一种基于 Swoogle 本体映射的改进算法①

An Improved Algorithm of Ontology Matching Based on Swoogle

刘应龙　江　杰　(中南大学 信息科学与工程学院 湖南 长沙 410075)

摘　要： 基于 Swoogle 的本体映射关键步骤在于通过本体搜索引擎动态的获取多个背景本体。然而，目前该步骤却存在无法找到与某些概念相关联的背景本体，以及由于词义的模糊性搜集了错误的背景本体的问题。针对第一个问题，提出利用基于虚拟文档的映射技术，提取在 WordNet 中与概念同义的同义词，把原先的对单个概念进行搜索转换成对同义概念集进行搜索，进而提高本体搜索面，获取更多背景本体；针对第二个问题，提出基于语义环境的动态本体映射的新映射方法，采用该方法来排除错误背景本体，使本体收集更加精确。实验显示改进后的方法有效的提高了映射的查全率和查准率。

关键词： 本体 映射 语义网

本体映射是在本体间建立语义关系的关键步骤。随着本体映射技术的发展，目前出现了基于 Swoogle 的本体映射方法[1]，该方法采用本体搜索引擎在网络中对包含待匹配概念对的本体文件进行搜索，收集这些本体文件，然后对这些本体文件进行语义挖掘，最后输出待匹配概念对的映射关系。

然而，目前该方法却存在以下两点不足：1)本体获取率不高这直接导致了本体映射的查全率不高，此点不足一方面是由于目前网络中本体文件数量还不是很大，另一重要的方面是网络资源没有充分得到利用。2)本体收集精度不高[2]导致了本体映射准确率未达到理想结果的主要因素，在本体收集的时候我们只考虑了概念名称相等，却没有考虑到概念在不同的环境中所表达的语义不同。针对第一点不足，本文采用基于虚拟文档本体映射技术[3]从 WordNet[4]中提取同义词概念集，把原先的对单个概念进行搜索转换成对同义概念集进行搜索，进而提高本体文件的搜索率。针对第二点不足，本文采用基于语义环境的动态本体映射方法对搜索后的本体进行选取，排除由于词汇的同名异义引起的错误本体，提高本体收集精度。

2　基于Swoogle的本体映射方法及不足

基于 Swoogle 本体映射主要思想是：通过本体搜索引擎动态的选取多个本体，然后利用推理工具挖掘单个本体或多个本体中所隐含待匹配概念对的语义关系。这种方法比基于人工选定背景本体映射方法[1]有着可重用资源更加丰富的优点，所以在一般情况下该方法有着更高的映射查全率。

在文献[1]中，讲述了基于 Swoogle 本体映射方法，然后上述方法却存在着两个主要的问题。1) 无法搜索到相关联的定位本体　一个重要的原因是由于我们并没有充分的利用现有的网络本体资源。2) 本体收集不精确　这主要是目标本体与搜索到的背景本体的语义环境不同引起的。

3　本体搜索改进算法

本体搜索改进算法主要为了提高本体搜索面，通过提取 WordNet 概念同义词集来实现。字典根据语义关系组织成一棵树的形式，其中树中的一个节点由一些同义词组成。由于一个词可能有多种意思，所以一个词可能在树中的多个节点上出现。所以必须考虑本体中的概念词汇定位在 WordNet 中的哪个节点上。这个定位过程通过映射来实现，把本体概念与 WordNet 存在映射关系的节点作为同义词集。基于上述讨论，本文采用了基于虚拟文档的映射方法。

3.1 基于虚拟文档的映射

3.1.1 虚拟文档的建立

虚拟文档是为了描述概念特点而建立起来的文

① 收稿时间:2008-11-20

档，每个概念都可建造自身的虚拟文档。虚拟文档主要包括本节点的描述信息和邻近节点的描述信息组成，由于本体和 WordNet 建模的不同，所提供的描述信息种类也是不同。为了方便讲述，现给出以下几个定义。

定义 1. 本体概念描述信息(Des) 它的描述信息有一些与概念有关的评论，注释，标签，和属性组成定义为：

$$Des(e)=\alpha_1*CWC(e)+\alpha_2*CAC(e)+\alpha_3*CLC(e)+\alpha_4*CPC(e)$$

其中，α_1，α_2，α_3，α_4是四个代表权值的有理数。CWC(e)、CAC(e)、CLC(e)、CPC(e)分别代表概念 e 的评论，注释，标签，和属性的词汇集合。值得注意是，这里的描述信息并没有包括概念名称，这主要是映射的目的是排除同名异义，概念的名称对这种映射目的并没有帮助。

