第35卷 第17期 计 算 机 工 程 2009年9月
Vol.35 No.17 Computer Engineering September 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)17—0175—03 文献标识码：A 中图分类号：TP311.5

基于遗传禁忌算法的Ontology划分

李 广，谢 强，丁秋林
(南京航空航天大学信息科学与技术学院，南京 210016)

摘 要：为解决企业实际应用中需要部分使用本体(Ontology)的问题，提出一种基于遗传禁忌算法的Ontology自动划分方法。按Ontology划分的要求，将概念被划分进的子Ontology编号组成的数字串作为一条染色体，设计遗传禁忌算法的适应度函数，给出Ontology划分算法的具体步骤。对比实验结果表明，该方法的划分平衡度和准确性优于其他方法。

关键词：本体；遗传禁忌算法；划分

Ontology Partition Based on Tabu and Genetic Algorithm

LI Guang, XIE Qiang, DING Qiu-lin
(College of Information Science & Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016)

【Abstract】In order to solve the problem of using partial content of the huge Ontology effectively, a method of ontology partition is proposed based on Tabu and Genetic Algorithm(TGA). In this method, the digital sequence composed of the number of the sub-Ontology whose concepts will be allocated in is taken as a chromosome according to Ontology partition request. The fitness degree function of TGA is designed, and the concrete steps of Ontology partition algorithm are presented. According to the correlation contrast experiment, it finds that its division balance degree and accuracy are better than other methods.

【Key words】Ontology; Tabu and Genetic Algorithm(TGA); partition

1 概述

现代企业中存在着大量的零散知识源，它们以不同的形式存在于生产的各个环节中，格式混乱，无法进行有效的抽取和查询。构建企业Ontology是解决这一问题的有效方法。但创建企业Ontology时总希望模型建立得尽量准确和完整，这往往导致Ontology的规模巨大。而实际应用中，人们往往主要关注某一方面的工作，只使用部分的Ontology，此时如果还是使用整个Ontology进行查询处理，会大大影响Ontology的使用效率，因此，在企业的实际应用中对规模大的Ontology进行划分，形成各个子领域的局部Ontology以便操作是一个亟待解决的问题。

近来，Ontology划分逐渐成为研究热点，主要包括3种方法：(1)网络图划分方法，代表有Stuckenschmidt等研究的基于层次机制的划分方法[1]，这种方法在计算概念相关性时，仅仅考虑Ontology中的概念层次关系，没有考虑其他语义关系，因此，不适用于实际划分；(2)基于查询的方法，代表有SparQL查询语言[2]，这种方法在划分时对概念间的语义关系考虑不够，因此，只适用于小规模Ontology的简单划分；(3)遍历划分，文献[3]使用了这种方法，但利用这种方法生成的多个子Ontology中概念平衡度不高，而且没有利用子Ontology之间的耦合度来优化划分效果，实际使用中也需要用户做出许多手动选择。

2 Ontology划分

定义1 Ontology为元组 $O=(C, Root, A^C, R^C, I, R^I)$，其中，$C$表示概念的集合；$|C|$表示概念的数目；$Root$表示Ontology中的根概念，一个Ontology有且只有一个根概念；A^C表示概念$c \in C$的属性集合；$|A^C|$表示概念属性的数目；R^C表示概念之间关系的集合；$R^C=\{r(c_i,c_j,\lambda_{ij})|c_i \in C, c_j \in C, i \neq j\}$；$\lambda_{ij}$为关系的权重，不同性质的关系都赋予不同的权重，层次关系、恒等关系、函数关系、逆关系、层次继承关系等重要关系被赋予高权重，使之在划分时被优先保留；I表示概念实例集合；概念$c \in C$的实例集合记为$I(c)$；$|I(c)|$表示概念实例的数目；R^I是不同实例之间的关系组成的实例关系集合：$R^I: I \times I$。

Ontology的划分就是按照要求把一个规模大的Ontology划分成几个规模小的Ontology，本文将划分后的Ontology称为子Ontology。

定义2 子Ontology表示为元组 $O_S=(C_S, Root_S, A^{Cs}, R^{Cs}, I_S, R^{Is})$，且 $C_S \neq \phi$, $C_S \subseteq C$, $A^{Cs} \subseteq A^C$, $R^{Cs} \subseteq R^C$, $I_S \subseteq I$, $R^{Is} \subseteq R^I$，$Root_S$为O_S的唯一根节点[4]。

