

基于改进遗传规划算法的本体匹配方法

申建朝¹,戴可涛²,薛醒思³,李何²,吕青²

(1. 潞安化工集团慈林山煤业有限公司夏店煤矿,山西 长治 046299;

2. 太原理工大学 电气与动力工程学院,山西 太原 030024;

3. 福建理工大学 计算机科学与数学学院,福建 福州 350118)

摘要: 受到遗传规划 (genetic programming, GP) 在特征构建领域成功应用的启发,提出一种基于改进遗传规划算法的本体匹配方法 (ESMPD-GP)。为了解决 GP 个体简化压力问题、避免算法在进化过程中陷入局部最优,ESMPD-GP 采用了两种新的算法组件,即双目标精英选择机制和种群多样性增强策略。前者通过同时考虑个体适应度和编辑距离,选出种群中多样性最强的个体集合;后者通过创建额外种群引入优秀解简化压力,增加种群多样性、防止早熟收敛。基于国际本体匹配竞赛上 Benchmark 和 Anatomy 测试集的实验结果表明,所提方法在 Benchmark 上的 f-measure 平均值为 0.96,优于现有先进方法。

关键词: 本体匹配;遗传规划;精英选择;种群多样性增强

中图分类号: TP301.6

文献标志码: A

文章编号: 2097-3853(2025)04-0344-10

Ontology matching method based on improved genetic programming algorithm

SHEN Jianchao¹, DAI Ketao², XUE Xingsi³, LI He², LYU Qing²

(1. Xiadian Coal Mine of Lu'an Chemical Group Cilinshan Coal Industry Co., Ltd., Changzhi 046299, China;

2. School of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan 030024, China;

3. School of Computer Science and Mathematics, Fujian University of Technology, Fuzhou 350118, China)

Abstract: Inspired by the successful application of genetic programming (GP) in the field of feature construction, an ontology matching method based on improved genetic programming algorithm (ESMPD-GP) is proposed. In order to solve the problem of GP individual simplification pressure and avoid the algorithm falling into local optimum in the evolution process, ESMPD-GP adopts two new algorithm components, namely, dual-objective elite selection mechanism and population diversity enhancement strategy. The former selects the individual set with the strongest diversity in the population by considering both individual fitness and edit distance. The latter introduces excellent solutions to simplify the pressure by creating additional populations, increasing population diversity and preventing premature convergence. Experimental results based on Benchmark and Anatomy test sets on the Ontology Alignment Evaluation Initiative (OAEI) show that the proposed method outperforms existing state-of-the-art methods with an average f-measure of 0.96 on Benchmark.

Keywords: ontology matching; genetic programming; elite selection; population diversity enhancement

计算机领域的本体 (ontology) 作为语义网 (semantic web, WB) 基础技术,通过类、属性、实例等元素对领域知识进行形式化表达^[1],支撑自然语言处理、智能搜索及机器学习等应用。其构

建过程的主观性易导致系统、语法和层次异构问题^[2],而本体匹配技术通过关联不同本体的概念映射,可有效消除语义异质性,最终实现知识共享与语义融合。

收稿日期:2024-11-18

基金项目:山西省省筹资金资助回国留学人员科研项目 (2023061)

第一作者简介:申建朝 (1986—),男,山西晋城人,工程师,硕士,研究方向:本体匹配、知识图谱的故障诊断。

近年来,进化算法如遗传算法 (genetic algorithm, GA) 和遗传规划 (genetic programming, GP) 等在本体匹配的应用取得了显著进展。与 GA 相比,GP 通过树型编码可以自由灵活地组合相似度方法,扩展搜索空间并提升全局搜索能力,有利于处理本体匹配问题。然而 GP 会出现陷入局部最优使其性能下降等问题,GP 的可变大小染色体会导致个体规模随迭代增加,形成 GP 个体简化压力问题^[3],从而增加计算成本和内存,导致搜索效率降低。

算法陷入局部最优问题的解决思路是增加种群多样性,XUE 等^[4]提出基于分群策略的混沌粒子群算法,通过 PSO 分群和混沌分群粒子交叉替代,增加种群多样性以提高算法全局寻优能力。Baker 等^[5]提出一种多样性阈值,定义多重多样性度量量化种群的多样性,并根据当前种群的多样性水平动态调整交叉变异操作。为平衡探索与开发,Mei 等^[6]引入精英选择标准,动态调整每个非支配前沿中进入下一代父代种群的精英解数量,从而增加了种群的多样性。

然而,现有的方法很少同时关注算法性能与 GP 个体简化压力问题。为此,本研究提出了一种包含精英选择策略与种群多样性增强策略的遗传规划算法 (ESMPD-GP)。通过进化单目标和多目标两个种群,增强单目标种群算法性能,帮助算法跳出局部最优;设计了一种新型精英选择机制 (AESE) 筛选出代表个体,为确定 AESE 选择半径,使用两个指标共同选择种群精英个体;还使用了一种树型编码方式,通过 GP 自由组合相似度度量方法及聚合过滤方法,提高匹配性能。

1 新的串行匹配框架

新的串行进化本体匹配框架由单目标种群和多目标种群两部分各自承担不同的进化方向,如图 1 所示。单目标种群关注匹配质量,而多目标种群同时关注匹配质量和树的规模。首先初始化两个种群的个体,每个个体由新的树型编码方式构建。多目标种群首先进行交叉和变异生成后代,根据多目标种群公式计算出个体适应度值;再利用 NSGA-II^[7] 的思想,对种群进行非支配排序和拥挤度计算;记录当前代的非支配个体并选择

