CAAI Transactions on Intelligence Technology

Special Issue: Advances in Bio-inspired Heuristics for Computing

Journals
The Institution of Engineering and Technology

Using NSGA-III for optimising biomedical ontology alignment

ISSN 2468-2322
Received on 26th February 2019
Revised on 10th April 2019
Accepted on 22nd April 2019
doi: 10.1049/trit.2019.0014
www.ietdl.org

Xingsi Xue[1,2,3,4] ✉, *Jiawei Lu*[1,2], *Junfeng Chen*[5]

[1]*College of Information Science and Engineering, Fujian University of Technology, Fuzhou, Fujian, People's Republic of China*
[2]*Intelligent Information Processing Research Center, Fujian University of Technology, Fuzhou, Fujian, People's Republic of China*
[3]*Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fujian University of Technology, Fuzhou, Fujian, People's Republic of China*
[4]*Fujian Key Laboratory for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou, Fujian, People's Republic of China*
[5]*College of IOT Engineering, Hohai University, Changzhou, Jiangsu, People's Republic of China*
✉ *E-mail: jack8375@gmail.com*

Abstract: To support semantic inter-operability between the biomedical information systems, it is necessary to determine the correspondences between the heterogeneous biomedical concepts, which is commonly known as biomedical ontology matching. Biomedical concepts are usually complex and ambiguous, which makes matching biomedical ontologies a challenge. Since none of the similarity measures can distinguish the heterogeneous biomedical concepts in any context independently, usually several similarity measures are applied together to determine the biomedical concepts mappings. However, the ignorance of the effects brought about by different biomedical concept mapping's preference on the similarity measures significantly reduces the alignment's quality. In this study, a non-dominated sorting genetic algorithm (NSGA)-III-based biomedical ontology matching technique is proposed to effectively match the biomedical ontologies, which first utilises an ontology partitioning technique to transform the large-scale biomedical ontology matching problem into several ontology segment-matching problems, and then uses NSGA-III to determine the optimal alignment without tuning the aggregating weights. The experiment is conducted on the anatomy track and large biomedic ontologies track which are provided by the Ontology Alignment Evaluation Initiative (OAEI), and the comparisons with OAEI's participants show the effectiveness of the authors' approach.

1 Introduction

Over the recent years, ontologies have been extensively used in biomedical domains [1] such as annotation of medical records [2], medical knowledge representation and sharing [3], clinical data integration and medical decision making [4]. The vast usage of ontologies in biomedical domain has compelled researchers to develop more biomedical ontologies such as gene ontology (GO) [5], National Cancer Institute (NCI) thesaurus [6], Foundation Model of Anatomy (FMA) [7] and the Systemised Nomenclature of Medicine (SNOMED-CT) [8]. However, because of human subjectivity, various biomedical ontologies may use different terms for the same meaning or may use the same term to mean different things, yielding ontology heterogeneous problem. For example, when describing the muscles surrounding the human heart, NCI ontology uses the term 'Myocardium' but FMA utilises 'Cardiac Muscle Tissue'. Thus, to integrate the knowledge regarding human heart, it is necessary for a biomedical system to determine the correspondences between NCI and FMA. Similarly, finding correspondence between GO and FMA can be used by molecular biologist in understanding the outcome of proteomics and genomics in a large-scale anatomic view [9]. Moreover, the correspondences between ontologies have also been used for heterogeneity resolution among various health standards [10]. The biomedical concept mapping set between two ontologies is called the alignment and the process of discovering it is termed as ontology matching.

Matching biomedical ontologies is an open challenge in the ontology matching domain because biomedical concepts are usually complex and ambiguous. Frequently, the same entity has several names (e.g. gluconeogenesis, glucose synthesis and glucose biosynthesis, all refer to the same metabolic process), a common word refers to a biomedical concept (e.g. hedgehog and fruity are both gene names) or even the same word can be applied to two different biomedical concepts (e.g. lingula can either be a structure of the brain or the lung). Since none of the similarity measures can distinguish the same biomedical concepts in any contexts independently, the ontology matching systems actually apply several similarity measures to determine the correspondences between particular biomedical concepts. The most common composition of multiple similarity measures is the parallel composition, where the similarity measures are executed independently from each other and the aggregated correspondence is computed afterwards [11]. Currently, researchers mainly focus on how to tune the aggregating weights for various similarity measures to improve the quality of the ontology alignments [12]. However, the ignorance of the effects brought about by different biomedical concept mapping's preferences on some similarity measures significantly reduce the alignment's quality. For example, it is better to use the linguistic-based similarity measure instead of syntactic-based similarity measure to distinguish two terms 'Myocardium' and 'Cardiac Muscle Tissue', and weights tuned in this way could be problem specific, which means they might not be reused in other matching scenarios. Moreover, existing matching techniques can only deal with small-scale ontologies, and their runtime and memory consumption are always long and huge when matching biomedical ontologies which often possess tens of thousands of concepts. To effectively match the biomedical ontologies, in this paper, we propose a non-dominated sorting genetic algorithm (NSGA)-III-based [13] ontology matching technique to optimise the biomedical ontology alignment. In particular, the contributions made in this paper are as follows:

- A large-scale biomedical ontology matching framework is proposed.
- A many-objective optimal model is constructed for the biomedical ontology matching problem.
- A problem-specific NSGA-III has presented to optimise the biomedical ontology alignment, which can improve the convergence as well as maintain the diversity during the matching process.

