动态多策略本体映射与集成方法研究

李 鹏,徐德智,尹 艳
LI Peng,XU De-zhi,YIN Yan

中南大学 信息科学与工程学院,长沙 410083
College of Information Science and Engineering,Central South University,Changsha 410083,China
E-mail:lp_4555015@hotmail.com

LI Peng,XU De-zhi,YIN Yan.Research on method of dynamic multi-strategies ontology mapping and integration.Computer Engineering and Applications,2009,45(30):150-153.

Abstract: Various problems have existed in the current multi-strategies ontology mapping.For example,most mapping strategies cannot do different operations on ontology mapping according to the differences existed in ontologies and the semantic information of the ontology is not made to full use when integrating each strategy.This paper proposes a dynamic method of multi-strategies ontology mapping and integration.In the process of mapping,this method improves some key strategies,then the different mapping strategies are assembled by the AHP.The experimental results show that this method improves the recall and precision of ontology mapping while maintaining currency and stability.

Key words: ontology mapping;multi-strategies;Analytic Hierarchy Process(AHP);integration

摘 要:针对目前多策略本体映射中各种映射策略不能根据待映射本体间的差异进行不同的映射处理、多策略集成时也没有充分利用本体包含的语义信息等问题。提出了一种动态多策略本体映射与集成方法,该方法首先对部分关键策略进行了动态地改进,然后利用层次分析法对不同的映射策略进行集成,从而得到最终的映射结果。实验结果表明,该方法在保证通用性和稳定性的同时,提高了映射结果的查全率和查准率。

关键词:本体映射;多策略;层次分析法;集成
DOI:10.3778/j.issn.1002-8331.2009.30.046 **文章编号:**1002-8331(2009)30-0150-04 **文献标识码:**A **中图分类号:**TP18

1 引言

语义 Web 的发展导致本体数量激增，然而由于本体的创建者不同、使用的建模方法不同,不同的领域专家开发出来的本体必然存在着差别。本体映射的目的就是要找到本体之间的语义联系,以便于知识共享和重用。而实现本体映射的关键是计算本体间实体的语义相似度。目前,大多数本体映射系统都综合利用多策略[1]来发现映射,但是各种映射策略都还存在着不足，在对多策略进行集成时也不能充分利用本体语义信息,使得各种策略的价值有所降低,影响了映射结果的查全率和查准率,因此还不能满足大多数映射任务和语义检索的需求。为了弥补以上缺陷，提出了一种动态多策略本体映射与集成方法,该方法在传统方法的基础上对各种策略进行了改进,并依据本体包含的语义信息,利用层次分析法对各策略的重要性进行赋权,以进行多策略的集成,得到最终的映射结果。实验结果验证了该方法的有效性。

2 相关工作

2.1 传统的多策略映射方法

随着映射思想的提出,多策略映射已成为该领域研究的热点。由于实际的待映射本体常常包含多种信息,将几种策略集成使用可以更充分的利用本体信息,因此往往会比采用单一策略产生更好的结果;另外,由于本体构建者的习惯、水平、拥有的资源、针对的应用领域有很大不同,某种映射策略所需要的信息可能恰恰是某些本体所缺乏的,多策略集成则具有更广泛的适应性。但是目前,各种映射策略也存在着缺陷,例如,传统的名称映射策略是一种基于字符串处理的方法(如编辑距离、相同字符个数),但名称相同的两概念可能是同名异义;而当两者的名称完全不同时,也可能表示相同的语义;当待映射本体的结构差异程度很大时,传统的自底向上的遍历方式来发现映射(如 cupid[2]),得出的映射结果很多是不正确的。这是因为下层概念是上层概念的细化，而上层概念是下层概念的抽象,当

基金项目:国家自然科学基金重点项目(the Key National Natural Science Foundation of China under Grant No.60433020);湖南省自然科学基金(the Natural Science Foundation of Hunan Province of China under Grant No.06JJ50142)。

作者简介:李鹏(1983-),男,硕士,研究方向为本体映射;徐德智(1963-),男,教授,主要研究方向为 Web 计算、语义网等;尹艳(1982-),女,硕士,研究方向为本体映射。

