doi:10.11835/j.issn.1000.582X.2023.06.012

基于进化算法的生物医学本体匹配技术

王 颖[1],薛醒思[1],卢家伟[1],黄艺坤[2]
(1.福建工程学院 信息科学与工程学院,福州 350118; 2.福建师范大学协和学院 信息技术系,福州 350117)

摘要:由于生物医学本体拥有规模庞大的概念和复杂概念间关系,已有本体匹配技术难以高效确定生物医学本体匹配结果。为解决这一问题,构建了生物医学本体匹配问题优化模型,提出基于进化算法的生物医学本体匹配技术来确定最优匹配结果。在求解生物医学本体匹配问题时,采用一种新的生物医学本体概念相似度度量来确保匹配结果质量,并通过基于推理的概念对剪枝技术缩小算法的搜索空间,提高算法效率。实验结果表明,基于进化算法的生物医学本体匹配技术能有效匹配生物医学本体。

关键词:进化算法;生物医学本体匹配;概念对剪枝
中图分类号:TP182 **文献标志码:**A **文章编号:**1000-582X(2023)06-130-06

Evolutionary algorithm based biomedical ontology matching technique

WANG Ying[1], *XUE Xingsi*[1], *LU Jiawei*[1], *HUANG Yikun*[2]
(1. School of Information Science and Engineering, Fujian University of Technology, Fuzhou 350118, P. R. China; 2. Department of Information Technology, Concord University College Fujian Normal University, Fuzhou 350117, P. R. China)

Abstract: Since biomedical ontologies own large-scale concepts and complex relationships among them, the existing ontology matching techniques are not able to determine the biomedical alignment efficiently. To tackle this challenge, a mathematical optimal model for biomedical ontology matching problem is first constructed, and then an evolutionary algorithm (EA) based biomedical ontology matching technique is proposed to determine the optimal alignment. In particular, when solving the biomedical ontology matching problem, a novel biomedical concept similarity measure is utilized to ensure the quality of the alignment, and a reasoning-based concept pruning approach is used to reduce the algorithm's search space and improve its efficiency. The experimental results show that EA-based biomedical ontology matching technique is able to match the biomedical ontologies effectively and efficiently.

Keywords: evolutionary algorithm; biomedical ontology matching; concept pair pruning

收稿日期:2020-04-21
基金项目:国家自然科学基金资助项目(62172095);福建省自然科学基金资助项目(2020J01875);福建省本科高校教学改革研究项目(FBJG20190156)。
Supported by National Natural Science Foundation of China (62172095), National Natural Scientific Foundation of Fujian Provnce (2020J01875) and Fujian Province Undergraduate Universities Teaching Reform Research Project (FBJG20190156).
作者简介:王颖(1980—),高级实验师,主要从事智能计算和本体匹配技术方向研究,(E-mail)2020774@qq.com。
通信作者:薛醒思,教授,(E-mail)jack8375@gmail.com。

生物医学本体是对生物医学领域中存在的概念、实例及它们之间关系的规范化描述，使基于生物医学知识的智能系统之间准确理解彼此数据的真实含义，在语义层面上实现系统间的交互与协作[1-2]。近年来，生物医学本体被广泛应用在诸如病历的语义标注[3]、医学数据格式标准化[4]、医疗知识表示和共享[5]、临床数据集成和辅助诊疗等[6]应用领域。为满足不同领域需求，本体工程师开发了如基因本体(gene ontology, GO)[7]、人类表型本体(human phenotype ontology, HPO)[8]、国家癌症研究术语本体(national cancer institute thesaurus, NCI)[9]和医学系统术语本体(systemized nomenclature of medicine, SNOMED-CT)[10]等众多生物医学本体。由于本体工程师们对于客观事物的认识、描述角度各不相同，导致不同生物医学本体之间存在严重异质问题，阻碍生物医学智能系统间的交互与协作。

生物医学本体匹配技术可通过确定本体中异质概念间的对应关系来解决生物医学本体异质问题。AgreementMakerLight[11]、YAM-BIO[12]、XMap[13]和 LogMapBio[14]等目前已有的本体匹配技术在求解生物医学本体匹配问题时需要消耗大量时间且无法保证匹配结果质量。因此，如何有效识别异质的生物医学概念、提高生物医学本体匹配过程效率是求解生物医学本体匹配问题的关键。为有效且高效求解生物医学本体匹配问题，研究构建了生物医学本体匹配问题优化模型，并利用进化算法[15]确定最优匹配结果。笔者采用一种新的生物医学本体概念相似度度量技术确保匹配结果质量，通过基于推理的概念对剪枝技术来缩小算法的搜索空间并提高算法效率。

