   | P   | F   |
|---------|-----|-----|-----|
| LogMap  | 0.83| 0.87| 0.85|
| LogMapBio| 0.84| 0.86| 0.85|
| AML     | 0.80| 0.88| 0.84|
| LogMapLt| 0.60| 0.99| 0.75|
| POMAP++ | 0.57| 0.85| 0.68|
| Lily    | 0.64| 0.68| 0.66|
| XMap    | 0.31| 0.99| 0.47|
| DOME    | 0.30| 0.99| 0.47|
| CDE     | 0.72| 0.91| 0.80|
| HCDE    | 0.87| 0.93| 0.89|

4.2. Experimental results

As can be seen from Tables 1 and 2, HCDE’s f-measure is the best in all testing cases. In general, our precision value is high, which show the effectiveness of the proposed concept similarity metric. Comparing with CDE, which is the nonlocal search version of HCDE, the recall

| Systems | R   | P   | F   |
|---------|-----|-----|-----|
| AML     | 0.77| 0.93| 0.84|
| LogMap  | 0.79| 0.89| 0.84|
| LogMapBio| 0.79| 0.87| 0.83|
| LogMapLt| 0.61| 0.98| 0.75|
| XMap    | 0.54| 0.96| 0.70|
| KEPLER  | 0.57| 0.88| 0.69|
| Lily    | 0.78| 0.58| 0.67|
| AML     | 0.87| 0.51| 0.64|
| DOME    | 0.43| 0.99| 0.60|
| CDE     | 0.71| 0.88| 0.78|
| HCDE    | 0.81| 0.90| 0.85|
and precision are both improved significantly, which further show the effectiveness of the local search algorithm. To conclude, HCDE can effectively match the biomedical ontologies.

5. Conclusion

To efficiently match the biomedical ontologies, an HCDE-based technique is proposed to efficiently determine the identical biomedical concepts. In particular, HCDE combines CDE (global search) and binomial crossover (local search) to address the biomedical ontology matching problems. In addition, a novel biomedical ontology similarity metric is presented to distinguish the heterogeneous concepts and an optimal model of biomedical ontology matching problem is constructed. The experimental results show that the introduction of a local search strategy can significantly improve the converging speed of CDE, and the performance of HCDE outperforms the state-of-the-art ontology matchers of OAEI 2018.

Notes

1. http://oaei.ontologymatching.org/2018/phenotype/.
2. http://oaei.ontologymatching.org/2018/biodiv/index.html.
3. http://oaei.ontologymatching.org/2018/.
4. http://www.obofoundry.org/ontology/hp.html.
5. http://www.obofoundry.org/ontology/mp.html.
6. http://www.obofoundry.org/ontology/doid.html.
7. http://www.orphadata.org/cgi-bin/index.php#ontologies.
8. http://www.obofoundry.org/ontology/envo.html.
9. https://biportal.bioontology.org/ontologies/SWEET.
10. http://www.obofoundry.org/ontology/flopo.html.
11. http://www.obofoundry.org/ontology/to.html.
12. http://oaei.ontologymatching.org/2018/.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is supported by the National Natural Science Foundation of China (Nos. 61503082 and 61403121), the Natural Science Foundation of Fujian Province (No. 2016J05145), the Fundamental Research Funds for the Central Universities (No. 2019B22314), the Program for New Century Excellent Talents in Fujian Province University (No. GY-Z18155), the Program for Outstanding Young Scientific Researcher in Fujian Province University (No. GY-Z160149) and Scientific Research Foundation of Fujian University of Technology (No. GY-Z17162).

References

Acampora, G., Loia, V., & Vitiello, A. (2013). Enhancing ontology alignment through a memetic aggregation of similarity measures. *Information Sciences*, 250, 1–20.

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Suppl_1), D267–D270.

Cross, V., & Hu, X. (2018). Semantic similarity: A key to ontology alignment. In *Ontology matching: OM-2018: Proceedings of the ISWC workshop* (pp. 61–72). California, US.

Detwiler, L. T., Mejino, J. L., & Brinkley, J. F. (2016). From frames to OWL2: Converting the foundational model of anatomy. *Artificial Intelligence in Medicine*, 69, 12–21.

Faria, D., Pesquita, C., Mott, I., Martins, C., Couto, F. M., & Cruz, I. F. (2018). Tackling the challenges of matching biomedical ontologies. *Journal of Biomedical Semantics*, 9(1), 4.

Filice, R. W., & Kahn, C. E. (2019). Integrating an ontology of radiology differential diagnosis with ICD-10-cm, RadLex, and SNOMED CT. *Journal of Digital Imaging*, 32(2), 206–210.

Gincsa, A.-L., & Iftene, A. (2010). Using a genetic algorithm for optimizing the similarity aggregation step in the process of ontology alignment. In *9th Roedunet international conference* (pp. 118–122). Sibiu, Romania.

Martinez-Gil, J., & Montes, J. F. A. (2011). Evaluation of two heuristic approaches to solve the ontology meta-matching problem. *Knowledge and Information Systems*, 26(2), 225–247.

Mininno, E., Neri, F., Cupertino, F., & Naso, D. (2011). Compact differential evolution. *IEEE Transactions on Evolutionary Computation*, 15(1), 32–54.

Naya, J. M. V., Romero, M. M., & Loureiro, J. P. (2010). Improving ontology alignment through genetic algorithms (pp. 240–259). Hershey, PA: Information Science Reference.

Oliveira, D., & Pesquita, C. (2018). Improving the interoperability of biomedical ontologies with compound alignments. *Journal of Biomedical Semantics*, 9(1), 1.

Stoilos, G., Stamou, G., & Kollias, S. (2005). A string metric for ontology alignment. In International semantic web conference (pp. 624–637). Galway, Ireland.

Storn, R., & Price, K. (1997). Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4), 341–359.

Van Rijsbergen, C. J. (1986). A non-classical logic for information retrieval. *The Computer Journal*, 29(6), 481–485.

Wang, J., Ding, Z., & Jiang, C. (2006). Gaom: genetic algorithm based ontology matching. In *Proceedings of IEEE Asia-Pacific conference on services computing* (pp. 617–620). Guangzhou, China.

Xue, X., & Pan, J.-S. (2018). A compact co-evolutionary algorithm for sensor ontology meta-matching. *Knowledge and Information Systems*, 56(2), 335–353.

Xue, X., & Wang, Y. (2015). Optimizing ontology alignments through a memetic algorithm using both matchformance and unanimous improvement ratio. *Artificial Intelligence*, 223, 65–81.