收稿日期:2008-06-12 **修回日期:**2008-10-08

上层概念差异很大时,此时从最底层开始比较会造成最初的映射结果就存在偏差,而后面的映射结果是通过在前面的基础上迭代得来的。

2.2　传统的多策略集成方法

对多个映射策略集成是本体映射的关键,而集成方法的优劣将直接影响映射结果的查全率和查准率。传统的集成方法都假设使用不同策略所获得的相似度值可以累加,对其得到的相似度赋以相应权值,再采用加权平均或 Sigmoid[3]函数对各个策略计算所得相似度值进行合并,这实际上是把多个结果进行数值上的集成,而并没有真正在语义层面综合考虑各种策略对映射结果的重要性,因此,待映射本体对缺失了某一方面的特征或特征发生变化时(随着本体的进化这是很可能的),影响了映射结果的质量。例如,加权平均法,由于用户或领域专家容易受主观因素的影响,因此,加权时常常会因考虑不周全、顾此失彼而使得集成后的相似度的值不能反映两概念间的真实相似程度,而 Sigmoid 函数法将较高的相似度赋予高权值比例,将较低的权值赋予低权重值,它在语义关系众多且比较弱时,突出了主要语义成分,因此比简单加权法有一定的优势,但当语义关系稀疏时,如果参数调整不适合的话,通过该方法得出的聚合结果常常比实际结果要偏低,导致大量的映射关系被遗漏掉。

3　动态多策略本体映射与集成

针对各个映射策略和策略的集成方法进行了改进。在单个映射策略方面,针对待映射本体间的差异,动态地进行映射处理,以避免本体某一方面信息的缺失对映射结果的影响;在多策略集成方法上,分析了各种策略在当前映射任务中的相对重要性,提出利用层析分析法对各个策略进行集成,以使各个策略的语义价值得到充分利用。下面分别对各个映射策略和多策略的集成做详细介绍。为了方便叙述,这里先给出本体的定义:

定义 1　本体:$O=(C,P,R,I,T)$

其中,C 表示概念的集合;P 表示属性的集合;R 表示关系的集合;I 表示实例集合;T 表示公理集合。概念和属性通称为实体。假设 A 和 B 分别表示源本体 O_1 和目标本体 O_2 中的某实体。

3.1　基于名称的策略

本体中相似的实体对其名称通常也存在相似性,传统的基于编辑距离或字串的方法来进行名称映射的准确度都比较低,不适合使用。基于 WordNet 来进行实体间的名称相似度计算,其核心思想是:如果两实体的 URI 一致或两实体互为同义词则相似度为 1;否则,通过计算两实体的同义词集在 WordNet 中的路径距离来计算相似度,取其最大值。

定义 2　在 WordNet 层次图中,实体 A,B 之间的语义距离 $Dist(A,B)$ 为连接它们的最短路径上 n 条边的权值的总和,即:

$$Dist(A,B)=\sum_{i=1}^{n} weight_i \tag{1}$$

其中,$Weight_i$ 是连接 A,B 的最短路径上第 i 条边的权值。考虑到自顶向下,实体的分类是由大到小,大类间的相似度肯定要小于小类间的,所以处于不同深度的实体的边赋予不同的权值,当实体由抽象逐渐变得具体,连接它们的边对语义距离计算的影响将逐渐减小。即边的权值可以表示为:

$$Weight(E)=\frac{1}{2^{Dep(E)}} \tag{2}$$

其中,$Dep(E)$ 表示边 E 在 WordNet 层次树中的深度,对于根节点来说,$Dep(E)$ 为 0。

另外还有两个关键因子需要考虑:边的强度和边的密度。一般地,一个父节点对某一子节点相对于其他子节点越重要,即边的强度越大,则该父子节点相连的边的权值越大;随着边的密度增加,边的权值越大。