The rest of this paper is organised as follows: Section 2 describes the related works; Section 4 shows the biomedical ontology partitioning technique; Section 5 defines many-objective similarity measure combining problem and presents the NSGA-III-based ontology matching technique; Section 6 presents the experimental studies and analysis; and finally, Section 7 draws the conclusions and presents the future work.

2 Related work

In general, the basic similarity measures can be divided into three broad categories, i.e. syntactic-based similarity measure, linguistic-based similarity measure and structure-based similarity measure. In particular, syntactic-based similarity measure computes the edit distance between ontology entities such as similarity measure for ontology alignment (SMOA) [14]. Linguistic-based matcher utilises synonymy, hypernymy and other linguistic relations to calculate the similarity score between ontology entities which require a lexicon and thesauri such as WordNet [15]. Structure-based matcher computes a similarity score between two ontological entities based on their ontology taxonomy hierarchy structure, and the common intuition is that two distinct ontology entities are similar when their adjacent entities are similar. The most popular structure-based similarity measures are the well known similarity flooding (SF) algorithm [16] and the profile-based similarity measure [12]. Although both of them utilise the ontology's taxonomy structure to calculate the similarity value, SF executes an iterative fix-point computing process, while the profile-based similarity measure first constructs for each entity a profile by collecting the data properties from its direct descendants and itself, then, the similarity value between two entities is measured by calculating the similarity of their corresponding profiles.

Usually, similarity measure combination and tuning are tackled by setting appropriate weight set through different methods. The most outstanding approach in this area is COMA++ [17] which utilises two kinds of similarity measures: simple similarity measure such as the syntactic-based similarity measure and linguistic-based similarity measure and hybrid similarity measure that combines multiple similarity measures. COMA++'s aggregating weights are determined by an expert. Lately, the focus is placed on the heuristic techniques for combining different similarity measures. The first method is called harmonic adaptive weighted sum which is presented in the PRIOR+ [18]. The harmony value is calculated through a similarity matrix and further assigned as the weight to the similarity measure associated with that matrix. PRIOR+ integrates the syntactic-based similarity measure and structure-based similarity measure. The second method is called the local confidence weighted sum, which is the core method for combining individual similarity measures in the AgreementMaker [19]. This measure is defined for an entity by considering the average of similarity values of entities that are associated (or not associated) with that entity. Finally, the selection of the final candidates from the set of candidates is performed by a greedy selection strategy. In particular, AgreementMaker utilises the syntactic-based similarity measure and linguistic-based similarity measure. For a given matching scenario, YAM++ [20] evaluates the degree of reliability of these similarity measures and assigns appropriate weight values to them. More recently, Benaissa and Khiat [21] propose a heuristic strategy to estimate the weights for different similarity measures, which is of a statistical nature and estimates the weights by an estimation of the precision standard metric. Particularly, the similarity measures they use are the linguistic-based similarity measure and structure-based similarity measure.

Recently, evolutionary algorithms (EAs) are appearing as an effective methodology to determine the optimal aggregating weights for different similarity measures. Genetic algorithm based ontology alignment (GOAL) [22] is the first matching system that utilises EA to determine the optimal weight configuration for a weighted average aggregation of several similarity measures by considering a reference alignment. A similar idea of combining multiple similarity measures is also developed by Naya *et al.* [23], Alexandru Lucian and Iftene [24] and Gulić *et al.* [11]. To improve efficiency, a hybrid EA is presented to tune the parameters for aggregating various similarity measures [12, 25]. More recently, Xue and Liu [26] present an approach based on a multi-objective EA to determine the optimal weights being assigned to the profile-based similarity measure, WordNet-based similarity measure and structure-based similarity measure. All these methods dedicate to tune the weights for aggregating different similarity measures, which ignore the effects brought about by different entity mappings' preferences on different similarity measures, and thus, decrease the quality of the alignment. In this work, a many-objective matching technique is proposed to further improve the alignment's quality, which takes into consideration each mapping's preference on various similarity measures and determine the optimal alignment without tuning the aggregating weights.

3 Large-scale biomedical ontology matching framework

The proposed large-scale biomedical ontology matching framework is shown in Fig. 1. As shown in this figure, our proposal first utilises an ontology partitioning technique to transform the biomedical ontology matching problem into several ontology segment-matching problems and then uses NSGA-III to combine various similarity measures and optimise the quality of the ontology alignment. The former technique can transform the large-scale biomedical ontology matching problem into several ontology segment-matching problems, which can improve the efficiency of the matching process hereafter. The latter can trade-off each biomedical concept mapping's preference on various similarity measures, and determine the optimal alignment without tuning the aggregating weights. Finally, the segment alignments are aggregated into a final alignment which is further evaluated with the reference alignment.