出保留到下一代个体后,执行单目标种群的进化。单目标种群使用进化算子生成后代及进行评估,然后应用 AESE 精英选择算子筛选出代表解。当单目标种群达到设定条件时,根据所记录的非支配解集引入非支配个体,增加单目标种群多样性来帮助算法跳出局部最优,引导算法进化。若未达到多样性增强条件,则直接进行锦标赛选择,保留 num_s 个个体进入下一代。最终,当单目标种群精英个体适应度值达到 1.0 或达到最大代数 Max-gen 时,算法终止。算法返回单目标种群的最优 GP 树将通过源本体和目标本体的输入后,得到最终的匹配结果。

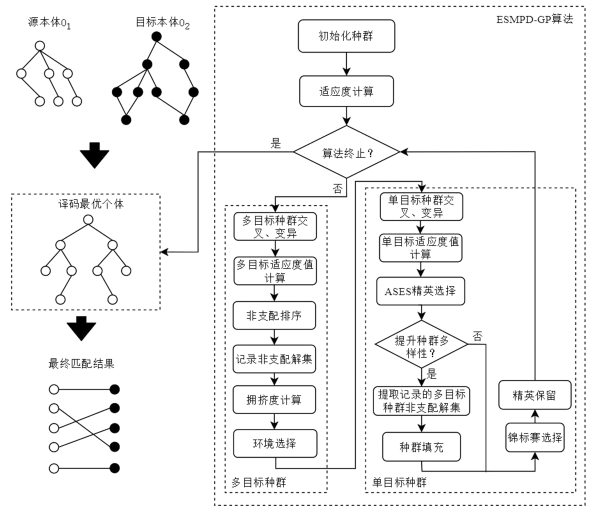


图 1 新的串行进化本体匹配框架

Fig.1 New sequential framework of evolutionary ontology matching

2 基于改进遗传规划算法的本体匹配方法

2.1 个体编码和解码方式

为了提升算法在本体实体匹配领域的适用性与匹配效能,使用了树型编码方式将相似度方法的集成、计算及筛选过程集成于单一树结构中,确保了个体表达的灵活性与适应性。在生成树时使用 Grammar-based GP 进行约束,利用简单的 Grammar 来引导两层 GP 树的生成。采用串行方式先生成相似度方法选择与组合层,再生成匹配结果提取层。各个函数种类及方法见表 1 (各类相似度度量方法具体可参考文献[8])。

表 1 语法树中的函数及终端描述

Tab.1 Function and terminal descriptions in grammar tree

层类型	方法类型	功能	函数集
相似度方法选择与组合层	相似度方法	基于字符串的相似度度量方法 (Syntax-based similarity feature)	SMOA、Jaro-Winkeler、LCS、Damerau-Levenshtein、Jaccard similarity、N-gram、Dice's coefficient、Cosine similarity
		基于语义的相似度度量方法 (Semantic-based similarity feature)	Resnik distance、Wu and Palmer
		基于结构的相似度度量方法 (Structure-based similarity feature)	SimRank、Numbered Hierarchy Distance、Hierarchy Distance、Instance-based Distance
	计算方法	加、减、乘、除	(+、-、*、/)
匹配结果提取层	聚合方法	取两个矩阵中同一位置最大值	Max
		取两个矩阵中同一位置平均值	Mean
		矩阵前 30%置为 1,其余置为 0	Top30
		矩阵前 50%置为 1,其余置为 0	Median
	过滤方法	矩阵每行每列的最大值置为 1,其余置为 0	Maximum
		大于阈值置为 1,其余置为 0	{ 0.80, 0.85, 0.90, 0.95 }
		启发式方法	Naive descending extraction

新树具体结构如图 2 所示,其中左侧为架构概览,右侧为实例演示。

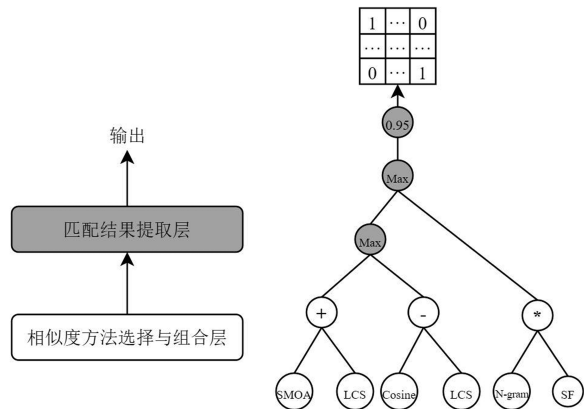


图 2 所提出 GP 树结构以及示例

Fig.2 Proposed GP tree structure and an example

新树框架包括相似度方法选择与组合层、匹配结果提取层。相似度方法选择与组合层负责灵活集成相似度度量方法。节点由相似度度量方法和计算方法构成,输入为源本体和目标本体中所有的实体。相似度度量方法包括基于字符串、基于语义及基于结构的相似度度量。该层通过计算输入的所有实体之间的相似度值,生成反映实体间联系的不同相似度方法的相似度矩阵。匹配结