1　生物医学本体匹配问题

生物医学本体是生物医学概念及概念间关系集合，生物医学本体匹配结果是2个本体中语义相同的概念对集合。本体匹配结果的质量通常利用查全率、查准率和F度量[16]来评价，但需要专家提供标准的本体匹配结果。由于生物医学中的概念规模庞大，专家无法事先提供标准本体匹配结果，笔者提出一种近似度量技术评价生物医学本体匹配结果质量。通过实验观察发现，生物医学本体匹配结果的质量同匹配结果中的概念对数量和平均相似度值成正比。给定一个生物医学本体匹配结果A，提出如下公式近似评价生物医学本体匹配结果质量

$$f(A)=2\times\frac{r(A)\times p(A)}{r(A)+p(A)}, \tag{1}$$

式中：$r(A)=\frac{|A|}{\mathrm{M}}$；$|A|$是$A$中概念匹配对的数量；$M$是大的正整数；$p(A)=\frac{\sum \mathrm{sim}(a_i)}{|A|}$，$\mathrm{sim}(a_i)$表示$A$中第$i$个概念对的相似度值。在此基础上，给定2个生物医学本体O_1和O_2，生物医学本体匹配问题的数学优化模型定义如下

$$\begin{cases}\max \quad f(\boldsymbol{X})\\ \text{s.t. } \boldsymbol{X}=\left(x_1,x_2,\cdots,x_{|O_1|}\right)^{\mathrm{T}},\\ \quad x_i\in\left\{0,1,2,\cdots,|O_2|\right\}\end{cases} \tag{2}$$

式中：$|O_1|$和$|O_2|$分别表示O_1和O_2中的概念数量；$x_i=j,j=1,2,\cdots;|O_2|$表示O_1中第i个概念同O_2中第j个概念形成概念对（若$x_i=0$，则O_1中第i个概念没有匹配上任何一个概念）；$\boldsymbol{X}$表示一个本体匹配结果，该模型的目标是最大化$f(\boldsymbol{X})$的值。

2　生物医学概念相似度度量技术

概念相似度度量技术是本体匹配技术的基础，生物医学概念的异质性高、专业性强、结构复杂，因此已有概念相似度度量技术难以有效识别语义相同的生物医学概念。在基于概念名称、背景知识库和本体概念体系关系结构这三类相似度度量技术基础上，提出混合度量技术以识别异质的生物医学概念。给定2个生物医学概念c_1和c_2，利用本体概念体系关系结构获取二者直接的子概念集合C_1和C_2，分别抽取出C_1和C_2中所有概念的名称和属性名称构建二者对应的信息档案p_1和p_2，通过其对应的信息档案p_1和p_2的相似度值来度量c_1和c_2的相似程度，相关的计算公式如下

$$\mathrm{sim}(c_1,c_2)=\frac{\sum_{i=1}^{|p_1|}\max_{j=1,2,\cdots,|p_2|}\left(\mathrm{sim}'(p_{1i},p_{2j})\right)+\sum_{j=1}^{|p_2|}\max_{i=1,2,\cdots,|p_1|}\left(\mathrm{sim}'(p_{1i},p_{2j})\right)}{|p_1|+|p_2|},\tag{3}$$

式中：$|p_1|$和$|p_2|$分别是p_1和p_2中元素的个数；p_{1i}和p_{2j}分别是p_1和p_2中第i个和第j个元素。当p_{1i}和p_{2j}在生物医学知识库Unified Medical Language System (UMLS)[16]中是同义词时，$\mathrm{sim}'(p_{1i},p_{2j})=1$，否则$\mathrm{sim}'(p_{1i},p_{2j})=\text{N-gram}(p_{1i},p_{2j})$，其中N-gram距离[17]是用于度量生物医学概念名称编辑距离最有效技术。