4 Biomedical ontology partition

Partitioning the large-scale biomedical ontology into various segments, where the term 'segment' is referred to as a fragment of an ontology, is an efficient way of reducing the algorithm's search space [27]. In this work, an alignment-oriented ontology partition technique [28] is introduced to partition the ontologies into various similar ontology segment pairs. First of all, the ontology with better reliability is selected as the source ontology. The reliability of an ontology is measured by the semantic accuracy, which is computed through the average of the squared semantic distance between each concept c_i and the ontology O's taxonomic root node ROOT. In particular, the formula for calculating semantic accuracy is presented as follows:

$$\mathrm{semAccuracy}(O) = \frac{\sum_{c_i \in C} \mathrm{semDistance}(c_i,\ \mathrm{ROOT})^2}{|C|} \qquad (1)$$

where $\mathrm{semDistance}(c_i,\ \mathrm{ROOT}) = \log_2\left(1 + \left((|\mathrm{Ances}(c_i)| - 1)/|\mathrm{Ances}(c_i)|\right)\right)$ calculates the semantic distance between the concept c c_i and

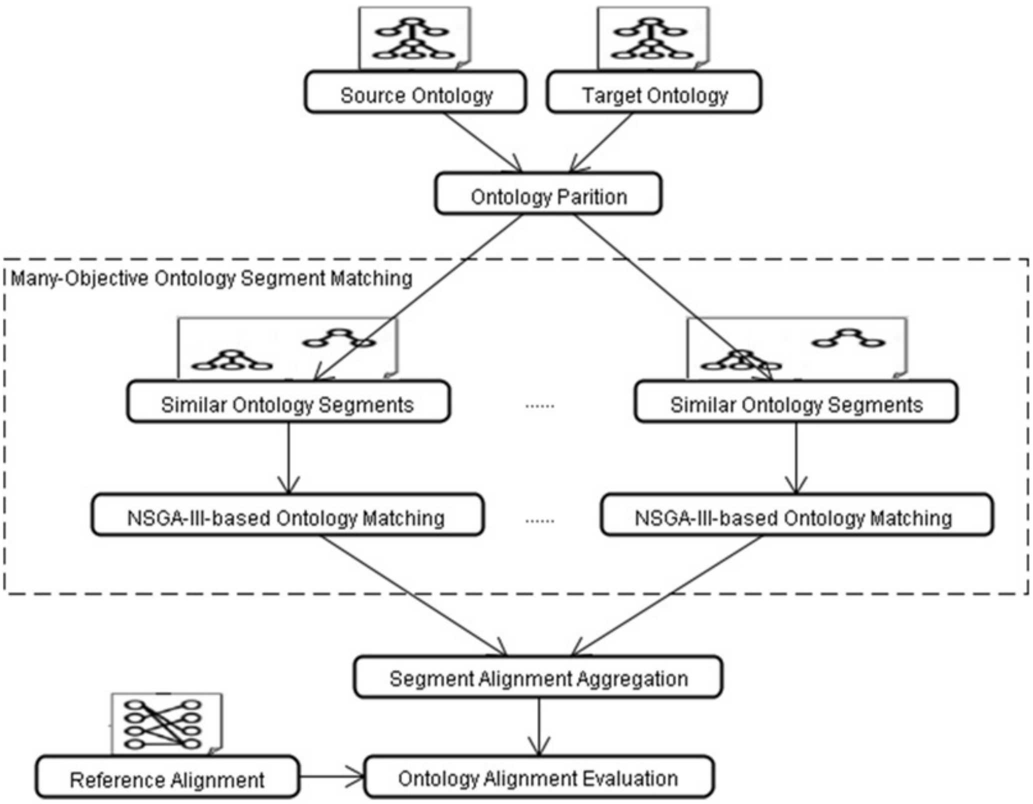
Source Ontology
Target Ontology
Ontology Parition
Many-Objective Ontology Segment Matching
Similar Ontology Segments
......
Similar Ontology Segments
NSGA-III-based Ontology Matching
......
NSGA-III-based Ontology Matching
Segment Alignment Aggregation
Reference Alignment
Ontology Alignment Evaluation

Fig. 1 *Large-scale biomedical ontology matching framework*

ROOT. Ances(c_i) refers to the set of taxonomic ancestors of concept c_i in the ontology including itself.

The source ontology is partitioned into disjoint segments through an ontology partition algorithm which is extended from structural clustering algorithm for network (SCAN) [29]. Then, a concept relevance measure-based approach is adopted to determine the similar target ontology segments of each source ontology segment seg_{src}. Particularly, for each target ontology concept c_i, the similarity value sim_{c_i} between c_i and seg_{src} is calculated by summing up every SMOA(c_i, c_j) (see also Section 7.1). If sim_{c_i} is larger than the threshold, c_i will be added to the candidate concept set C_{cand}. If the relevance value of a concept in C_{cand} is bigger than the threshold, it will be added to the final target segment. Given a concept $c_m \in C_{\text{cand}}$, the relevance value of c_m to source ontology segment can be calculated by the following formula:

$$\text{relevance}(c_m) = \text{sim}_{c_m} \times \sum_{c_n \in C_{\text{cand}}} \text{sim}_{c_n} \times \text{e}^{-(p(\text{c}_m,\, \text{c}_n))^2} \tag{2}$$

where sim_{c_m} and sim_{c_n}, respectively, denote the similarity values of c_m and c_n to seg_{src} and $p(c_m,\, c_n)$ is the shortest length between their corresponding vertexes in ontology taxonomy structure.

After partitioning the ontologies, the matching process only needs to deal with the similar biomedical ontology segments' matching problem, and all the similarity values obtained in the process of ontology partitioning are stored in the hash map to avoid repeating calculations in the hereafter matching process. With respect to the details of the alignment-oriented ontology partition algorithm, please see also [30].