果提取层负责提取相似度方法选择与组合层输出的相似度融合矩阵,得到映射实体匹配关系的二进制矩阵。

解码过程是从叶子节点开始到根节点结束,根据不同节点所选用的函数对矩阵进行对应的操作。

2.2 新的精英选择方法

传统单目标进化过程仅依赖适应度值的选择压力容易导致种群结构趋同,使算法容易陷入局部最优而导致早熟收敛。本体匹配问题有众多局部最优解,若是采用传统方法,会加剧该现象。因此如何促进收敛的同时使算法收敛到全局最优解成为本研究的一大挑战。

个体基因型的相似程度可用编辑距离的大小来表示,参考文献[9]的研究,编辑距离的计算如式(1)所示。

$$\text{Edit}(t_1, t_2) = \begin{cases} \text{dist}(n_1, n_2), & \text{如果 } n_1 \text{ 或 } n_2 \text{ 为叶子节点} \\ \text{dist}(n_1, n_2) + \sum_{k=0}^n \text{Edit}(t_1, k, t_2, k), & \text{其他} \end{cases}$$

(1)

其中, t_1, t_2 表示需要距离计算的两棵树, n_1, n_2 分

别为 t_1 、 t_2 的根节点; n 表示 t_1 、 t_2 中最大的节点个数; $t_{1,k}$ 、 $t_{2,k}$ 分别表示 t_1 、 t_2 的第 k 个节点的子树。如果节点 n_1 、 n_2 不相同, $\text{dist}(n_1, n_2)$ 的值为 1, 反之则为 0。

本研究提出精英选择方法 AESE, 根据个体的基因型和表现型筛选出单目标种群中的代表解。图 3 是 AESE 工作过程示意图, 横坐标表示适应度值, 横坐标的距离反映了个体间适应度值的差距, 纵坐标表示两个个体之间的编辑距离。首先, 将种群中适应度值最大的个体挑出放入坐标系中, 标记为灰色。然后以该解为顶点, 在长为 $\text{Threshold}_{\text{fitness}}$ 、宽为 $\text{Threshold}_{\text{edit}}$ 的矩形区域内, 形成一个黑色虚线矩形, 记作矩形 1。计算所有解与最优个体的编辑距离, 若编辑距离小于 $\text{Threshold}_{\text{edit}}$ 且适应度值差距小于 $\text{Threshold}_{\text{fitness}}$, 则将该解加入到矩形 1 中, 并标记为黑色。接着忽略种群中已标记在矩形 1 中的所有解, 重新找到种群中的最优个体标记为灰色。按照上述方法构建矩形 2, 找到相似个体标记为黑色。重复以上操作至所有解都被标记完成, AESE 将灰色标记的解作为精英保留到下一代。

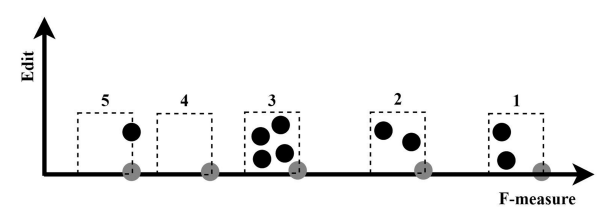


图 3 新精英选择机制图解

Fig.3 New elite selection mechanism diagram

2.3 适应度函数

适应度函数能够使算法确定哪些个体可以更好地适应环境, 从而引导算法的进化方向, 对算法的性能有着决定性的影响。对于单目标种群, 关注优化匹配质量, 适应度函数设置为 F-measure^[10]。

设置多目标种群的目的在于增加单目标种群的多样性, 利用树规模较小且匹配结果不差的个体引导算法的进化方向, 因此多目标种群的适应度函数还要考虑树的规模。多目标优化通过同时最大化 F-measure 和 GP 个体规模的倒数, 使得种群中具有高匹配质量又较小规模的个体。这些个体有助于解决单目标种群陷入局部最优和 GP 个体简化压力问题。

2.4 进化算子

在进化操作过程中, 多目标和单目标种群采用同一方式进行交叉和变异。对于所选择出的两个亲本, 以交叉概率 rate_c 决定是否进行交叉。若不交叉, 子代直接复制两亲本; 若交叉, 则采用子树替换策略, 随机交换双亲同类型子树以产生新个体, 期间所有亲本均被标记以确保种群遍历。交叉完成后, 亲本与子代共同构成规模为 $2 * \text{num}_s$ 的中间种群, 随后对所有个体施加变异操作。

就变异而言, 对交叉得到的规模为 $2 * \text{num}_s$ 的中间种群中的每个个体经过变异概率 rate_m 进行判断。若不变异, 则该个体保持不变; 若进行变异, 则子树变异或节点变异的概率各为 50%。节点变异通过改变若干个节点的函数生成新的个体, 子树变异将用一棵新生成的子树替换原个体中的一棵子树生成新个体(以上变异均在满足语法的情况下进行), 变异后的个体将替换原个体。

2.5 种群多样性增强策略

当单目标种群连续 δ 代的最优个体的适应度值没有改变时, 执行种群多样性增强策略, 引入多目标种群帕累托解集。该解集包含了规模小或者匹配性能优秀的个体, 匹配性能优秀个体有助于提高算法整体匹配性能; 规模小的个体使进化过程更有可能产生规模小的个体, 从而缓解简化压力。通过该策略增强单目标种群多样性, 帮助其跳出局部最优。