定义 2. WordNet 概念描述信息 它的描述信息是与实体有关的评论和注释组成，定义为：

$$Des(e)=\alpha_1*CWC(e)+\alpha_2*CAC(e)$$

其中，α_1，α_2是两个代表权值的有理数。

定义 3. 虚拟文档(VD) 虚拟文档由实体的描述信息和实体子父概念集节点组成，定义为：

$$VD(e)=Des(e)+r_1*\sum_{e'\in FN(e)}Des(e')+r_2*\sum_{e'\in CN(e)}Des(e')$$

FN(e)代表关于实体 e 的父概念集合，CN(e)代表关于实体 e 的子概念集合,r_1,r_2是两个有理数权值。

3.1.2 相似度计算

相似度计算在文献[3]中具体的介绍，这里不再赘述。

3.1.3 映射发现

经过了相似度计算，产生了本体中的一个概念与 WordNet 中多个节点的多个相对应相似度值，选取相似度中最大的作为映射对。

$$Sim(A,A_j)=Max\sum_{k=1,2\cdots n}sim(A,A_k)$$

即选择满足上述公式的(A,Aj)做为映射对。其中，A 是本体的概念，Aj 是 WordNet 中的第 j 个相对应节点。

3.2 同义概念集确定

通过 3.1 节的映射，可以找到搜索概念的 WordNet 同义概念集，以此概念集为搜索同义概念集。

4 本体收集的改进算法

排除错误背景本体，排除的方法的基本思想与章节三中相似，但这里采用了基于语义环境的动态本体映射的方法。

4.1 基于语义环境的动态本体映射

基于语义环境的动态本体映射的主要思想是：相似度计算公式的确定和权值的设置是一个动态的过程，它不仅要考虑各个策略的自身特征，还需考虑各个策略所处的语义环境。例如注释信息的丰富程度对基于注释策略的影响，在名称相同情况下对策略组合方式以及策略自身变化的影响等等。该映射方法比传统的映射方法能够更有效的利用语义信息，特别是针对在对排除同名异义为主要映射目的的映射上有着更好的效果。

下面就对几个具体策略进行介绍，为了简单描述起见，现假设<A,A'>为待匹配概念对。

4.1.1 基于名称的策略

该策略采用 lenvenshtein's 编辑距离与 WordNet 相结合的方法来计算概念间的相似度。因篇幅有限不做具体讲述。其具体的方法已在文献[6]中有详细介绍。

4.1.2 基于实例的策略

在概念同名的情况下，实例是判断两个概念是否相关最直接最有效的依据。因为实例是概念的实例化结果，概念的属性都包含在实例中，如果存在两个相同的实例，则说明两个概念也有许多属性相同，这也说明两者概念是有一定的相关性的。在现实中也很难找出两个同名异义的概念，会有相同的实例。所以在概念同名的情况下基于该策略的相似度计算公式为：

$$Sim_{ins}(A,A')=\begin{cases}1 & Ins(A)\cap Ins(A')\neq\varnothing\\ 0 & Ins(A)\cap Ins(A')\in\varnothing\end{cases}$$

其中，Ins(A)、Ins(A'),分别为元素 A,A‘的实例集合。

在概念不同名的情况下，传统的方法是通过计算两个概念的相同实例集与共有实例集的比值来作为相似度值。然而这种方法却忽略了实例信息之间的差异，一般来说实例越丰富，语义信息越多，基于该策略的相似度计算也就越重要、越可靠以及概念间的实例信息越均衡该策略的可信度也更高。考虑到这一点，本文给出了如下的一个定义：

定义 4. 支持度：

$$Support(A,A^{'}) = k*(N_A + N_{A'})(1 - Min(N_A, N_{A'}) / Max(N_A, N_{A'}))\ k \in Q^{+}$$

其中，N_A,表示概念 A 所对应的实例个数。支持度是待匹配概念实例数目之和的正比例函数，与实例的均衡程度成正比。总的实例数目越多，实例之间的数目越均衡，基于该实例的策略的可信度越高，权值也相应增高。

经过调整后的在概念不同名的情况下基于实例策略的相似度计算公式为：

$$Sim_{ann}(A,A^{'}) = Support(A,A^{'}) * \frac{Ins(A) \cap Ins(A^{'})}{Ins(A) \cup Ins(A^{'})}$$