3 求解生物医学本体匹配问题的进化算法

生物医学本体匹配问题是一个复杂的大规模优化问题，进化算法具有全局寻优能力、自动获取和指导优化搜索空间并自适应调整搜索方向，是求解该问题的有效方法。提出的用于求解生物医学本体匹配问题的进化算法框架如表1所示。

表1 进化算法框架

Table 1 The framework of Evolutionary Algorithm

```
t = 0; //初始化进化代数
initialize the Population P_t; //初始化种群
evaluate(P_t); //评价种群
while t < t_max
    S_t = select(P_t); //选择操作
    C_t = crossover(S_t); //交叉操作
    P_{t+1} = mutation(C_t); //变异操作
    evaluate(P_{t+1});
    save elite(); //保留精英个体
    if elite is updated
      pruning mapping pairs; //概念对剪枝
    end if
    t = t+1;
end while
output elite
```

该算法初始化进化代数t并随机初始化种群P_t，对种群中每个个体的质量进行评价；在每一代的进化过程中，通过赌轮盘方法来选出新一代种群，依据交叉概率对种群中的个体执行单点交叉操作以实现个体间的信息交换，依据变异概率对种群中的个体执行位点变异操作以保证种群多样性；最后更新精英个体（历史最优解）并将精英个体取代种群中适应度值最低的个体以保证精英个体不会在进化过程中丢失，当精英个体被更新后，算法依据新的精英个体信息对概念进行剪枝以缩小算法的搜索区域，当算法进化到最大代数$t_{\max}$后终止，输出精英个体elite。

3.1 编码机制

假设$|C_1|$和$|C_2|$分别是2个生物医学本体中概念集C_1和C_2中元素的个数，进化算法中的每个个体可表示为长度为$|C_1|$的一维数组$N_1N_2\cdots N_{|C_1|}$，其中，$N_i\in\{0,1,2,\cdots,|C_2|\}$。当$N_i=j\in\{1,2,\cdots,|C_2|\}$时，表示$C_1$中的第$i$个概念同$C_2$中的第$j$个概念匹配上；当$N_i=0$时，表示$C_1$中的第$i$个概念没有匹配上$C_2$中的任何一个概念。

3.2 基于推理的生物医学概念对剪枝

针对大规模本体匹配问题，目前是通过本体划分算法将大规模生物医学本体划分为若干本体分块，问

题转化为等价的若干个小规模的本体分块匹配问题。本体划分算法存在3个局限：1) 本体划分算法的时空复杂度同后续大规模本体匹配算法的时空复杂度一样，无法从本质上提高匹配过程的效率；2) 本体划分算法无法控制本体分块规模，使本体分块的规模不是太大就是太小，使得匹配过程效率不高； 3) 本体划分算法会导致位于分块边缘概念丢失一定程度的语义信息，使本体匹配结果质量不高。为缩小匹配过程中进化算法的搜索空间，提出一种基于推理的生物医学概念对剪枝方法，利用生物医学本体的概念体系结构减少匹配过程中所需的概念相似度值计算次数，提高生物医学本体匹配过程效率。

通过实验发现：1) 生物医学本体的概念体系结构通常是通过“is-a”和“part-of”关系来构建的，正确的匹配结果同该体系结构一致；2) 生物医学本体在某个区域内的大部分概念会同另一个生物医学本体在某个区域的概念匹配。在此基础上，假设$\left(c_i,c_j\right)$是精英个体中确定的一个拥有高概念相似度值的生物医学概念对，则c_i所有的直接子概念（或父概念）同c_j所有的直接父概念（或子概念）为不相似概念，即将c_j所有的直接父概念（或子概念）编号从c_i所有的直接子概念（或父概念）对应基因位可行域中移除。

4　实验结果与分析

实验中采用国际本体匹配竞赛(ontology alignment evaluation initiative, OAEI)提供的Anatomy测试数据集和Large Bio测试数据集。表2-3分别给出了研究方法的匹配结果（30次独立运行的均值）和OAEI参与者的匹配结果。进化算法采用的配置如下：种群规模=100，交叉概率=0.85，变异概率=0.02，最大进化代数=3 000，相似度值阈值=0.95。算法的配置是通过实验确定的，该配置可以确保获取的匹配结果质量最好。