定义3 将Ontology划分为多个子Ontology $O_{S1}, O_{S2}, \cdots, O_{Sn}$，如果它们满足：

(1) $C_{S1} \cup C_{S2} \cup \cdots \cup C_{Sn} \subseteq C$

(2) $R_{S1} \cup R_{S2} \cup \cdots \cup R_{Sn} \subseteq R$

(3) $\forall i, j, i \neq j$, $C_{Si} \cap C_{Sj} = \phi$

(4) $\forall i, j, i \neq j$, $R_{Si} \cap R_{Sj} = \phi$

则称$O_{S1}, O_{S2}, \cdots, O_{Sn}$为一个划分。

3 基于遗传禁忌算法的Ontology划分方法

3.1 遗传禁忌算法

遗传算法(Genetic Algorithm, GA)是基于“适者生存”的一种随机、自适应的优化算法。其显著特点是全局解空间搜索，但存在进化缓慢、算法参数敏感等缺点。文献[5]提出了遗传禁忌算法(Tabu and Genetic Algorithm, TGA)，给出了一

作者简介：李 广(1981－)，男，硕士研究生，主研方向：主动知识服务，本体划分；谢 强，副教授、博士；丁秋林，教授、博士生导师

收稿日期：2009-02-23 **E-mail：**andrewli2004@163.com

个算法框架。本文在此基础上，采用遗传禁忌算法进行 Ontology 划分。同时为了保证划分的精确度，对遗传禁忌算法进行了一些改进。

3.2 基于 TGA 的 Ontology 划分方法

Ontology 划分实质上是决定每个概念该被分配到哪个子 Ontology 中的过程，这是一个离散空间问题的求解过程。由于二进制编码的遗传禁忌算法不足以描述 Ontology 划分问题，概念所在的位置不能用 0 或 1 表示，而是用子 Ontology 的标号表示，因此本文采用了实数编码的遗传禁忌算法。

划分 Ontology 的一个解(染色体)X 表示为一个整数串，$X=x_1x_2\cdots x_n$，其中，整数 $x_i\in[1,k]$, $1\leqslant k\leqslant n$；k 为将要划分成的子 Ontology 个数；n 为 Ontology 中概念的个数；x_i ($1\leqslant i\leqslant n$)则表示第 i 个概念被划分进第 x_i 个子 Ontology 中。根据经验，取种群数为染色体长度的 2 倍，则划分的种群为 $C=(C_1,C_2\cdots,C_{2n})$。

子 Ontology 个数 k 的确定有 2 种策略: (1)由构建企业本体的领域专家对企业知识进行分类，利用分类对企业 Ontology 进行初步划分，同时在子 Ontology 的应用中由参与人员提出分类修正；(2)采用神经网络等智能算法对企业核心概念进行预分类，确定分类数。出于企业实际应用的需要，本文使用前一种方法。

3.2.1 算法的适应度函数

在进化过程中，评价每条染色体的适应度并记录整个种群的全局最优解是算法的关键之一。本文采用划分后形成的子 Ontology 内部的内聚度和子 Ontology 之间的耦合度进行评价。子 Ontology 内部的内聚度越高，子 Ontology 之间的耦合度越低，划分的效果越好。

为了计算子 Ontology 的内聚度，采用概念图的方法，把子 Ontology 抽象成一张概念图。并以初始构建 Ontology 时创建的关系权重作为有向边的权值，不同的关系有不同的权重，如层次关系权重为 10，恒等关系为 5，函数关系为 3。

定义 4 Ontology M 的概念图模型表示为一个三元组(O_M, N_M, A_M)，O_M 代表本体自身；N_M 代表一个节点集合，其中，每个节点代表 Ontology 中的一个概念；A_M 代表图中有向边的集合，$A_M=\{r(n^i_M, n^j_M, rw), n^i_M\in N_M, n^j_M\in N_M, i\neq j\}$；每个有向边代表一个关系；$rw$ 为有向边的权值，如果节点为孤立点，则权值 rw 设为 0。

