第35卷 第13期 计 算 机 工 程 2009年7月
Vol.35 No.13 Computer Engineering July 2009
·软件技术与数据库· 文章编号：1000—3428(2009)13—0049—03 文献标识码：A 中图分类号：TP311

本体候选映射集搜索方法的研究

尹 艳，徐德智，李 鹏
(中南大学信息科学与工程学院，长沙 410083)

摘 要：针对传统的本体候选映射集搜索方法时间复杂度偏高、且容易得出错误的候选映射集等问题，提出一种改进的候选映射集搜索方法。该方法通过比较概念间的名称相似度获得初始的候选映射集，利用概念间的相关度对其进行扩展，从而得出最终的候选映射集。实验结果验证了该方法的有效性。

关键词：本体；相关度；候选映射集

Research on Ontology Candidate Mapping Sets Search Approaches

YIN Yan, XU De-zhi, LI Peng
(College of Information Science and Engineering, Central South University, Changsha 410083)

【Abstract】The traditional search methods suffer from the problems of high time complexity and easily getting wrong candidate mapping sets. This paper proposes an improved method of search candidate mapping sets. This method obtains the initial candidate mapping sets through comparing the similarity between the different concepts, and extends them by making use of correlation to get the final candidate mapping sets. Compared with traditional methods, experimental results prove that this method is valid.

【Key words】ontology; correlation; candidate mapping sets

1 概述

语义 Web 的发展导致本体数量激增，然而通过不同途径开发出来的本体经常存在差别。本体映射的目的就是要找到这些本体之间的语义联系，以便于知识共享和重用。实现本体映射的关键是计算本体间实体的语义相似度，但是由于本体中实体的数目十分庞大，因此在计算相似度时将本体中的每对实体都考虑在内是不现实的。为了得到更好的映射结果，很有必要在映射算法执行之前缩小待比较实体对的范围，以降低映射过程的时间复杂度和空间复杂度。因此，寻找一种合适的搜索候选映射集的方法意义重大，通过对它的研究，从而为下一步映射工作打下基础。

2 相关工作

为了叙述方便，首先给出一些基本概念的定义。

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(本文仅考虑本体中概念的候选映射集)。A 和 B 分别表示源本体 O_1 和目标本体 O_2 中的任一概念。

定义 2 候选映射概念集：指不同本体之间可能存在语义关系的概念对的集合。

目前，国内外关于候选映射集搜索方法的研究[1-2]大体可分为以下 2 类：(1)利用实体名称的相似性[3]来搜索，该方法由于仅考虑了语法上的信息，因此过于片面，容易遗漏掉或得出不一致的候选映射集，例如，源概念 human 和目标概念 fellow，通过名称相似度得出的值通常很低，如果此时设定的阈值较高，那么认为不属于候选映射而排除掉，但是实际上它们表达的语义是一致的；(2)根据启发式规则[4]来搜索，该方法在本体规模过于庞大、关系十分复杂时，容易得出冗余的或相互冲突的候选映射集，且随候选映射空间的不断扩充，搜索次数也将大大增加，从而造成搜索效率低下、复杂度偏高等缺点，其现实意义不大。为了弥补上述缺陷，本文提出了一种改进的候选映射集搜索方法，并通过实验验证了该方法的有效性。

3 改进的搜索方法

3.1 核心思想

为了克服传统方法存在的缺陷，全面且准确地找出不同本体中概念蕴含的语义关系，本文在权衡了时间复杂度的基础上，综合利用本体的名称信息和结构信息来寻找候选映射概念集。核心思想是：对于 O_1 中的任一概念 A，通过名称相似度比较，在 O_2 中找出与其名称最相似的概念 B，加入到候选映射关系中；以 B 概念为锚点，将 B 与 B 的概念集中的概念进行相关度计算，如果其概念集中某一概念与 B 的相关度大于阈值，则认为 A 概念与 B 概念集中的该概念也存在候选映射关系。

3.2 名称相似度

待映射本体的规模通常很庞大，为了减小搜索过程的复杂度，本文首先通过计算概念间的名称相似度找出锚点，从而为搜索算法提供一个最可靠的入口以提高搜索的效率。传统的基于编辑距离和相同字符的个数的方法的准确度都比较底，不适合使用。本文在对概念进行预处理、URI 比较的基础上，结合 WordNet 来进行名称相似度计算，其核心思想是：如果待比较概念对的 URI 相同或待比较概念对互为同义词，则名称相似度为 1，否则，通过计算它们在 WordNet 中的语