假设某一条边 E 的子节点为 C,父节点为 F,边的强度可以表示为:

$$edge_{important}(E)=|IC(C)-IC(F)| \tag{3}$$

其中 $IC(C)$ 和 $IC(F)$ 分别表示节点 C 和节点 F 包含的信息量,信息量的计算参考文献[4]。

边的密度可以表示为:

$$edge_{density}(E)=\frac{1}{Wid(F)} \tag{4}$$

其中 $Wid(F)$ 表示由节点 F 引出的边的数目。

依据上述分析,任意一条边 E 的权值经过修正后可表述为:

$$Weight(E)=\frac{1}{2^{Dep(E)}}*edge_{important}(E)*edge_{density}(E) \tag{5}$$

上述公式保证了随着实体的边在 WordNet 层次结构中所处深度、强度和密度的增加,其权值会减小。

对于同义词集中的词汇,其名称相似度由同义词在 WordNet 层次树中的语义距离确定,分情况考虑:如果待比较实体的同义词存在着公共上位词(cuw),则两者的距离由它们分别与最近公共上位词的 $Dist$ 值之和确定;如果不存在公共上位词,则由两者的最短路径距离确定。其语义距离度量公式为:

$$S(A,B)=\begin{cases} Dist(A,cuw(A,B))+Dist(B,cuw(A,B)), & AB\text{ 存在公共上位词} \\ Dist(A,B), & \text{其他} \end{cases} \tag{6}$$

依据公式(6)转化得名称相似度计算公式为:

$$Sim_{name}(A,B)=\begin{cases} 1, & A\text{ 和 }B\text{ 的 }URI\text{ 相同,}A\text{ 和 }B\text{ 互为同义词} \\ \max(1-S(Synset_A,Synset_B)), & \text{其他情况} \end{cases} \tag{7}$$

其中,$Synset_A$,$SynsetB$ 分别表示实体 A,B 的同义词集合,由于 WordNet 是依据实体之间的语义组成的同义词典,因此该方法不仅在实体名称完全或者部分相同的情况下有效,而且在实体名称完全不同但存在一定语义关联的情况下也非常有效。

3.2　基于结构的策略

本体的结构蕴含着丰富的语义信息,在本体映射中起到非常重要的作用。针对传统方法存在的缺陷,考虑了本体间结构特征的差异,将结构级映射机制分为两种情况:自顶向下映射和自底向上映射。自顶向下映射算法花费的代价较小,因为一开始所要比较的对象比较少,以后的比较也只要用到前面的比较结果,而自底向上映射由于利用的是本体中最终描述的原子数据,因此更有可能得到好的映射结果。但是,对于具体的应用环境,需要做出不同的处理以保证映射的质量。下面举例说明,给出两个本体部分节点的树形层次结构图(图 1,图 2):

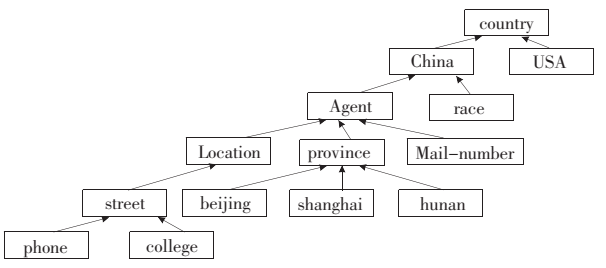

图 1　源本体(片段)

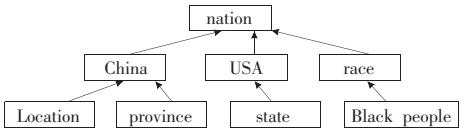

图2 目标本体(片段)

对于图1,图2所示的源本体和目标本体,如做自底向上的映射,先计算两本体叶子节点的相似度,再不断向上迭代计算,但是由于两者的结构存在较大的差异(实体个数相差很大、本体树高度不一致、叶子节点数目不同),它们的叶子节点并没有多大的相似性,映射结果显得意义不大。相反,此时如进行自顶向下的映射,效果会更好。依据上述分析,结构级映射时,遍历方式的选择需谨慎考虑,如果遍历方式不当,则有可能出现计算复杂度增加但得出的映射结果却不理想的情况。针对这个问题,提出利用本体异构度来评价待映射本体间结构特征的差异程度,当本体异构时,通过层次遍历,做自顶向下映射;反之,通过后序遍历,做自底向上映射。

定义3 本体异构度:

$$isomerous(O_1,O_2)=1-\frac{Entity(O_1,O_2)+Depth(O_1,O_2)+Leaf(O_1,O_2)}{3} \tag{8}$$