定义 5 Ontology M 的内聚度 $f(O_M)$为 M 内部关系权重的总和，即 $f(O_M)=\sum rw$。

定义 6 一个 Ontology 划分的内聚度 $f(X_i)$可定义为各个子 Ontology 的内聚度总和的倒数，即

$$f(C_i)=\frac{1}{\sum_{i=1}^{k} f(O_i)}$$

定义 7 Ontology 划分的耦合度为子 Ontology 间、概念与概念之间相似度的总和，即 $g(X_i)=\sum Sim(O_i: C_n, O_j: C_m)$, $C_n\in O_i$, $C_m\in O_j$, $0<i\neq j<k$， X_i 是 Ontology 的一个划分。

Ontology 划分效果随 $f(X_i)$升高而升高，随 $g(X_i)$升高而降低，因此，必须综合考虑这 2 个指标，形成 TGA 的适应度函数，即 $Fit(X_i)=f(X_i)-g(X_i)$。

3.2.2 算法步骤

用于 Ontology 划分的 TGA 算法具体如下:

(1)初始化参数。设置种群大小 P (解的个数)、杂交率 α、变异率 β、禁忌表长度 ω、划分适应度期望值 Y 以及达不到期望值时的最大迭代次数 τ (终止准则)，迭代次数 $t=0$。

(2)随机产生初始种群 X，种群中的每条染色体 X_i 代表一个可能的解($x_1x_2\cdots x_n$)，即 n 个概念被划分进的子 Ontology 编号的集合。

```
For i=1 to 2n
    For j=1 to n // n 为 Ontology 概念的个数
      生成随机整数  x∈[1,k],1≤k≤n
    Next  j
     得到一个染色体
Next i
得到初始化的种群 X=(X1,X2,⋯,X2n)
```

(3)计算适应度。求出种群中每条染色体的适应度，即应用此染色体中的划分原则后，生成的子 Ontology 的内聚度和耦合度之和 $Fit(X_i)$，并记录最佳适应度的染色体 X_{BEST} 和适应度总和 $Fit(X)_{TOTAL}$。

```
For each Xi in X
    计算  Fit(Xi)
    If Fit(Xi)≥Fit(Xi)MAXthen
     Fit (Xi) MAX = Fit(Xi)
Fit(X)TOTAL = Fit(X)TOTAL+ Fit(Xi)
    End if
End For
```

(4)运用选择算子选出 2 个父代染色体，把 2 个父代染色体进行交叉变异，产生一个新的后代，即一个新的 Ontology 划分，计算这个后代的适应度。检查父代染色体，如果同时满足：1)2 个父代染色体的标识符不相等；2)2 个父代染色体的标识符不在对方的 Tabu list 中，或者在 Tabu list 中且后代染色体的适应度大于目前最好的适应度(藐视准则)，则此后代被加入到子种群中，父代 2 个染色体把对方的标识符放入自己的 Tabu list 中，并且让后代继承父代中一个染色体的标识符和 Tabu list。重复第(4)步，直到子种群满。为了防止始终没有合格的 2 个父代染色体被选出并且子种群未满而导致死循环，利用 regeneration 算子直接把其中一个父辈变异，即产生一个新的 Ontology 划分，并赋予其新的标识符，然后再次重复第(4)步，直到子种群满。

```
While population size p not filled do
n =n+1
XMUN = select from X
XDAD = select from X
XBABY = Crossover(XMUM, XDAD)
Mutation(XBABY)，求 Fit(XBABY)
  If XMUN.id ≠ XDAD.id AND
(Fit(XBABY)≥Fit (Xi) MAX OR not in tabu list)then
  XBABY 放入子种群
   XMUN, XDAD 的标识符放入对方的 Tabu list，并让 XBABY 继承
  end if
If n>deadlock then
Regeneration()
End if
End while
```

(5)把子种群替换父种群，$t=t+1$，检查迭代终止条件，如果 $Fit(X_i)<Y$ 且 $\tau<t$，则重复步骤(3)～步骤(5)，否则算法结束。

迭代结束后，全局最佳适应度的染色体 X_{BEST} 就是 Ontology 划分的求解结果，一般情况下，还需要领域专家对

自动划分求解结果进行评估和修正。

4 试验及结果分析

为了验证本文算法的有效性，选择了某航空研究所构建的质量控制 Ontology 作为试验对象，实验数据见表 1。

表 1 实验数据

名称	概念数目	实例数目	层次关系数目	最大层次	语义关系数目
质量控制 Ontology	657	2 612	856	15	246

采用 Matlab R2006 实现本文算法，并在 2.66 GHz Celeron、512 MB 内存上做实验，操作系统为 Windows XP SP2。

根据经验，设定 TGA 算法的参数为 P =1 314, α =1, β =0.05, ω =263，让算法循环 5 000 代。首先采用公认比较好的遍历划分法划分质量控制 Ontology，再采用本文方法划分，使 2 种方法的划分数目相等，以方便比较。获得的划分结果见表 2、表 3。