5 Many-objective similarity measure combination

5.1 *Many-objective similarity measure combining problem*

Although the alignment evaluation measures recall, precision and f-measure [31] can reflect the quality of the resulting alignment, the reference alignment between two ontologies is usually unknown for real-life match problems [32]. In this work, 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 [33], we utilise the following metric to measure the quality of an alignment:

$$f(A) = \frac{2 \times \phi(A) \times \left(\sum_{i=1}^{|A|} \delta_i / |A|\right)}{\phi(A) + \left(\sum_{i=1}^{|A|} \delta_i / |A|\right)} \tag{3}$$

where $|A|$ is the number of correspondences in A; ϕ is a function of normalisation in [0,1]; and δ_i is the similarity value of the ith correspondence in A.

On this basis, the many-objective optimal model of combining various similarity measures can be defined as follows:

$$\begin{cases} \min \quad F(A) = (1 - f_1(A),\ 1 - f_2(A),\ \ldots,\ 1 - f_m(A)) \\ \text{s.t.} \quad A = (a_1,\ a_2,\ \ldots,\ a_{|C_1|})^{\text{T}} \\ \qquad a_i \in \{1,\ 2,\ \ldots,\ |C_2|\},\ i = 1,\ 2,\ \ldots,\ |C_1| \end{cases} \tag{4}$$

where m is the number of similarity measures; $f_i(A)$, $i = 1,\ 2,\ \ldots,\ m$, calculates the alignment A's quality with respect to the ith similarity measure; $|C_1|$ and $|C_2|$, respectively, represent the cardinalities of source concept set C_1 and target concept set C_2; and a_i, $i = 1,\ 2,\ \ldots,\ |C_1|$ represents the ith pair of correspondence.

Similarity measure takes as input two concept sets C_1 and C_2 and output a $|C_1| \times |C_2|$ similarity matrix $\boldsymbol{S}$, whose element s_{ij} is the similarity score between the ith concept in $|C_1|$ and the jth concept in $|C_2|$. Since the number of elements in biomedical ontology is large, we should avoid allocating an $n_1 \times n_2$ similarity matrix, where n_1 and n_2 are the cardinalities of two concept sets. On the basis of the observation that a correct alignment should be consistent with the concept hierarchies organised by 'is-a' [34], if two concepts c_1 and c_2 have high similarity value, so-called anchors in the partitioning process, the sub-concepts (/super-concepts) of c_1 and super-concepts (/sub-concepts) of c_2 can be skipped or directly set as 0. Then, considering the similarity matrix is a typical sparse matrix, the compression techniques can be further adopted to replace it. It usually compresses a similarity matrix into several mega bytes (MBs). In our approach, we first replace the two-dimensional (2D) reduction set with 1D style, then merge the continuous number of elements as a link.

5.2 NSGA-III for optimising biomedical ontology alignment

NSGA-III is a many-objective algorithm proposed by Deb *et al.* [35], which introduces a well-distributed reference points based clustering operator to replace the crowding distance operator in NSGA-II. In this work, NSGA-III [13] is utilised to automatically combine various similarity measures and determine the optimal biomedical ontology segment alignment. Original NSGA-III emphasises that the solutions should be Pareto non-dominated and closed to the reference line of each reference point. However, with the growing number of the objectives, selection pressure based on Pareto dominance would be too small to pull the population toward Pareto front, and in this case, NSGA-III indeed emphasises diversity more than convergence. To this end, we present a problem-specific NSGA-III to improve the convergence as well as maintain the diversity when matching the biomedical ontology segments.

Next, three key components of NSGA-III are presented in details, i.e. encoding mechanism, uniform design-based reference points generation and θ-dominance. Finally, the outline of problem-specific NSGA-III is given.

5.2.1 Encoding mechanism: Let $|C_1|$ and $|C_2|$ be the cardinalities of the source concept set C_1 and target concept set C_2, respectively. Each chromosome in the population would be a 1D array with $|C_1|$ elements, and the elements are denoted as: $N_1 N_2 \cdots N_{|C_1|}$, where $N_i \in \{0, 1, \ldots, |C_2|\}$, $i \in \{1, \ldots, |C_1|\}$, which means the ith concept in C_1 is mapped to the N_ith concept in C_2. In particular, when $N_i = 0$, the ith concept is not mapped to any concept in C_2.

5.2.2 Uniformly distributed reference points: In the original NSGA-III, the Das and Dennis's systematic approach [36] is used to generate reference points. However, when the number of objectives is high, the number of reference points generated by this approach would become very large [37]. In our work, we propose to use a uniform design [38], which aims at determining a set of points that are uniformly distributed over the design space, to produce uniformly distributed reference points in a unit sphere $S = \{(s_1, s_2, \ldots, s_m) | \sum_{i=1}^{m} s_i^2 = 1, s_i \geq 0, i = 1, 2, \ldots, m\}$. First, we need to generate a set of Q uniformly distributed points on $C = \{(c_1, c_2, \ldots, c_m) | 0 \leq c_1, c_2, \ldots, c_m \leq 1\}$. Let Q be the number of uniform distributed points in C; m be the dimension of the problem that is equal to the number of basic similarity measures in this work; δ be the number that yields the smallest discrepancy of generated point set (see also [39]), an integer matrix so-called uniform array $[\boldsymbol{M}_{ij}]_{Q \times m}$ can be calculated with $\boldsymbol{M}_{ij} = i\delta^{j-1} \bmod Q + 1$, $i = 1, 2, \ldots, Q$, $j = 1, 2, \ldots, m$, where the ith row of it can define a point $C_i = (c_{i,1}, c_{i,2}, \ldots, c_{i,m})$ with $c_{ij} = (2\boldsymbol{M}_{ij} - 1)/2Q$, $i = 1, 2, \ldots, Q$, $j = 1, 2, \ldots, m$. Next, a set of Q reference points uniformly distributed on S, denoted by $P(Q, m) = P_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,m})$, can be calculated as follows:

$$p_{i,j} = \begin{cases} \prod_{s=1}^{m-1} \cos(0.5c_{i,s}\pi) & j = 1 \\ \sin(0.5c_{i,m-j+1}\pi) \prod_{s=1}^{m-j} \cos(0.5c_{i,s}\pi) & 1 < j < m \\ \sin(0.5c_{i,1}\pi) & j = m \end{cases} \quad (5)$$