基金项目：湖南省自然科学基金资助项目(06JJ50142)；湖南省国土资源厅科技计划基金资助项目(200718)

作者简介：尹 艳(1982－)，女，硕士，主研方向：Web 计算，语义网；徐德智，教授；李 鹏，硕士

收稿日期：2009-01-10 **E-mail：**xiuxiannanzi@163.com

义距离来确定名称相似度。

定义 3 在 WordNet 层次结构中，两概念 *A,B* 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。另外，考虑到自顶向下，概念的分类由大到小，大类间的相似度肯定要小于小类间，所以，处于不同深度的概念的边赋予不同的权值，当概念由抽象逐渐变得具体时，连接它们的边对语义距离计算的影响将逐渐减小。

定义 4 在 WordNet 层次结构中，一个概念 *C* 引出的边的权值为

$$weight(C)=\frac{1}{2^{Dep(C)}} \tag{2}$$

其中，*Dep*(*C*)表示概念 *C* 在 WordNet 层次树中的深度，对于根节点来说，*Dep*(*C*)为 0。上述定义保证了随着概念的边在 WordNet 层次结构中所处深度的增加，其权值会减小。

另外，提出 2 个关键因子：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；概念节点的密度越大，边的权值越大。假设某一条边 *E* 的子节点为 *C*，父节点为 *F*，则有：

定义 5 边的强度为

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

其中，*IC*(*C*)和 *IC*(*F*)表示子节点 *C* 和父节点 *F* 包含的信息量，信息量的计算过程如下：设 *O* 为本体，*P* 为任一概念节点，$arcs_i$ 表示与 *P* 相连的边的数目，则本体中 *P* 包含的信息量可由与 *P* 相连的边的数目和各边的重要性确定。与 *P* 相连的第 *i* 条边对 *P* 的信息量贡献值(即边的信息量)为

$$H(a_i)=-\log P(x=a_i) \tag{4}$$

其中，$P(x=a_i)=\frac{1}{arcs_i}$。

由此得，节点 *P* 包含的信息量为

$$H(P)=\sum K_i\times H(a_i) \tag{5}$$

其中，K_i 为边的重要性系数(不同关系类型的边的重要性不同，is-a 关系的重要性要大于其他关系)。

定义 6 边的密度：

$$edge_{\text{density}}(E)=\frac{sum_C+sum_F}{2\times sum_{edge}} \tag{6}$$

其中，sum_{edge} 表示边的总数；sum_C 和 sum_F 分别表示由节点 *C* 和节点 *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以表示为

$$weight(E)=\frac{1}{2^{Dep(E)}}\times edge_{\text{important}}(E)\times edge_{\text{density}}(E) \tag{7}$$

对于待比较的概念对，其名称相似度由两概念在 WordNet 层次树体系中的语义距离确定，分情况考虑：如果待比较概念存在公共上位词(rcw)，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

$$S(A,B)=\begin{cases}Dist(A,rcw(A,B))+Dist(B,rcw(A,B)) & A,B\text{存在公共上位词}\\ Dist(A,B) & \text{其他}\end{cases} \tag{8}$$

依据式(8)转化得名称相似度计算公式为

$$Sim_{name}(A,B)=\begin{cases}1 & A\text{和}B\text{的}URI\text{相同，}A\text{和}B\text{互为同义词}\\ 1\text{-}S(A,B) & \text{其他情况}\end{cases} \tag{9}$$

由于 WordNet 是依据概念之间的语义组成的同义词典，因此该方法不仅在概念名称完全或者部分相同的情况下有效，而且在概念名称完全不同但存在一定语义关联的情况下也非常有效。

3.3 相关度

相关度(relation)指概念之间的语义关联程度。

在同一本体中，相关度的大小在结构上体现为概念间存在连通的路径距离的长短。下面分为概念集的确定和相关度的计算 2 步来阐述相关度。

本体结构包含了十分丰富的语义信息，本文在得到初始候选映射集的基础上，综合利用了本体的结构信息对其进行扩展。考虑到本体中概念的数量十分庞大，本文提出利用概念集对扩展的范围进行限制，在保证搜索质量的同时提高了搜索的效率。核心思想是：确定目标概念的概念集，计算目标概念与其对应概念集中的概念的相关度，将其概念集中的与目标概念的相关度大于阈值的概念加入候选映射集中。