为了确定 δ 取值, 进行多样性及收敛性测试。在测试单目标 GP 算法的收敛性的同时, 也测试了种群随迭代进行的种群多样性变化, 其中种群多样性的指标设置为种群适应度值的方差。测试结果如图 4 所示, 图中, a 、 b 、 c 3 点分别表示算法跳出当前停滞点或者区域。算法在迭代初期, 种群多样性不断增加, 伴随种群的适应度值持续上升。但随着迭代进行至 a 点前, 种群多样性急剧降低, 算法适应度值缓慢上升。通过进化算子在 a 点时多样性有所上升, 但算法仍陷入了局部最优, 该状态持续到了 b 区域前。 b 区域开始时种群多样性上升并保持, 使算法跳出了当前局部最优, 但随后陷入停滞。直到 c 点多样性最后一次上升使算法再次打破原先停滞状态, 种群多样性下降使其无法跳出局部最优, 算法适应度值达到极限。

从测试观察到,当算法性能出现停滞时,若种群多样性无法提升,则算法性能也无法进一步提高。结合图中出现多次超过 20 代算法性能仍陷入停滞的情况,将 δ 设置为 10。

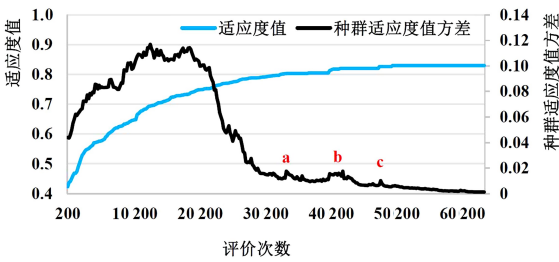


图 4 多样性及收敛性测试

Fig.4 Diversity and convergence tests

3 实验结果以及分析

为了评估所提出的 ESMPD-GP 算法在本体匹配上的性能,使用了 Ontology Alignment Evaluation Initiative (OAEI) 的 101~262 基准测试集及 Anatomy 测试集。为了比较的公平性,单目标种群适应度函数设置为 F-measure,多目标种群适应度函数设置为 F-measure 和 GP 树规模的倒数,在验证最优 GP 树时使用真实 f-measure^[10] 公式。

3.1 实验参数

本研究将所提出的 ESMPD-GP 算法与 OAEI 竞赛的参与者进行了比较。这些参与者包括 AOTL(ontology alignment at terminological and linguistic level)、IAMA(institute of automation's matcher)、MapSSS(map with syntactic, semantic, and structural metrics)、YAM++(yet another matcher)、CroMATCH 以及 Lily。以上方法在 Benchmark 及 Anatomy 测试集的匹配结果均可在 OAEI 官网上查阅。

此外,还将 ESMPD-GP 与其他几种基于进化算法的方法进行了比较,包括 GA^[11]、HGA(genetic algorithm with hill climbing)^[12] 以及常规的 GP^[13]。为确保对比公平,配置的最大代数、种群规模都与 ESMPD-GP 一致。算法设置最大代数 Maxgen 为 1 000;每个种群的种群规模 num_s 为 100;变异概率 rate_m 为 0.15;交叉概率 rate_c 为 0.8;初始生成树的最大深度为 7 层;锦标赛规模为 3;适应度阈值 Threshold_{fitness} 为 0.05;编辑距离阈值 Threshold_{edit} 为 40。

3.2 参数敏感性测试

AESE 所设置的 Threshold_{fitness} 和 Threshold_{edit} 两个阈值分别是为了筛选适应度值相似和结构相似的个体,过高的 Threshold_{fitness} 和 Threshold_{edit} 会导致算法选择的精英个体不足,造成优秀基因的缺失;过低的 Threshold_{fitness} 和 Threshold_{edit} 会导致算法选择出过多的精英个体,从而使 AESE 失效。AESE 中的阈值参数 Threshold_{fitness} 和 Threshold_{edit} 的参数敏感性测试结果如图 5 所示。从图 5 可见,当 Threshold_{fitness} 为 0.05 且 Threshold_{edit} 设为 40 时,测试结果达到了最优的性能。

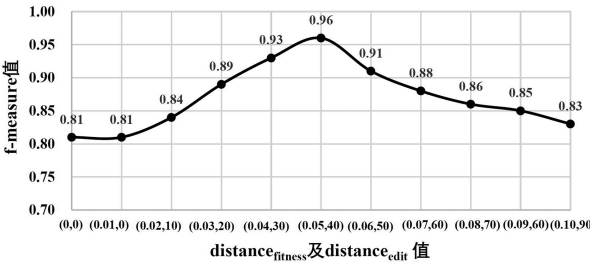


图 5 Threshold_{fitness} 及 Threshold_{edit} 参数敏感性测试

Fig.5 Sensitivity testing on Threshold_{fitness} and Threshold_{edit}

在进化算法中,变异概率和交叉概率对算法性能的影响非常明显。ESMPD-GP 中的变异概率 rate_m 和交叉概率 rate_c 的参数敏感性测试结果分别如图 6、图 7 所示。