Equation (5) is a hyper-sphere formula, and in particular, it becomes a circular formula when $m = 2$ and a spherical formula when $m = 3$.

5.2.3 θ-Dominance: Given reference points $P(Q, m)$ which can be denoted by $\{P_i, P_2, \ldots, P_Q\}$, a reference line is defined by joining a reference point with the origin. After that, each individual is associated with a reference point by calculating the perpendicular distance of it from each of the reference line. The reference point whose reference line is closest to a solution is considered to be associated with this solution. In this way, the population can be split into Q clusters $C = \{C_1, C_2, \ldots, c_Q\}$ where the cluster C_j is represented by the reference point P_j, $j = 1, 2, \ldots, Q$.

Given a solution x and its objective vector $\boldsymbol{f}(x)$ which can be denoted by $[f_1(x), f_2(x), \ldots, f_m(x)]$, reference line L_j passing through the origin point Z and P_i, a penalty function [40] can be defined as $D_j(x) = \|(\boldsymbol{f}(x) - Z)s\| + \theta d_{j,\text{perpendicular}}(x)$, $j = 1, 2, \ldots, Q$, where $d_{j,\text{perpendicular}}(x)$ calculates the perpendicular distance between $\boldsymbol{f}(x)$ and L_j

$$d_{j,\text{perpendicular}}(x) = \left\| (\boldsymbol{f}(x) - Z) - \frac{\|(\boldsymbol{f}(x) - Z)^{\text{T}} P_j\|}{\|P_j\|} \left(\frac{P_j}{\|P_j\|} \right) \right\| \quad (6)$$

Given $m = 2$, an example of the perpendicular distance is shown in Fig. 2.

In this work, $\theta > 0$ is a predefined penalty parameter, which is set as 2 to achieve the best mean quality of alignment on all testing cases. It is obvious that the smaller $\|\boldsymbol{f}(x)\|$ and $d_{j,\text{perpendicular}}(x)$, respectively, lead to better convergence and better diversity. Given two solutions $x, y \in \Omega$, x is said to θ-dominate y, denoted by $x \prec_\theta y$, if $x, y \in C_j$ and $D_j(x) < D_j(y)$, $j \in \{1, 2, \ldots, Q\}$ [37]. Then, we utilise the θ-dominance to implement the fast non-dominated sorting [35] on the population to partition it into different θ-non-domination levels.

5.2.4 Flowchart of NSGA-III: The flowchart of NSGA-III is presented in Fig. 3. First, we apply a uniform design-based method to generate any number of reference points, and the common one point crossover operator and the bit mutation operator. Before calculating the perpendicular distance between a population and each of the reference lines, NSGA-III needs to normalise objectives' values and supplied reference points, which can ensure they have an identical range. In this work, since all the objective's values are in the same range [0, 1] and the ideal point is the zero vector, we do not need to carry out the normalisation in each generation. In addition, replace the Pareto dominance in NSGA-III with θ-dominance to trade-off the convergence and diversity in many-objective optimisation, and utilise the θ-dominance based fast non-dominated sorting is employed on the population clusters to divide them into different θ-non-domination levels. Finally, we determine the next generation's population by including one θ-non-domination at a time, which starts from