其中,$Entity(O_1,O_2)$指待映射本体间实体个数的比值(总取较小的一个为分子,下同),$Depth(O_1,O_2)$指待映射本体树高度的比值,$Leaf(O_1,O_2)$指待映射本体树叶子节点的比值。当 $isomerous>p$(规定值)时,则认为源本体和目标本体异构。P值由领域专家或实验确定,该文取值0.6时实验效果最好。下面给出结构级映射的算法描述:

```
Structure-Match(SourceTree S,TargetTree T){
sim_structure(s_i,t_j)=datatype_compatibility(s_i,t_j); //初始化操作
If (isomerous<P){
  Down_Top Matching(S,T){ //做自底向上映射
  //按后序遍历来唯一列举本体树的元素
  ArrayList s'=post_order(S);   //s'={s0,s1,s2,…,sn}
  ArrayList t'=post_order(T);   //t'={t0,t1,t2,…,tm}
sim(S,T)=W_structure*sim_structure(s_i,t_j)+(1-W_structure)*sim_name(s_i,t_j)
    if(sim>P1) 增加叶子集的相似度;
    if(sim>P2) 降低叶子集的相似度;
    } }
Else{
  Top_Down Matching(S,T){ //做自顶向下映射
  //按层次遍历来唯一列举本体树的元素
  ArrayList s'=layer_order(S);   //s'={s0,s1,s2,…,sn}
  ArrayList t'=layer_order(T);   //t'={t0,t1,t2,…,tm}
sim(S,T)=W_structure*sim_structure(s_i,t_j)+(1-W_structure)*simname(s_i,t_j)
    if(sim>p3) 迭代计算直到收敛于固定点
    } }
}
```

3.3 基于实例的策略

当两个实体具有共同的实例时,这两个实体可能是相似的。根据这种思想,传统的方法利用Jaccard[3]来进行相似度计算,但是这种方法没有考虑实例个数的差异程度,当实体间的实例数目不均衡时,得出的映射结果不准确。在传统的方法上进行改进,提出两个关键因子:丰富度和差异度。

定义4 丰富度:

$$richness=\mathrm{Min}\left\{1-\frac{1}{(sum_A+a)},1-\frac{1}{(sum_B+a)}\right\} \tag{9}$$

其中,sum_A、sum_B分别表示A、B的实例集(下同),$richness$值随实例数目的增大而增大,但随实例数目的增大,$richness$的增长应该放缓(因为两个实例比一个实例要可信得多,但50个实例与40个实例则没有明显区别)。a的作用是令$richness$值在实例数目为1时不会过小。当两实体拥有的实例数目越丰富时,则基于实例的策略得出的结果越可靠。

定义5 差异度:

$$difference=1-\frac{sum_A}{sum_B} \tag{10}$$

总取较小的一个为分子。差异度反应两实体实例丰富程度的差异,差异越大,$difference$值越大。当$difference$值较大时,就算$|sum_A\cap sum_B|=|sum_A|$,即$A$的实例完全被映射,最终计算出的实例相似度都可能永远达不到阈值。为了避免这种情况,$difference$值较大时,分母用$2\times\min(|sum_A|,|sum_B|)$来取代$|sum_A\cup sum_B|$。所以,基于实例的实体相似度计算公式为:

$$Sim_{instance}(A,B)=\begin{cases}richness*JaccardSim(A,B) & difference\leqslant E\\ richness*\dfrac{|sum_A\cap sum_B|}{2*\min(|sum_A|,|sum_B|)} & difference>E\end{cases} \tag{11}$$

3.4 多策略映射的集成

多策略本体映射的关键是怎么最优地集成各个策略,以充分利用各个策略包含的语义价值。考虑了本体的语义信息,提出使用层次分析法对各策略的重要性进行判断,对各个策略进行赋权并结合,从而得到最终的映射结果。

层次分析法(AHP)[5]是一种定性和定量相结合的决策分析方法,它按照分解、比较判断、综合的思维方式进行决策,特别适用于那些难于完全定量分析的问题。将该方法应用到语义Web领域,从定性与定量相结合的角度对多策略进行集成,避免了传统定性分析方法所带来的缺陷。下面举例说明如何基于层次分析法来进行多策略的集成:

假设,语义相似度为:$sim(A,B)=W_{name}*sim_{name}(A,B)+W_{structure}*sim_{structure}(A,B)+W_{instance}*sim_{instance}(A,B)$;其中$W_{name}$,$W_{structure}$,$W_{instance}$分别为三种策略计算出的相似度对映射结果的相对重要程度。对于以上三种相似度计算方法,影响最终映射结果的因子主要有三个:名称信息、结构信息、实例数目。对于不同的待映射本体,通过解析其特征(名称信息、结构信息和实例信息),确定它们对于映射结果的影响。根据AHP原理,首先构造出如下的相似度计算模型(图3)。

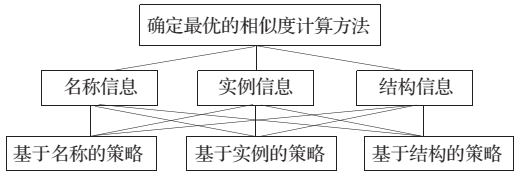

图3 多策略相似度计算模型图

根据AHP的评价尺度,构造三种影响因子对于确定最优相似度计算方法的判断矩阵[6],作相对重要度判断,并进行一致

性检验。设 $C1$,$C2$,$C3$ 分别表示三种影响因子:名称信息、结构信息、实例数目;W 表示三种影响因子对于确定最优相似度计算方法的重要程度;CI 为一致性指标,见表 1。

表 1　影响因子判断矩阵

G	$C1$	$C2$	$C3$	W	CI
$C1$	1	1/7	1/5	0.076	
$C2$	7	1	2	0.591	0.001<0.1
$C3$	5	1/2	1	0.333	

最后,分别给出三种策略的判断矩阵为:($\boldsymbol{A}$,$\boldsymbol{B}$,$\boldsymbol{C}$ 分别表示三种策略的判断矩阵)。

$$\boldsymbol{A}=\begin{pmatrix}1 & 2 & 3\\ \frac{1}{2} & 1 & 2\\ \frac{1}{3} & \frac{1}{2} & 1\end{pmatrix}\quad \boldsymbol{B}=\begin{pmatrix}1 & \frac{1}{5} & \frac{1}{3}\\ 5 & 1 & 2\\ 3 & \frac{1}{2} & 1\end{pmatrix}\quad \boldsymbol{C}=\begin{pmatrix}1 & \frac{1}{3} & \frac{1}{5}\\ 3 & 1 & \frac{1}{2}\\ 5 & 2 & 1\end{pmatrix}$$

计算得矩阵 $\boldsymbol{A}$,$\boldsymbol{B}$,$\boldsymbol{C}$ 对应的归一化特征向量为(计算过程略):

$$\boldsymbol{V}_A=\begin{bmatrix}0.539\\ 0.297\\ 0.164\end{bmatrix}\quad \boldsymbol{V}_B=\begin{bmatrix}0.106\\ 0.563\\ 0.331\end{bmatrix}\quad \boldsymbol{V}_C=\begin{bmatrix}0.106\\ 0.331\\ 0.563\end{bmatrix}$$

对于 $\boldsymbol{V}_A$,$\boldsymbol{V}_B$,$\boldsymbol{V}_C$,依据 AHP 原理,分别计算三种策略归一化特征向量的一致性比率:CRA=0.003;CRB=CRC=0.088。三者的值都小于 0.1,即通过了一致性检验。则可得:基于名称的策略的权重为:W_{name}=0.539*0.076+0.106*0.591+0.106*0.333=0.139;基于结构的策略的权重为:$W_{structure}$=0.297*0.076 +0.563*0.591 + 0.331*0.333=0.466;基于实例的策略的权重为 $W_{instance}$=0.164* 0.076+0.331*0.591+0.563*0.333=0.395。即三种策略相似度计算方法的优先排序为:基于结构的策略>基于实例的策略>基于名称的策略。依据上述计算得到的各个策略的权值来进行多策略集成,即得到最终的映射结果。