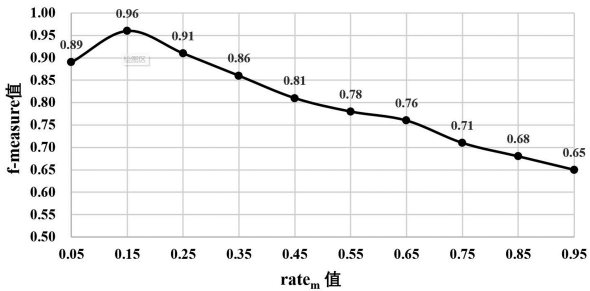


图 6 rate_m 参数敏感性测试

Fig.6 Sensitivity testing on rate_m

对于变异概率而言,过低的变异概率会使种群中的个体变得高度相似,缺乏探索新的解空间的能力,从而使算法陷入局部最优;过高的变异概率导致种群中的个体频繁变化,降低种群的稳定性,使算法陷入随机搜索的状态。在所提的 ESMPD-GP 算法中,rate_m 取 0.15 时,算法性能最优。对于交叉概率而言,过低的交叉概率意味着

个体之间的基因的交流很少,导致算法在搜索过程中无法有效利用已有的优质基因,使算法性能下降;而过高的交叉概率使得个体的基因信息迅速传播,从而降低个体间的多样性,种群会失去探索新解空间的能力,容易陷入局部最优。在 ESMPD-GP 算法中,rate_c 取 0.8 时,算法性能最优。

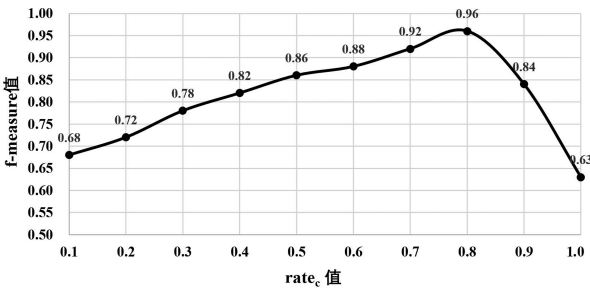


图 7 rate_c 参数敏感性测试

Fig.7 Sensitivity testing on rate_c

3.3 匹配结果质量对比

表 2 展示了 ESMPD-GP 与其他最先进的本

体匹配方法在 OAEI 基准数据集上获得的平均 f-measure。表格最后一列列出了 ESMPD-GP 在 30 次独立运行中的平均值,每个其他最先进的本体匹配方法旁的符号“(+)”/“(=)”/“(=)”表示使用 α 值为 0.05 的 t 检验后 ESMPD-GP 与之相比是否表现得更好、更差或与之相当。最后一行显示了 ESMPD-GP 相对于其他算法在性能上表现出明显更好、更差或等效的测试用例数量。

从表 2 可见,ESMPD-GP 在几乎所有测试集上的性能都不亚于甚至明显优于 OAEI 的优秀参与者。在 248 之前的测试集中,由于本体相对简单,ESMPD-GP 表现出色,几乎所有测试集的 f-measure 值均为 1。而在 248~262 这些较复杂的测试集中,ESMD-GP 依然取得了令人满意的结果。在 248~262 这 13 个测试集上,ESMPD-GP 的表现远远胜过其他算法。综合所有测试集,ESMPD-GP 平均 f-measure 为 0.96,其他方法中 CroMatch 的平均 f-measure 值最高,仅为 0.81。

表 2 ESMPD-GP 在 30 次独立运行后的平均 f-measure 值结果与 OAEI 参与者 f-measure 值对比

测试集	f-measure						
	AOTL	Lily	MapSSS	YAM++	CroMather	IAMA	ESMPD-GP
101~202	0.91 (+)	0.99 (=)	0.94 (+)	0.92 (+)	0.95 (+)	0.70 (+)	0.99
221~247	0.98 (+)	1.00 (=)	0.99 (+)	1.00 (=)	1.00 (=)	1.00 (=)	1.00
248~262	0.35 (+)	0.63 (+)	0.44 (+)	0.59 (+)	0.65 (+)	0.17 (+)	0.91
Avg.	0.66	0.80	0.72	0.78	0.81	0.60	0.96
+ / = / -	24/8/0	14/17/1	20/12/0	15/17/0	14/18/0	15/17/0	

表 3 展示了 ESMPD-GP 与 4 种进化算法 (GA、HGA、GP 和 LGP^[14]) 在 34 个测试集上的性能比较。ESMPD-GP 在 34 个测试集中的表现均优于 GA,在 19 个测试集中优于 HGA,在 18 个测试集中优于 GP,在其他测试集中表现相当。ESMPD-GP 在 9 个测试集的结果高于 LGP,在测试集 254 上低于 LGP,观察其未找到的实体缺少充足的结构信息,并且在语义和字符串上与所匹配的实体有较大差异,故而结果不理想。

表 3 ESMPD-GP 与其他进化算法 30 次独立运行的平均 f-measure 的性能比较

Tab.3 Performance comparison of of ESMPD-GP with other evolutionary algorithms in terms of mean f-measure over 30 independent runs

测试集	f-measure				
	GA	HGA	GP	LGP	ESMPD-GP
101~104	0.93 (+)	1.00 (=)	1.00 (=)	1.00 (=)	1.00
201~202	0.84 (+)	0.89 (+)	0.90 (+)	0.94 (+)	0.98
22~1247	0.90 (+)	0.97 (+)	0.97 (+)	1.00 (=)	1.00
248~262	0.47 (+)	0.30 (+)	0.87 (+)	0.89 (+)	0.91
+ / = / -	34/0/0	19/15/0	18/16/0	9/24/1	