3.3.1 概念集的确定

为了叙述方便，先给出一些概念的含义。

定义 7 后代概念集(CCS)：它是目标概念的下层概念，与目标概念的层次距离不大于某一规定值 P_1，它的后代概念必须是目标概念的后代概念。

定义 8 祖先概念集(FCS)：它是目标概念的上层概念，与目标概念的层次距离不大于某一规定值 P_2，它的后代概念必须是目标概念的祖先概念。

下面仅给出确定后代概念集的算法，确定祖先概念集的算法类似：

算法 1 确定目标概念 *B* 的后代概念集 B_{CCS} 的算法

Step 1 将目标概念的后代概念加入后代概念集。

Step 2 如后代概念集中的概念 *C* 有后代概念，且与目标概念的层次距离不大于规定值，则将 *C* 的后代概念加入后代概念集。

Step 3 循环执行，直到没有新的后代概念加入。

算法 1 说明：在确定目标概念的概念集时，概念集的范围大小(如 P_1, P_2)需要谨慎考虑，本文认为概念集的范围大小与该概念的重要性有关，一个概念越重要，则作用于该概念和被该概念作用的关系就越多，即与它语义相关的概念也就越多，限于篇幅，概念重要性评价方法不再详述。

3.3.2 相关度的计算

为了更好地说明相关度的计算，给出本体部分节点的树形层次结构如图 1 所示。

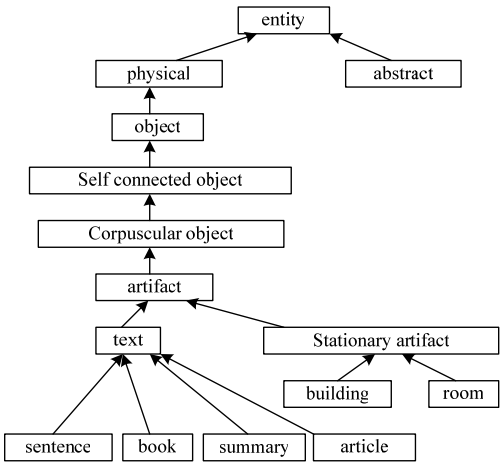

图 1 (本体片段)树状结构

对于目标概念与其后代概念集中的概念来说，影响两者相关度的因子主要有以下几个：(1)目标概念所在的层次，层次越深，与该概念相关的概念就越多，如图 1 所示，与概念 text 相关的概念(book, fiction, article, summary)要多于与概念 aitifact 相关的概念(text, stationary artifact)；(2)目标概念与其后代概念集中概念的层次差，层次差越大，则两者的语义距离越大，相关度越小；(3)后代概念集中概念的分类程度，分类越趋向于细致，则相关度越小。

依据以上分析，后代概念集相关度的计算表达式为

$$relation_{CCS}(B,B_{CCS})=\sum_{i=1}^{k}\left\{[1-Dist(B,B_i)]\times\sqrt{\frac{2\mathrm{dep}(B_i)}{\mathrm{dep}(B)+\mathrm{dep}(B_i)}}\times\frac{1}{Wid(B_{CCS})}\right\} \tag{10}$$

祖先概念集相关度的计算类似：

$$relation_{FCS}(B,B_{FCS})=\sum_{i=1}^{k}\left\{[1-Dist(B,B_i)]\times\sqrt{\frac{2\mathrm{dep}(B_i)}{\mathrm{dep}(B)+\mathrm{dep}(B_i)}}\times\frac{1}{Wid(B_{FCS})}\right\} \tag{11}$$

其中，*Dist* 表示目标感念与其概念集中概念的语义距离，其值的计算参考式(1)、式(2)、式(3)、式(6)、式(7)综合得来；$\sqrt{\frac{2\mathrm{dep}(B)}{\mathrm{dep}(B)+\mathrm{dep}(B_i)}}$ 和 $\sqrt{\frac{2\mathrm{dep}(B_i)}{\mathrm{dep}(B)+\mathrm{dep}(B_i)}}$ 分别表示目标概念与其后代概念集和祖先概念集中概念的层次差；$1/Wid(B_{CCS})$ 和 $1/Wid(B_{FCS})$ 分别表示后代概念集和祖先概念集中概念的细分程度(即宽度)，*i* 表示对应的概念集中概念的个数。

由于相关度的计算不具有对称性，一般来说，概念与其后代概念的相关度要大于与其祖先的相关度，如图 1 所示，text 与 book 的相关度要大于 text 与 artifact 的相关度，因此 $relation_{CCS}(B, B_{CCS})$ 集合中的元素个数通常要大于 $relation_{FCS}(B, B_{FCS})$中元素的个数，实验结果也证明了这一点。