4　实验结果及分析

利用 OAEI2007 的标准测试数据集进行实验,程序实现了该文的方法,得出的映射结果取名为 Map,将其与目前国内外典型的映射系统的结果进行比较,其中,#1××~3××表示标准测试数据集 benchmarks 中的本体编号,R 表示查全率,P 表示查准率。实验结果如表 2,表 3。

表 2　Map 与其他测试系统的查全率比较

System	Falcon	Rimom	OntoDNA	ASMOV	Map
1××	1.00	1.00	1.00	1.00	1.00
2××	0.87	0.80	0.76	0.82	0.84
3××	0.77	0.86	0.78	0.86	0.78
Total	0.88	0.89	0.85	0.89	0.87

表 3　Map 与其他测试系统的查准率比较

System	Falcon	Rimom	OntoDNA	ASMOV	Map
1××	1.00	1.00	0.94	1.00	1.00
2××	0.93	0.94	0.75	0.85	0.95
3××	0.82	0.90	0.90	0.75	0.93
Total	0.92	0.95	0.86	0.88	0.96

从上面的实验结果可以看出,Map 的查准率最高,查全率只比 Falcon[7]、Rimom[8]和 ASMOV[9]等优秀映射系统稍低,但高于 OntoDNA[10]。也就是说该文提出的方法在保证了查全率的同时,较为明显地提高了查准率,即提高了映射结果的质量。另外,对于各组不同的测试数据,Map 的性能比较稳定。仔细对比分析每组数据的特点和不同映射算法的实验结果,主要有以下两方面的原因:(1)在映射过程中充分利用了本体包含的语义信息来进行相似度计算,并针对本体结构、实例数目的差异进行了动态地处理,使得映射算法更精确地反映了实体之间的语义关系;(2)通过利用层次分析法对不同映射策略进行集成,避免了传统方法中由于人为干预或简单加权造成的结果偏差,从而使得集成后的映射结果更加合理、准确。从整体实验结果均衡来看,该方法是有效的,达到了预期的目的。

5　结论

提出了一种动态多策略本体映射与集成方法,该方法对单个策略进行了动态地改进,弥补了策略所固有的缺陷,并通过层次分析法集成各个策略,使得各策略包含的语义价值得到充分利用,弥补了传统多策略集成方式的缺陷。从实验结果来看,该文的方法比传统的多策略映射与集成方法有所改进。

参考文献:

[1] Giunchiglia F,Yatskevich M,Shvaiko P.Semantic matching:Algorithms and implementation[J].Journal on Data Semantics,2007,IX:1-38.

[2] Madhavan J,Bernstein P A.Generic schema matching with cupid[J].VLDB,2001:49-58.

[3] 程勇,黄河,邱莉榕,等.一种基于相似度计算的动态多维实体映射算法[J].小型微型计算机系统,2006(5):975-979.

[4] 文敦伟,樊小虎.基于信息理论的启发式本体匹配框架[J].计算机技术与发展,2007(10):43-46.

[5] 黄显勇,毛明海.运用层次分析法对水利旅游资源进行定量评价[J].浙江大学学报,2001,28(3):327-332.

[6] 吕跃进,郭欣荣.群组 AHP 判断矩阵的一种有效集结方法[J].系统工程理论与实践,2007,30(8):132-136.

[7] Hu Wei,Zhao Yuan-yuan,Li Dan,et al.Falcon-AO:Results for OAEI 2007[C]//Proceedings of the ISWC'2007 Workshop on Ontology Matching OM-2007,Bexco,Korea,2007:160-167.

[8] Li Yi,Zhong Qian,Li Juan-zi,et al.Result of ontology alignment with RiMOM at OAEI2007[C]//Proceedings of the ISWC'2007 Workshop on Ontology Matching OM-2007,Bexco,Korea,2007:216-224.

[9] Jean-Mary Y R,Kabuka M R.ASMOV results for OAEI 2007[C]//Proceedings of the ISWC'2007 Workshop on Ontology Matching OM-2007,Bexco,Korea,2007:141-150.

[10] Kiu C C,Lee C S.OntoDNA:Ontology alignment results for OAEI 2007[C]//Proceedings of the ISWC'2007 Workshop on Ontology Matching OM-2007,Bexco,Korea,2007:185-195.