为了验证 ESMPD-GP 的先进性,本实验将 ESMPD-GP 与过去两年内包括神经网络和机器学习等先进的本体匹配技术^[15-17]进行了比较。表 4 展示了上述方法与 ESMPD-GP 在 Benchmark 测试集的 f-measure 平均值。

实验结果表明,无论是面对相对简单的 101~104 数据集,还是面对较为复杂的 248~262 数据集,ESMPD-GP 的表现始终不亚于甚至显著优于其他先进技术。这是由于与包括 Neurofuzzy 在内的神经网络方法相比,ESMPD-GP 进化过程(如基因编码、选择、交叉和变异等)具有更优秀的透明度和可解释性。尤其是在解决本体匹配问题时,优化过程和每代的变化可以被清晰观察。与包括 DRL-OMT 在内的机器学习方法相比,在处理如实体匹配问题时,ESMPD-GP 拥有更为强大的解决组合优化问题能力。

表 4 ESMPD-GP 与其他前沿方法的 f-measure 比较

Tab.4 Comparison of f-measure of ESMPD-GP with other advanced methods

测试集	f-measure			
	DRL-OMT ^[15]	SR ^[16]	Neurofuzzy ^[17]	ESMPD-GP
101~104	1.00 (=)	1.00 (=)	1.00 (=)	1.00
201~202	0.98 (=)	0.90 (+)	0.95 (+)	0.98
221~247	1.00 (=)	0.97 (+)	0.99 (+)	1.00
248~262	0.89 (+)	0.86 (+)	0.89 (+)	0.91
+/=/-	1/3/0	3/1/0	3/1/0	

3.4 Anatomy 测试集结果

为验证 ESMPD-GP 算法在其他测试集上的性能,在 Anatomy 测试集上进行了实验。Anatomy 测试集通常包含来自生物医学领域的本体,这些

本体可能包括不同的解剖结构、器官、组织等。

表 5 展示了 OAEI 参与者、基于 BERT 方法 (OLaLa)^[18] 以及其他进化算法的对比结果。OAEI 参与者的方法与结果均可在 OAEI 官网查询。ESMPD-GP 的平均 f-measure 值为 0.96,相比起其他方法性能更加优秀。

表 5 与 OAEI 参与者、BERT 方法及前沿进化算法在 Anatomy 测试集上 30 次独立运行 f-measure 平均值比较

Tab.5 Comparison of 30 independent runs of f-measure averages on Anatomy test set with OAEI participants, BERT method, and advanced evolutionary algorithms

方法	Anatomy 测试集平均 f-measure 值
AML	0.94 (+)
LogMap	0.88 (+)
LogMapbio	0.89 (+)
ATMatch	0.79 (+)
ATBox	0.80 (+)
Wiktionary	0.84 (+)
Lily	0.90 (+)
OLaLa ^[18]	0.90 (+)
CMP SO-K ^[19]	0.94 (0.01) (+)
cFA ^[20]	0.94 (0.01) (+)
ESMPD-GP	0.96 (0.01)

3.5 消融实验

在 ESMPD-GP 中,使用的多样性增强策略、AESE 机制是极为重要的组成部分。为了验证其有效性,对 ESMPD-GP 算法与表 6 中的 6 种方法进行了比较。

表 6 消融实验所对比方法简介

Tab.6 Introduction to comparison methods of ablation experiments

方法	多种群策略	精英选择机制	编码方式	适应度函数
ESMPD-GP _s	只使用单种群,种群规模设置为初始两倍	无 AESE	两层 GP 树	按实验配置
ESMPD-GP _T	多种群	无 AESE,每次迭代中将多目标前 30%非支配解加入单目标种群	两层 GP 树	按实验配置

续表

方法	多种群策略	精英选择机制	编码方式	适应度函数
ESMPD-GP _N	多种群	AESE	单层 GP 树	按实验配置
ESMPD-GP ₀	多种群	AESE	两层 GP 树	多目标种群优化 recall 和 precision ^[10]
ESMPD-GP ₁	多种群	两阶段精英选择	两层 GP 树	按实验配置
ESMPD-GP ₂	随机化初始新个体替换 多目标种群	AESE	两层 GP 树	按实验配置
ESMPD-GP	多种群	AESE	两层 GP 树	按实验配置

表 7 展示了在 Anatomy 测试集上,用上述方法在 30 次独立运行 f-measure 的平均值。实验结果由高到低排列次序为: ESMPD-GP、ESMPD-GP₁、ESMPD-GP_T、ESMPD-GP₂、ESMPD-GP₀、ESMPD-GP_S、ESMPD-GP_N,验证了算法各个组件的有效性。

表 7 消融实验(括号中数值为 30 次独立运行结果的标准差)

Tab.7 Ablation experiments (values in parentheses are standard deviations of results of 30 independent runs)