表 2 遍历划分结果

核心概念	划分中概念个数
试验管理	48
型号	242
供应商	20
扩散企业	24
人员	12
部件	193
质量文件	6
质量归零	37
质量考核	21
制造工艺	54

表 3 TGA 划分结果

子划分编号	划分中概念个数
1	56
2	74
3	33
4	67
5	102
6	80
7	98
8	58
9	45
10	44

从表 2、表 3 可以明显看出，使用本文方法划分 Ontology，其平衡度大于遍历划分法，这是因为本文提出的 TGA 划分方法充分考虑了子 Ontology 内部语义关系的内聚度和子 Ontology 之间概念的相似度，而遍历划分法只是从一个核心概念开始不断寻找与其相邻的概念，以人工设定的寻找深度作为终止准则，没有充分考虑子 Ontology 之间的相似度。

为了检测算法的效率，对本文算法与 GA 进行了比较，设定 TGA 算法的参数为 P =1 314, α =1, β =0.05, ω =263，让算法循环 500 代。GA 算法的参数设置为种群数 1 314，杂交率 0.8，变异率 0.02，也让算法循环 500 代。从图 1、图 2 可以看出，GA 的种群多样性明显低于 TGA 算法，并且在算法运行过程中逐渐降低，以致无法收敛到最优解，而 TGA 算法的计算性能和寻找最优解的能力均优于 GA。

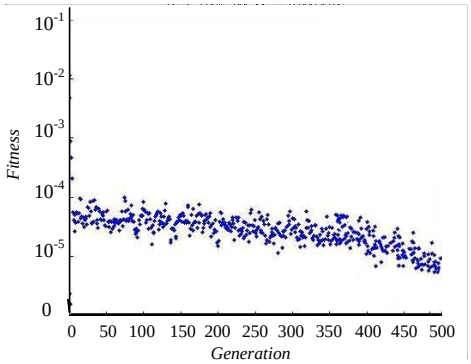

图 1 GA 算法的适应度

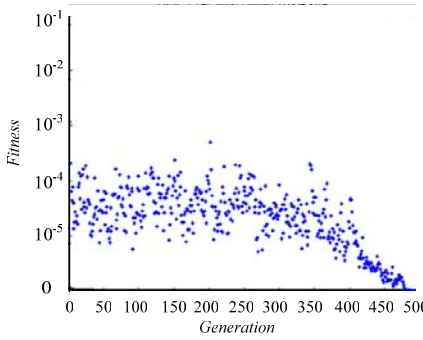

图 2 TGA 算法的适应度

5 结束语

本文提出了一种基于禁忌遗传算法的 Ontology 划分方法。实验表明，用该方法划分 Ontology 相比其他算法实际应用价值更高，有很好的平衡性和划分准确度。本文在 Ontology 划分中没有进一步考虑划分以后子 Ontology 之间的协同进化，这是今后需要研究的问题。

参考文献

[1] Stuckenschmidt H, Klein M. Structure-based Partitioning of Large Concept Hierarchies[C]//Proc. of the 3rd International Semantic Web Conference. Hiroshima, Japan: Springer, 2004.

[2] Seaborne A, Hommeaux E. SparQL Query Language for RDF[EB/OL]. [2008-06-02]. http://www.w3.org/TR/rdf-sparql-query/.

[3] Seidenberg J, Rector A. Web Ontology Segmentation: Analysis, Classification and Use[C]//Proc. of the 15th International Conference on World Wide Web. Edinburgh, UK: Springer, 2006: 13-22.

[4] 谢 强, 张 磊, 周 良. 基于改进粒子群优化算法的 Ontology 划分方法[J]. 华南理工大学学报: 自然科学版, 2007, 35(9): 118-122.

[5] Ting Chuankang, Lee Chungnan, Li Shengtun. TGA: A New Integrated Approach to Evolutionary Algorithms[C]//Proceedings of the 2001 Congress on Evolutionary Computation. Seoul, South Korea: Springer, 2001: 917-924

编辑 张正兴