4.1.3 基于注释的策略

该策略把注释的词汇看成实例，所以该策略与概念不同名的情况下的基于实例策略类似，也引用了实例支持度。然而除了上述改进外，我们还需对注释信息在概念同名与不同名情况下进行不同的处理。在概念不同名的情况下我们需要对注释进行停用词删除，而在概念同名的情况下我们除了删除停用词外还需删除与概念相同的词汇。因为这些词汇都是基于该策略的噪音信息，影响基于该策略相似度计算的正确性。

4.1.4 基于结构的策略

传统基于结构的策略的主要思想是：把给定两个元素的父类间和子类间的相似度值通过某种方法传播到给定元素间的相似度中。然而目前所使用的相似度传播方法却并未考虑元素与父类和子类之间的联系紧密程度所带来的传播影响，一般来说子父类的联系越紧密或者越相似所带来传播的值应该越大。基于此本文给出了语义传播因子的定义：

定义 5. 语义传播因子(SFF):

$$SFF=k*Min(Sim(A,A_f),Sim(A^{'},A_f^{'}))\ \ k \in Q^{+}$$

其中，A_f,A_f'分别是 A，A ‘的父概念，$Sim(A,A_f)$是我们采用基于名称、注释和属性计算所得的相似度。SFF 的语义是父子之间的相似度越高，越紧密，其相似度的传播应该越高。

改进后的父类对元素的相似度传播值为：

ISFsim=SFF* SFsim　其中 SFsim 是改进前的父类对元素的相似度传播值。

4.1.6 多策略合并

多策略的合并需要分为概念同名与不同名两种情况分析。

概念同名的情况下，要是能判定概念间是相关的，则可判定概念是同义的，相似度值定为 1，否则说明出现了同名异义的情况。判断两个概念是否相关，一种有效的方法就是判断两个概念的实例集中是否有相同的实例，如果有则说明相关，另外一种办法是通过基于注释和结构的相似度值我们取为 CL_{sim} 这个数值反应了概念间的相关性，如果该数值大于给定阈值的话，则可认为是相关的。在概念同名且不相关的情况下说明出现了同名异义，所以基于概念名称的策略不能采用，所以其相似度的值即为 CL_{sim}。固在概念同名的情况下相似度合并公式为其合并公式为：

$$Sim(A,A^{'}) = \begin{cases} 1 & I_A \cap I_{A^{'}} \notin \varnothing \text{ or } CL_{sim} \geq Lim_{sim} \\ CL_{sim} & CL_{sim} < Lim_{sim} \end{cases}$$

其中，I_A表示概念 A 的实例集和，LIM_{sim}是一个给定的阈值。

$$CL_{sim} = \sum_{k=1,2} w_k \sigma(Sim_k(A,A^{'})) / \sum_{k=1,2} w_k$$

其中，W_K 是某个策略的权值，σ 是一个 sigmoid 函数，sigmoid 是一个平滑函数，它使得合并结果偏向于预测值高的策略。

在概念不同名的情况下，则需要通过基于名称、实例、结构、注释相结合来计算相似度。其合并公式为：

$$Sim(A,A^{'}) = \sum_{k=1,2\cdots n} w_k \sigma(Sim_k(A,A^{'})) / \sum_{k=1,2\cdots n} w_k$$

4.2 错误背景本体排除

通过上面的基于语义环境的动态本体映射的相似度计算，我们知道在概念对同义的情况下，其相似度值为 1。下面我们给出错误背景本体排除的过程：

① 提取与待匹配概念对<A,B>在背景本体中对应的概念对<A',B'>；

② 分别对<A,A'>,<B,B'>进行基于语义环境的动态本体映射的相似度计算，获取 Simbs(A,A')、Simbs(B,B')；

③ 排除不满足 Simbs(A,A')=Simbs(B,B')=1 此条件的错误背景本体。

5 改进后的基于Swoogle的本体映射

改进后的映射系统的基本步骤：

输入：两个待匹配本体；

输出：本体中的概念对的匹配关系

Step1.在本体中提取待匹配概念对

Step2.使用本体搜索引擎在对包含待匹配概念对的本体进行搜索

Step3.对搜索后的背景本体进行筛选，排除错误误景本体

Step4.计算经过筛选后的本体文件数目， 如果 $O_{num} \leqslant O_{min}$ 进入 Step4，否则进入 Step5。其中 O_{num} 是筛选后的背景本体文件数目，O_{min} 是一个固定的整数阈值