3.4 候选映射集搜索算法

本文首先通过排除歧义、形式统一等预处理来标准化本体中的概念，与现有的自然语言预处理算法相比，本文做了一定的改进，例如，现有的缩写词扩展方式是依照领域缩写词词典，将由首字母表示的缩写词还原为完整形式。但是，常用的词汇缩写形式，除了取首字母外，还有 2 种：一种是取单词的前几个字母再加“.”，比如用 AI 表示 artificial intelligence；另一种是在多个词连写时，将部分或全部词只保留前几个字母，比如用 synset 表示 synonymy set，本文的预处理加入了对这 2 种情况的处理，在完成预处理的基础上，下面给出候选映射概念集的算法描述：

算法 2 候选映射集搜索算法

输入：OAEI 标准测试数据集中的本体

输出：候选映射集 candidateSet(*A*)

Step1 根据式(9)计算出 *A* 的最相似概念 *B*，将 *B* 加入候选映射概念集 candidature(*A*)中。

Step2 根据式(10)计算出 *B* 的后代概念集相关度，如果 $relation_{CCS}(B,B_{CCS})$>阈值 *T*，则将与 B_{CCS} 中大于 *T* 的概念加入 candidature(*A*)中。

Step3 根据式(11)计算出 *B* 的祖先概念集相关度，如果 $relation_{FCS}(B,B_{FCS})$>阈值 *T*，则将与 B_{FCS} 中大于 *T* 的概念加入 candidature(*A*)中。

Step4 循环执行 Step2, Step3，直到没有新的概念加入 candidature(*A*)中。

4 汇实验结果与分析

基于上述的候选映射集搜索方法，本文选取 OAEI2007 提供的标准数据集进行测试。采用 Java 语言程序实现了候选映射集搜索模块，并作为 SNAX 系统[5]中的一个子模块。

在实验进行前先做预处理，使得名称相同的概念不存在歧义，将#101 中的本体 onto.rdf 和#202 中本体 onto.rdf 输入，将根据本文方法得出的实验结果和依据实体名称相似性(记为方法 1)、启发式规则(记为方法 2)得出的实验结果进行比较，见表 1。

表 1 候选映射集结果比较表

System	发现候选映射对数目/对	所需时间开销/s
本文方法	208	12
方法 1	236	18
方法 2	252	24

从表 1 的实验结果可以看出，本文方法发现的候选映射集的数目相对于传统的方法稍低，搜索耗费的时间比传统的方法要少。通过分析，主要有 2 个方面的原因：(1)本文的方法在找出名称最相似的概念对时比传统的方法更加准确，在根据相关度进行扩展时，对相关与相似进行了很好地区分，把大量相关但语义关系较弱的候选映射对过滤掉了，从而避免了得出大量错误的候选映射关系，提高了搜索的准确率；(2)利用概念集代替所有概念进行相关度扩展，并考虑了本体间的概念层次差异、概念细分程度等问题，大大缩小了搜索的范围，从而提高了效率。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 结束语

本文综合考虑了本体的名称信息和结构信息，提出了一种改进的候选映射集搜索方法。该方法对相关但不一定相似的概念做了很好的区分，弥补了现有方法的不足。实验结果表明，该方法在保证搜索质量的基础上，时间复杂度较为合理，为以后的映射工作铺平了道路。下一步工作的主要任务包括以下 2 个方面：(1)利用推理技术对搜索方法进一步完善，并对得到的候选映射概念集进行冲突检测；(2)推广本文的方法以得到本体的实例和属性的候选映射集。

参考文献

[1] Yaghlane B B, Laamari N. OWL-CM: OWL Combining Matcher Based on Belief Functions Theory[C]//Proceedings of the 2nd International Workshop on Ontology Matching. NY, USA: [s. n.], 2007: 195-207.

[2] 梁晓涛. 基于语义 Web 的本体映射[D]. 合肥: 安徽大学, 2006.

[3] 曹泽文, 钱 杰, 张维明, 等. 一种改进的本体映射方法[J]. 科学技术与工程, 2006, 10(6): 3078-3082.

[4] Ehrig M, Staab S. QOM——Quick Ontology Mapping[C]// Proceedings of the 3rd International Semantic Web Conference. Arlington, VA, USA: [s. n.], 2004: 683-697.

[5] 郑春卉. 基于本体的概念语义相似度研究[D]. 长沙: 中南大学, 2006.

编辑 索书志