方法	recall	precision	f-measure
ESMPD-GP _S	0.74 (0.03) (+)	0.95 (0.02) (+)	0.83 (0.02) (+)
ESMPD-GP _T	0.85 (0.03) (+)	0.95 (0.02) (+)	0.90 (0.01) (+)
ESMPD-GP _N	0.70 (0.03) (+)	0.96 (0.02) (+)	0.81 (0.02) (+)
ESMPD-GP ₀	0.78 (0.03) (+)	0.96 (0.01) (+)	0.86 (0.02) (+)
ESMPD-GP ₁	0.90 (0.01) (+)	0.98 (0.01) (=)	0.94 (0.01) (+)
ESMPD-GP ₂	0.83 (0.03) (+)	0.95 (0.02) (+)	0.88 (0.02) (+)
ESMPD-GP	0.93 (0.01)	0.99 (0.02)	0.96 (0.01)

ESMPD-GP₁ 所采取两阶段的选择策略,首先选择出适应度高的一半个体,在余下的个体中比较与当前代最优个体的编辑距离,保留编辑距离高的个体作为精英。该策略会淘汰适应度值低个体,仅仅保留适应度值高的个体,导致算法多样性

下降,故算法性能不如 ESMPD-GP。与 ESMPD-GP₂ 相比,该算法采用随机初始化个体进行种群多样性增强。相比起 ESMPD-GP 所采用策略,提供的解在匹配性能或者在规模上都无法保障。这些解不能为单目标种群提供更有用的知识,匹配效果不如 ESMPD-GP₁ 与 ESMPD-GP。

分析比较 ESMPD-GP_N 与 ESMPD-GP 的实验结果,前者在提取方法上仅使用了 Maximum 方法,导致算法在 recall 上剧烈下降,说明所使用方法在相似度方法的集成上仍有不足,无法识别所有实体的差异,导致无法使用简单的提取方法便得到优秀的匹配结果。同时对比所有方法的查全和查准结果,发现查全率都低于查准率,说明相似度方法的集成及结果提取精度依旧不够,相关的工作在未来工作中需要进一步进行拓展。

3.6 简化 GP 个体压力能力验证

多样性增强策略作为连接两个种群交互的重要机制,通过将多目标种群中非支配个体复制到单目标种群中,增强单目标种群多样性。为验证多样性增强策略对种群中个体规模的影响,在 Anatomy 测试集上分别测试了 ESMPD-GP_S、ESMPD-GP_T、ESMPD-GP₀ 和 ESMPD-GP。各算法的适应度评估次数与种群中所有个体的总体规模的关系如图 8 所示。对比结果表明, ESMPD-GP 的个体规模最小。特别是对于 ESMPD-GP_S 及 ESMPD-GP₀,进化过程中种群 GP 个体简化压力问题加剧,规模不断增大。ESMPD-GP_T 引入了规模较小的个体,使个体规模有所下降,但总体仍大于 ESMPD-GP。且由于 ESMPD-GP_T 没有使用 AESE 精英选择机制,难以筛选出代表解。引入的多目标种群中规模小的个体无法很好保留下

来,因此个体规模大于 ESMPD-GP。综上所述, AESE 机制有效缓解 GP 简化压力,并对算法性能产生了积极影响。

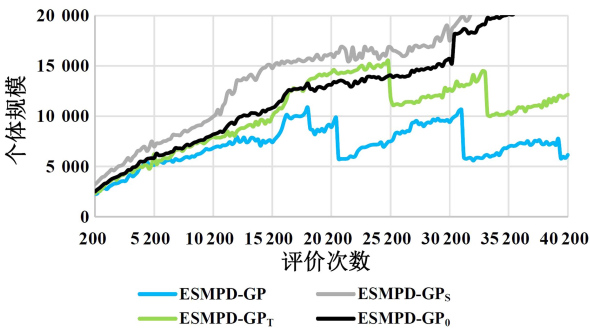


图 8 Anatomy 测试集上个体规模比较实验
Fig.8 Individual scale comparison experiment on Anatomy test set

3.7 收敛性分析

在 Anatomy 测试集上分别比较了 ESMPD-GP_s、ESMPD-GP_T、ESMPD-GP₂ 和 ESMPD-GP₄ 的适应度值随评估次数的变化,结果如图 9。由图可见,ESMPD-GP 的收敛速度最快。ESMPD-GP_s 在进化过程中长时间陷入局部最优导致算法晚于并低于 ESMPD-GP 达到最好性能。ESMPD-GP_T 增加的个体扰动使其比 ESMPD-GP_s 拥有更好性能;但单目标种群中仍可能包含冗余基因,在进化过程中干扰进化方向,从而导致算法收敛性慢于 ESMPD-GP。ESMPD-GP₂ 采用随机初始化个体替换多目标种群的非支配个体来增加单目标种群的多样性,这种不可控的多样性增强方式导致算法收敛缓慢,性能也不如 ESMPD-GP_T 及