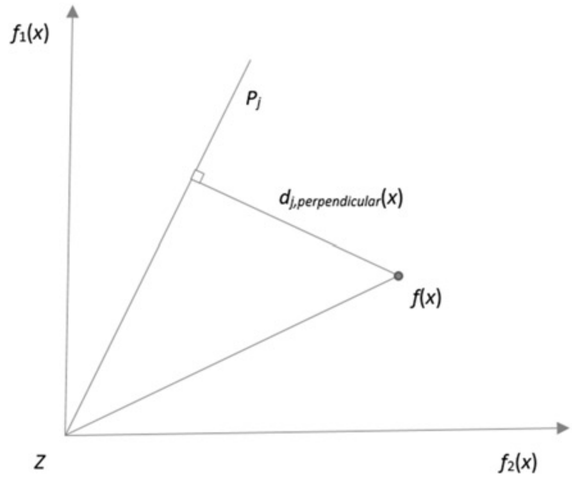

Fig. 2 *Example of the perpendicular distance*

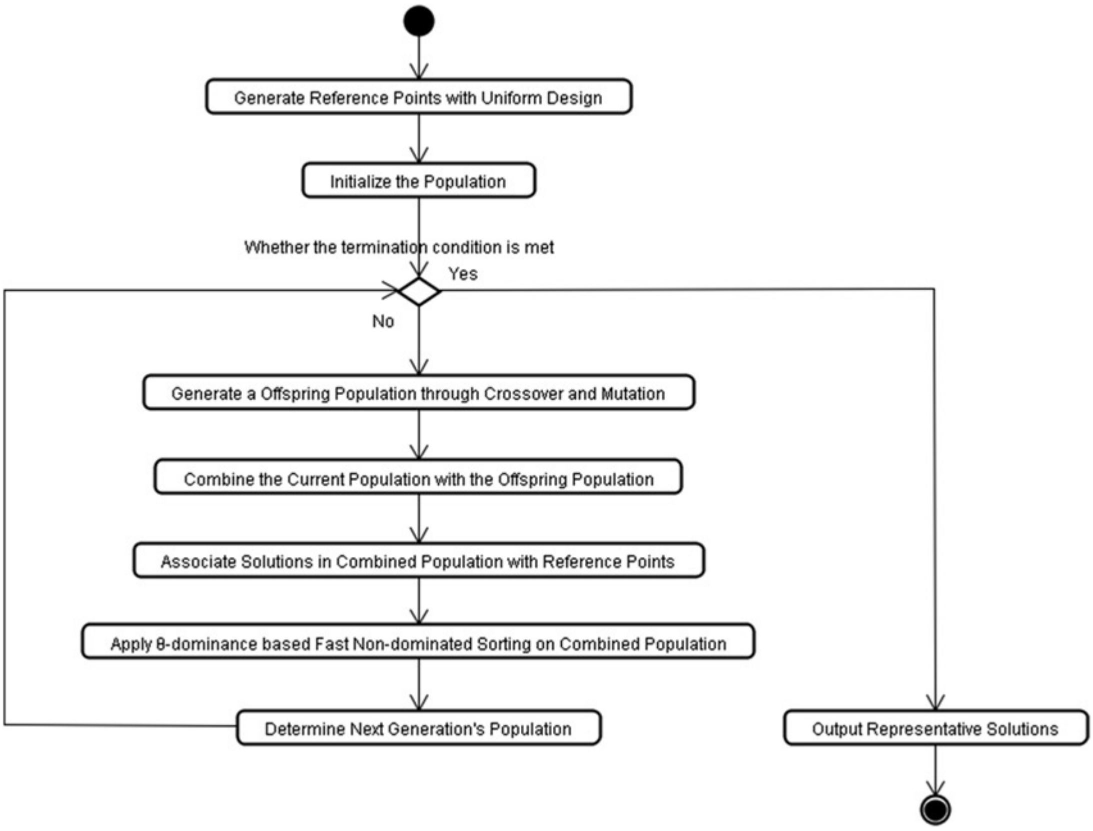
Generate Reference Points with Uniform Design
Initialize the Population
Whether the termination condition is met
Yes
No
Generate a Offspring Population through Crossover and Mutation
Combine the Current Population with the Offspring Population
Associate Solutions in Combined Population with Reference Points
Apply θ-dominance based Fast Non-dominated Sorting on Combined Population
Determine Next Generation's Population
Output Representative Solutions

Fig. 3 *Flowchart of NSGA-III*

the first level. With respect to the solutions in the last accepted level, we first sort them in ascending order according to their mean $f()$ values and then select the solutions sequentially. In this work, in order to compare with other ontology matching systems whose results are measured with f-measure, we pick up the solution in the Pareto front with the highest $\sum_{i=1}^{m} f_i/m$ as the representative solution.

6 Experimental studies and analysis

In this work, we exploit the Anatomy [http://oaei.ontologymatching.org/2017/anatomy/index.html.] and Large Biomed [http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/2017/.] track to study the effectiveness of our approach, which are provided by Ontology Alignment Evaluation Initiative (OAEI) 2017 [http://oaei.ontologymatching.org/2017.]. Tables 1 and 2 show the mean value of f-measure of the alignments obtained by our approach in 30 independent runs and the results obtained by the participants of OAEI.

Table 1 Comparison of anatomy track in OAEI 2017

Systems	*R*	*P*	*F*	Runtime, s
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
our approach	0.95	0.97	0.96	42

Three main categories of similarity measures are utilised in this work, i.e. SMOA (a syntactic-based similarity measure), Unified Medical Language System-based [41] similarity measure (a linguistic-based similarity measure) and profile-based similarity measure (a structure-based similarity measure) [12]. The parameters used by NSGA-III are as follows: numerical accuracy = 0.01, number of reference points = 20, population size = 25, crossover probability = 0.8, mutation probability = 0.02

Table 2 Comparison of large Biomed track in OAEI 2017

Systems	*R*	*P*	*F*	Runtime, s
Task 1: 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
our Approach	0.88	0.92	0.90	62
Task 2: 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
our Approach	0.82	0.93	0.87	165
Task 3: 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
our Approach	0.75	0.92	0.82	248

and maximum number of generation = 300. These parameters represent a trade-off setting obtained in an empirical way to achieve the highest average alignment quality on all test cases of the exploited dataset, which is robust against the heterogeneous situations in our experiment.

We run the anatomy track with a CPU @ 3.46 GHz × 6 with 8 GB allocated RAM, and the large biomed track with an Intel Core i9-8950HK CPU @ 2.90 GHz × 12 and 25 GB allocated RAM, which is the same with the OAEI's hardware configurations.

6.1 Anatomy track

The anatomy track is a large ontology matching task which is about matching the Adult Mouse Anatomy (2744 classes) and a part of the NCI thesaurus (3304 classes) describing the human anatomy. 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 42 s, which is less than AgreementMakerLight (AML), the best matcher of OAEI 2017 on Anatomy track. In this track, our approach's recall and precision are, in general, high, which further indicates the effectiveness of our approach.