4.1　Anatomy测试数据集

Anatomy要求是将成年鼠类解剖学本体（2 744个概念）同NCI中的人类解剖学本体（3 304个概念）进行匹配。从表2中可以看出，研究方法获取的匹配结果查全率、查准率和F度量值明显高于其他前沿本体匹配技术。研究方法的查全率最高，这说明提出的概念对剪枝可以在保证生物医学本体匹配结果质量前提下提高本体匹配过程的效率（研究方法的运行时同LogMap并列第1）。最后，方法的标准差较小，说明其稳定性较好。

表2　在Anatomy测试数据集上同OAEI参与者的比较

Table 2　Comparison among our approach and OAEI participants on Anatomy

匹配技术	查全率(标准差)	查准率(标准差)	F度量(标准差)	运行时/s (标准差)
AML	0.93 (0.0)	0.95 (0.0)	0.94 (0.0)	47 (0.0)
CroMatcher	0.90 (0.0)	0.94 (0.0)	0.92 (0.0)	573 (0.0)
Xmap	0.86 (0.0)	0.92 (0.0)	0.89 (0.0)	45 (0.0)
LogMapBio	0.89 (0.0)	0.88 (0.0)	0.89 (0.0)	758 (0.0)
FCA_Map	0.83 (0.0)	0.93 (0.0)	0.88 (0.0)	117 (0.0)
LogMap	0.84 (0.0)	0.91 (0.0)	0.88 (0.0)	**24** (0.0)
LYAM	0.87 (0.0)	0.86 (0.0)	0.86 (0.0)	799 (0.0)
Lily	0.79 (0.0)	0.87 (0.0)	0.83 (0.0)	272 (0.0)
LPHOM	0.72 (0.0)	0.70 (0.0)	0.71 (0.0)	1601 (0.0)
本文的方法	**0.94** (0.01)	**0.97** (0.01)	**0.96** (0.01)	**24** (3.0)

4.2　Large Biomed测试数据集

Large Biomed包含了3个任务，要求匹配3个大规模生物医学本体FMA（78 989个概念）、SNOMED CT（306 591个概念）和NCI（66 724个概念）。从表3中可以看出，方法在3个任务中的查准率和F度量值都最高，查全率和运行时在2个任务中最好，相应的标准差值较小。基于进化算法的生物医学本体匹配技术可以比前沿的生物医学本体匹配技术更有效匹配生物医学本体。

表3 在Large Biomed测试数据集上同OAEI参与者的比较

Table 3 Comparison among our approach and OAEI participants on Large Biomed

任务1: FMA vs NCI				
匹配技术	查全率	查准率(标准差)	*F*度量(标准差)	运行时/s (标准差)
Xmap	0.85 (0.0)	0.90 (0.0)	0.87 (0.0)	116 (0.0)
AML	**0.87** (0.0)	0.84 (0.0)	0.85 (0.0)	72 (0.0)
LogMap	0.80 (0.0)	0.85 (0.0)	0.83 (0.0)	80 (0.0)
LogMapBio	0.84 (0.0)	0.82 (0.0)	0.83 (0.0)	1,188 (0.0)
研究方法	**0.87** (0.01)	**0.92** (0.02)	**0.90** (0.01)	**65** (5.0)
任务2: FMA vs SNOMED				
匹配技术	查全率(标准差)	查准率(标准差)	*F*度量(标准差)	运行时/s (标准差)
Xmap	0.84 (0.0)	**0.97** (0.0)	0.90 (0.0)	366 (0.0)
AML	0.69 (0.0)	0.88 (0.0)	0.77 (0.0)	**166** (0.0)
LogMap	0.63 (0.0)	0.84 (0.0)	0.72 (0.0)	433 (0.0)
LogMapBio	0.64 (0.0)	0.81 (0.0)	0.71 (0.0)	2,156 (0.0)
研究方法	0.86 (0.02)	**0.97** (0.01)	**0.92** (0.01)	181 (25.0)
任务3: NCI vs SNOMED				
匹配技术	查全率(标准差)	查准率(标准差)	*F*度量(标准差)	运行时 /s (标准差)
AML	**0.67** (0.0)	0.90 (0.0)	0.77 (0.0)	376 (0.0)
LogMapBio	0.64 (0.0)	0.84 (0.0)	0.72 (0.0)	4,322 (0.0)
Average	0.62 (0.0)	0.85 (0.0)	0.72 (0.0)	1,353 (0.0)
LogMap	0.60 (0.0)	0.87 (0.0)	0.71 (0.0)	699 (0.0)
研究方法	0.65 (0.03)	**0.93** (0.02)	**0.81** (0.02)	**286** (23.0)