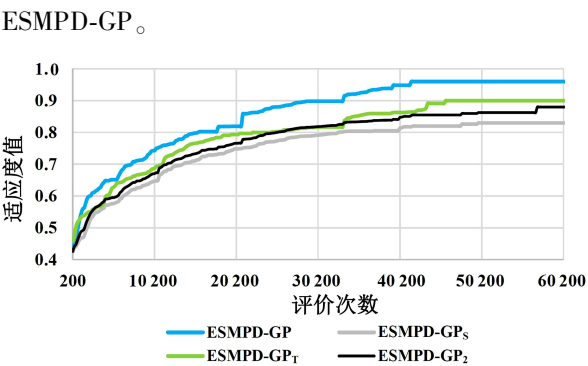


图 9 Anatomy 测试集上收敛性测试
Fig.9 Convergence test on Anatomy test set

4 结束语

本体匹配旨在处理多源异构数据,实现语义融合与语义消歧。但由于不同本体之间结构的复杂性,获取正确的匹配关系是十分困难的。为了提高本体匹配质量,本研究提出一种基于改进遗传规划算法的本体匹配方法。该方法通过新的精英选择机制挑选出多样性更强的代表解防止算法早熟收敛;所采用的多样性增强策略可以有效帮助算法增加种群多样性,并使算法跳出当前局部最优从而提升算法的整体性能。算法的多样性增强策略引入了帕累托前沿解以缓解 GP 个体简化压力,使算法探索出了个体规模更小的解。虽然该方法在匹配性能上取得了不错的结果,但是在实验中也发现其过度依赖所构建结果提取方法的正确性。未来工作将对如何提升集成相似度度量方法的精度开展进一步研究。

参考文献:

[1] IBRAHIM S,FATHALLA S,LEHMANN J,et al. Toward the multilingual semantic web:multilingual ontology matching and assessment[J]. IEEE Access,2023,11:8581-8599.

[2] XUE X S,YANG C F,JIANG C,et al. Optimizing ontology alignment through linkage learning on entity correspondences[J]. Complexity,2021,2021(1):5574732.

[3] KÖTZING T,GREGOR LAGODZINSKI J A,LENGLER J,et al. Destructiveness of lexicographic parsimony pressure and alleviation by a concatenation crossover in genetic programming[J]. Theoretical Computer Science,2020,816:96-113.

[4] 薛文,苏宏升. 基于分群策略的混沌粒子群优化算法[J]. 计算机工程与设计,2019,40(2):443-448.

[5] ABDALHAQ B,HAWASH A,AWAD A. Diversity enforced genetic algorithm (GA) for binary decision diagram (BDD) re-ordering[J]. Applied Soft Computing,2024,155:111453.

[6] MEI Z Y,LU Y J,LV L Y. Research on multi-objective low-carbon flexible job shop scheduling based on improved NSGA-II[J]. Machines,2024,12(9):590.

[7] DEB K,PRATAP A,AGARWAL S,et al. A fast and elitist multi-objective genetic algorithm:NSGA-II[J]. IEEE Transactions on Evolutionary Computation,2002,6(2):182-197.

[8] XUE X , LIN J C W , XU Z . Automatic similarity feature selection for ontology matching with semantic sampling[J] . Knowledge-Based Systems , 2024 , 302 : 112392 .

[9] BURKE E K , GUSTAFSON S , KENDALL G . Diversity in genetic programming : an analysis of measures and correlation with fitness[J] . IEEE Transactions on Evolutionary Computation , 2004 , 8 (1) : 47-62 .

[10] BOCK J , HETTENHAUSEN J . Discrete particle swarm optimisation for ontology alignment[J] . Information Sciences , 2012 , 192 : 152-173 .

[11] LIN W W , HAGA R . Matching cyber security ontologies through genetic algorithm-based ontology alignment technique[J] . Security and Communication Networks , 2021 , 2021 (1) : 4856265 .

[12] RYMA G , MOHAMED-KHIREDDINE K . Genetic algorithm with hill climbing for correspondences discovery in ontology mapping[J] . Journal of Information Technology Research (JITR) , 2019 , 12 (4) : 153-170 .

[13] MARTINEZ-GIL J , CHAVES-GONZALEZ J M . A novel method based on symbolic regression for interpretable semantic similarity measurement[J] . Expert Systems with Applications , 2020 , 160 : 113663 .

[14] XUE X S , LIN J C , SU T . Efficient ontology matching through compact linear genetic programming with surrogate-assisted local search[J] . Swarm and Evolutionary Computation , 2024 , 91 : 101758 .

[15] XUE X S , HUANG Y R , ZHANG Z Q . Deep reinforcement learning based ontology meta-matching technique[J] . IEICE Transactions on Information and Systems , 2023 , E106.D (5) : 635-643 .

[16] MARTINEZ-GIL J , YIN S Y , KUNG J , et al . Matching large biomedical ontologies using symbolic regression using symbolic regression[J] . Journal of Data Intelligence , 2022 , 3 (3) : 316-332 .

[17] MARTINEZ-GIL J , MOKADEM R , KÜNG J , et al . Neurofuzzy semantic similarity measurement[J] . Data & Knowledge Engineering , 2023 , 145 : 102155 .

[18] HERTLING S , PAULHEIM H . OLaLa : ontology matching with large language models[C] // Proceedings of the 12th Knowledge Capture Conference 2023 . Pensacola FL USA : ACM , 2023 : 131-139 .

[19] XUE X S , WU X J , CHEN J F . Optimizing biomedical ontology alignment through a compact multiobjective particle swarm optimization algorithm driven by knee solution[J] . Discrete Dynamics in Nature and Society , 2020 , 2020 : 4716286 .

[20] XUE X S . A compact firefly algorithm for matching biomedical ontologies[J] . Knowledge and Information Systems , 2020 , 62 (7) : 2855-2871 .

(责任编辑 : 方素华)