6.2 Large biomedic ontologies track

This track aims at finding alignments between the large and semantically rich biomedical ontologies FMA, SNOMED-CT and NCI, which contains 78,989, 306,591 and 66,724 classes, respectively. The track has been split into three matching problems: FMA–NCI, FMA–SNOMED and SNOMED–NCI, and each matching problem in three tasks involving different fragments of the input ontologies.

As can be seen from Table 1, in terms of f-measure and running time, our approach's results are the best in all three tasks. In this track, our approach outperforms AML, which is the top ontology matcher and developed primarily for the biomedical ontology matching, in all three tasks in terms of f-measure, and the runtime of our approach is also less than AML. The experimental results further show the effectiveness of our proposal when matching large-scale biomedical ontologies.

7 Conclusion and future work

An ontology matching framework is proposed to efficiently match biomedical ontologies, which first uses an ontology partition technique to reduce the matching algorithm's search space, and then utilises an NSGA-III-based biomedical ontology matching technique to directly determine the optimal alignment without tuning the aggregating weights. The experimental results show that our proposal is able to efficiently determine the high-quality biomedical ontology alignments. In continuation of our research, we are interested in combining more similarity measures. Moreover, some strategies which could remove the mappings that lead to logical conflicts can be introduced to further improve the alignment's quality.

In the future, we are interested in getting the user involved in our approach to guide the search direction, so that the alignment quality could be further improved. Since the similarity measures would lead to the opposing results on the same biomedical concepts, before combining them, we need to select the effective similarity measures based on the heterogeneous characteristics of biomedical ontologies. How to select, combine and tune these similarity measures to improve the alignment's quality is a challenge especially when the scale of similarity measures is huge. Therefore, we are also interested in carrying out a future study on such situation as combining more than 50 similarity measures to improve our proposal.

8 Acknowledgments

This work was 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. 2015B20214), 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 the Scientific Research Foundation of Fujian University of Technology (No. GY-Z17162).

9 References

[1] Jiménez Ruiz, E., Meilicke, C., Grau, B.C., *et al.*: 'Evaluating mapping repair systems with large biomedical ontologies', *Description Logics*, 2013, **13**, pp. 246–257

[2] López Fernández, H., Reboiro Jato, M., Glez Peña, D., *et al.*: 'Bioannote: a software platform for annotating biomedical documents with application in medical learning environments', *Comput. Methods Programs Biomed.*, 2013, **111**, (1), pp. 139–147

[3] Isern, D., SáNchez, D., Moreno, A.: 'Ontology-driven execution of clinical guidelines', *Comput. Methods Programs Biomed.*, 2012, **107**, (2), pp. 122–139

[4] De Potter, P., Cools, H., Depraetere, K., *et al.*: 'Semantic patient information aggregation and medicinal decision support', *Comput. Methods Programs Biomed.*, 2012, **108**, (2), pp. 724–735

[5] Consortium, G.O.: 'The gene ontology (GO) database and informatics resource', *Nucleic Acids Res.*, 2004, **32**, (suppl_1), pp. D258–D261

[6] Golbeck, J., Fragoso, G., Hartel, F., *et al.*: 'The National Cancer Institute's thesaurus and ontology', *Web Semant. Sci. Services Agents World Wide Web*, 2011, **1**, (1), pp. 75–80

[7] Rosse, C., Mejino, J.L. Jr.: 'A reference ontology for biomedical informatics: the foundational model of anatomy', *J. Biomed. Inf.*, 2003, **36**, (6), pp. 478–500

[8] Schulz, S., Cornet, R., Spackman, K.: 'Consolidating SNOMED CT's ontological commitment', *Appl. Ontol.*, 2011, **6**, (1), pp. 1–11

[9] Heymans, S., McKennirey, M., Phillips, J.: 'Semantic validation of the use of SNOMED CT in hl7 clinical documents', *J. Biomed. Seman.*, 2011, **2**, (1), p. 2

[10] Ganiyat, I.O., Soriyan, H.A., Ishaya, G.P.: 'Resolving semantic heterogeneity in healthcare: an ontology matching approach', *J. Comput. Sci. Eng.*, 2013, **17**, (2), pp. 28–34

[11] Gulić, M., Vrdoljak, B., Ptiček, M.: 'Automatically specifying a parallel composition of matchers in ontology matching process by using genetic algorithm', *Information*, 2018, **9**, (6), pp. 953–958

[12] Xue, X., Wang, Y.: 'Optimizing ontology alignments through a memetic algorithm using both matchFmeasure and unanimous improvement ratio', *Artif. Intell.*, 2015, **223**, pp. 65–81

[13] Deb, K., Jain, H.: 'An evolutionary many-objective optimization algorithm using reference-point-based non-dominated sorting approach, part i: solving problems with box constraints', *IEEE Trans. Evol. Comput.*, 2014, **18**, (4), pp. 577–601

[14] Stoilos, G., Stamou, G., Kollias, S.: 'A string metric for ontology alignment'. Int. Semantic Web Conf., 2005, pp. 624–637

[15] Miller, G.A.: 'WordNet: a lexical database for English', *Commun. ACM*, 1995, **38**, (11), pp. 39–41

[16] Melnik, S., Garcia Molina, H., Rahm, E.: 'Similarity flooding: a versatile graph matching algorithm and its application to schema matching'. 2002 Proc. 18th Int. Conf. Data Engineering, 2002, pp. 117–128