5 总 结

生物医学本体匹配技术能确定不同生物医学本体中异质概念，实现基于本体的生物医学智能系统之间协作。研究提出一种基于进化算法的生物医学本体匹配技术求解该问题，并确定最优的本体匹配结果。在算法求解过程中，采用新的生物医学概念相似度度量和基于推理的概念对剪枝来提高算法性能。实验结果表明，基于进化算法的本体匹配技术能够有效匹配生物医学本体。

参考文献

[1] 邱实. 基于领域本体的生物医学本体匹配算法研究[D]. 哈尔滨: 哈尔滨工业大学 2015.
Qiu S. Research on biomedical ontology matching algorithm based on domain ontology[D]. Harbin: Harbin Institute of Technology, 2015. (in Chinese)

[2] Chatterjee N, Kaushik N, Gupta D, et al. Ontology merging: a practical perspective[C]//International Conference on Information and Communication Technology for Intelligent Systems. Cham: Springer, 2018: 136-145.

[3] Yan S K, Wong K C. Elucidating high-dimensional cancer hallmark annotation via enriched ontology[J]. Journal of Biomedical Informatics, 2017, 73: 84-94.

[4] Ping P P, Hermjakob H, Polson J S, et al. Biomedical informatics on the cloud: a treasure hunt for advancing cardiovascular medicine[J]. Circulation Research, 2018, 122(9): 1290-1301.

[5] Strang J F, Meagher H, Kenworthy L, et al. Initial clinical guidelines for Co-occurring autism spectrum disorder and gender dysphoria or incongruence in adolescents[J]. Journal of Clinical Child & Adolescent Psychology, 2018, 47(1): 105-115.

[6] Heringa M, Floor-Schreudering A, De Smet P A G M, et al. Clinical decision support and optional point of care testing of renal function for safe use of antibiotics in elderly patients: a retrospective study in community pharmacy practice[J]. Drugs & Aging, 2017, 34(11): 851-858.

［7］Consortium T G O. Expansion of the gene ontology knowledgebase and resources[J]. Nucleic Acids Research, 2017, 45(D1): D331-D338.

［8］Taboada M, Rodriguez H, Gudivada R C, et al. A new synonym-substitution method to enrich the human phenotype ontology [J]. BMC Bioinformatics, 2017, 18(1): 446.

［9］Zheng L, Min H, Chen Y, et al. Auditing National Cancer Institute thesaurus neoplasm concepts in groups of high error concentration[J]. Applied Ontology, 2017, 12(2): 113–130.

［10］Sanz X, Pareja L, Rius A, et al. Definition of a SNOMED CT pathology subset and microglossary, based on 1.17 million biological samples from the Catalan Pathology Registry[J]. Journal of Biomedical Informatics, 2018, 78: 167-176.

［11］Cruz I F, Palmonari M, Caimi F, et al. Building linked ontologies with high precision using subclass mapping discovery[J]. Artificial Intelligence Review, 2013, 40(2): 127-145..

［12］Duchateau F, Bellahsene Z. YAM: A step forward for generating a dedicated schema matcher[J]. Transactions on Large-Scale Data-and Knowledge-Centered Systems XXV, 2016: 150-185.

［13］Djeddi W E, Yahia S B, Nguifo E M. A novel computational approach for global alignment for multiple biological networks[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2018, 15(6): 2060-2066.

［14］Harrow I, Jiménez-Ruiz E, Splendiani A, et al. Matching disease and phenotype ontologies in the ontology alignment evaluation initiative[J]. Journal of biomedical semantics, 2017, 8: 1-13.

［15］Rudolph G. An evolutionary algorithm for integer programming[C]//Parallel Problem Solving from Nature — PPSN III. Berlin, Heidelberg: Springer Berlin Heidelberg, 1994: 139-148.

［16］Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology[J]. Nucleic Acids Research, 2004, 32: D267-D270.

［17］Kondrak G. N-gram similarity and distance[C]//String Processing and Information Retrieval. Berlin, Heidelberg: Springer, 2005: 115-126.

（编辑　侯　湘）