Step5.提取同义概念集中的同义概念，进入 Step2。如果同义概念集替换结束，进入 Step1

Step6.通过推理工具对筛选后的背景本体提取待匹配概念对的语义关系

6 试验结果及分析

6.1 实验数据

考虑到基于语义网本体映射时间效率较低，在没有影响到结果的正确性基础上，本文的实验选用了规模较小的 university 数据集，该数据集分别包括 SWRC 和 LUBC 两个本体，描述的是大学本体。其中 SWRC 含有 56 个概念，LUBC 含有 43 个概念。

本文采用查全率和查准率作为评价标准对实验结果进行评估。我们手工建立了测试数据集的映射关系，用它们作为测评标准。

6.2 实验结果及分析

我们主要做了两组实验，第一组实验主要针对改进前的算法，查全和查准率分别由有图 2 中的 R1,P1 所示。第二组实验主要针对改进后的算法，通过动态的调整 Lim_{sim} 值，来测试改进后的算法的效果，其中查全和查准率分别有图 1 总的 R2，P2 所示。

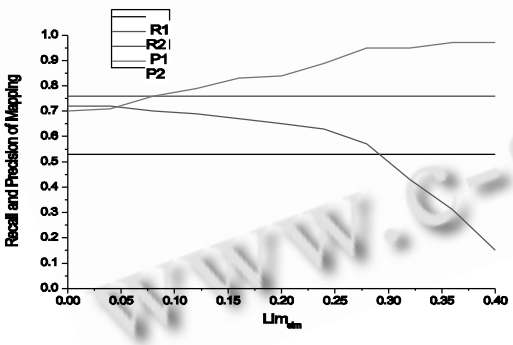

图 1 查全和查准率对比图

从图 1 中我们可以看出改进前算法的查全率和查准率分别为 53%、76%,从实验结果中看出，较低的查全率主要是有未找到背景本体的原因，特别是有三个以上词汇组成的概念，其背景本体就更少，在查准率不高的主要原因是部分收集了错误的背景本体。在改进后的算法试验结果中我们可以看出，当只进行搜索进行改进而未对收集未做处理时，查全率提高了 20%，达到了 73%。在实验结果中我们发现，上升的查全率主要集中在概念词汇数少于的 2 的概念中，其中单个词汇概念更加明显，本体中的所有单个词汇概念对都能找到映射关系，从中我们可以看出基于 Swoogle 本体映射在概念含有多个词汇中其效果不佳，在未做收集改进时，对应查全率的上升，查准率也相应有所下降，下降了 6%，从结果中看出下降的主要原因一方面是采用了搜索的改进算法后，搜索到了新的背景本体，而这些背景本体中也有部分是错误的背景本体，另外方面是搜索的改进算法在寻找同义词集中也出现了错误。接着对收集进行改进时，随着 Limsim 值的增加，其查准率也对应上升，而相反查准率却有所下降，特别是当 Lim_{sim} 大于 0.3 时下降的幅度比较大，这主要原因是由于 Lim_{sim} 的值过大，即使是正确的背景的本体也未能达到这个数值，所以错误的排除了正确的背景本体。总的来说，从图 1 中我们可以看出，改进后的映射算法，比较有效的提高了映射效果。

6.3 结论及未来工作

本文主要针对现有基于 Swoogle 本体映射过程中的背景本体搜索面窄和收集精度不高等问题，提出基于虚拟文档的本体映射技术提取 WordNet 同义词集，提高搜索面；以及通过基于语义环境的动态本体映射对收集后的本体进行筛选，提高收集精度。经过试验分析，改进后的算法在映射效率上要优于原先的算法。

我们的未来工作主要有以下两方面：第一个方面，降低改进后基于 Swoogle 本体映射算法的时间复杂度；第二方面，把基于 Swoogle 本体映射算法结合与传统映射方法中，提高映射的效果。

参考文献

1 Aleksovski Z, Klein M, ten Kate W. Matching Unstructured Vocabularies Using A Backgound Ontology. Proceedings of Knowledge Engineering and Knowledge Management, 2006:182－197.

2 Gracia1 J, Lopez V, Aquin M, et al. Solving Semantic Ambiguity to Improve Semantic Web based Ontology Matching,OAEI, 2007:4－27.

3 Qu YZ, Hu W, Cheng G. Constructing Virtual Documents for Ontology Matching. Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland, 2006:23－31.

4 Fellbaum C. WordNet - An Electronic Lexical Database. MIT Press, 1998.