[17] Aumueller, D., Do, H.H., Massmann, S., *et al.*: 'Schema and ontology matching with COMA++'. Proc. 2005 ACM SIGMOD Int. Conf. Management of Data, 2005, pp. 906–908

[18] Mao, M., Peng, Y., Spring, M.: 'An adaptive ontology mapping approach with neural network based constraint satisfaction', *Web Semant. Sci. Services Agents World Wide Web*, 2010, **8**, (1), pp. 14–25

[19] Cruz, I.F., Antonelli, F.P., Stroe, C.: 'Efficient selection of mappings and automatic quality-driven combination of matching methods'. Proc. Fourth Int. Conf. Ontology Matching, 2009, vol. 551, pp. 49–60

[20] Ngo, D., Bellahsene, Z.: 'Overview of yam++-(not) yet another matcher for ontology alignment task', *Web Semant. Sci. Services Agents World Wide Web*, 2016, **41**, pp. 30–49

[21] Benaissa, M., Khiat, A.: 'A new approach for combining the similarity values in ontology alignment'. IFIP Int. Conf. Computer Science and its Applications_x000D_, 2015, pp. 343–354

[22] Martinez Gil, J., Alba, E., Aldana Montes, J.F.: 'Optimizing ontology alignments by using genetic algorithms'. Proc. Workshop on Nature based Reasoning for the Semantic Web, Karlsruhe, Germany, 2008

[23] Naya, J.M.V., Romero, M.M., Loureiro, J.P., *et al.*: 'Improving ontology alignment through genetic algorithms'. Soft Computing Methods for Practical Environment Solutions: Techniques and Studies, 2010, pp. 240–259

[24] Alexandru Lucian, G., Iftene, A.: 'Using a genetic algorithm for optimizing the similarity aggregation step in the process of ontology alignment'. 2010 Ninth Roedunet Int. Conf. (RoEduNet), 2010, pp. 118–122

[25] Acampora, G., Loia, V., Vitiello, A.: 'Enhancing ontology alignment through a memetic aggregation of similarity measures', *Inf. Sci.*, 2013, **250**, pp. 1–20

[26] Xue, X., Liu, J.: 'Optimizing ontology alignment through compact moea/d', *Int. J. Pattern Recognit. Artif. Intell.*, 2017, **31**, (4), p. 1759004

[27] Rahm, E.: 'Towards large-scale schema and ontology matching', in (Eds.): 'Schema matching and mapping' (Springer, Berlin Heidelberg, Germany, 2011), pp. 3–27

[28] Xue, X., Chu, S.C.: 'An alignment-oriented segmenting approach for optimizing large scale ontology alignments', *J. Internet Technol.*, 2016, **17**, (7), pp. 1373–1382

[29] Yuruk, N., Mete, M., Xu, X., *et al.*: 'AHSCAN: agglomerative hierarchical structural clustering algorithm for networks'. Int. Conf. Advances in Social Network Analysis and Mining, Athens, Greece, 2009, pp. 72–77

[30] Xue, X., Pan, J.S.: 'A segment-based approach for large-scale ontology matching', *Knowl. Inf. Syst.*, 2017, **52**, (2), pp. 467–484

[31] Rijsberge, C.J.V.: 'Information retrieval' (University of Glasgow, Butterworth, London, 1975)

[32] Xue, X., Wang, Y., Hao, W., *et al.*: 'Optimizing ontology alignments through NSGA-II without using reference alignment', *Comput. Inf.*, 2015, **33**, (4), pp. 857–876

[33] Bock, J., Hettenhausen, J.: 'Discrete particle swarm optimisation for ontology alignment', *Inf. Sci.*, 2012, **192**, pp. 152–173

[34] Wang, P., Zhou, Y., Xu, B.: 'Matching large ontologies based on reduction anchors'. IJCAI, 2011, pp. 2343–2348

[35] Deb, K., Pratap, A., Agarwal, S., *et al.*: 'A fast and elitist multiobjective genetic algorithm: NSGA-II', *IEEE Trans. Evol. Comput.*, 2002, **6**, (2), pp. 182–197

[36] Das, I., Dennis, J.E.: 'Normal-boundary intersection: a new method for generating the Pareto surface in non-linear multicriteria optimization problems', *SIAM J. Optim.*, 1998, **8**, (3), pp. 631–657

[37] Yuan, Y., Xu, H., Wang, B.: 'An improved NSGA-III procedure for evolutionary many-objective optimization'. Proc. 2014 Annual Conf. Genetic and Evolutionary Computation, 2014, pp. 661–668

[38] Fang, K.T., Wang, Y.: 'Number-theoretic methods in statistics', vol. 51 (CRC Press, Chapman and Hall, London, 1993)

[39] Cai, D., Yuping, W.: 'A new uniform evolutionary algorithm based on decomposition and CDAS for many-objective optimization', *Knowl.-Based Syst.*, 2015, **85**, pp. 131–142

[40] Zhang, Q., Li, H.: 'Moea/d: a multiobjective evolutionary algorithm based on decomposition', *IEEE Trans. Evol. Comput.*, 2007, **11**, (6), pp. 712–731

[41] Bodenreider, O.: 'The Unified Medical Language System (UMLS): integrating biomedical terminology', *Nucleic Acids Res.*, 2004, **32**, (suppl_1), pp. D267–D270