高效的异构本体的映射算法研究*

An Efficient Heterogeneous Ontology Mapping Algorithm

王松 [1,2] 马勇 [1] 王刚 [1] 刘晓光 [1]

Wang Song, Ma Yong, Wang Gang, Liu Xiao-guang

（1.南开-百度联合技术实验室，南开大学信息学院，天津，300071）

（2.装备保障系，军事交通学院，天津，300161）

（1.Nankai-Baidu Joint Lab, College of Information Science, Nankai University, Tianjin, 300071）

（2.The Equipment support Department, Military Transportation University, Tianjin, 300161）

摘要： 基于本体的概念间相似度计算已经在信息检索等诸多领域成为当今信息技术研究的热点问题之一。本文的工作是针对描述同一领域的多个本体间存在的异构问题，设计一种快速、高效的映射算法来实现异构本体的融合。本文提出了一种基于异构本体的相似度计算方法，通过字面概念相似度和语义结构（包括节点深度、节点密度、边权重、信息量等）相似度等方面的综合计算，可以准确地得到异构本体间的概念映射关系，同时，通过对映射方法的优化，算法的匹配速度也有很大程度的提高。实验结果表明，该算法可以有效地排除本体异构的影响，得到较好的概念相似性计算效果。

Abstract: The research of ontology-based similarity calculation between concepts has already been a hot issue of information technology in the fields of information retrieval, and so on. In this paper, the contents of the study is to find a fast and efficient mapping algorithm for heterogeneous ontologies in the same field.This paper put forward a method of similarity calculation based on heterogeneous ontologies,considering the factors of similarity of literal meaning and semantic structure(including the depth of the node, node density, edge weight,information content, etc.) can get concept mapping between heterogeneous ontologies more accurate.Simultaneously,taking into account the optimization of mapping method,the speed of matching has also been improved to a large extent. the problem of how to improve the speed of matching more effectually has been mentioned in this paper.The experiment results show this method can effectively get better effectiveness with concept similarity computing, excluding the effects of heterogeneous ontologies.

关键词： 异构本体; 概念相似度; 映射

Keywords : heterogeneous ontologies; concept similarity; mapping

中图法分类号:TP391 **文献标识码:** A

1. 引言

近年来，随着 Internet 的飞速发展，以及人们对网络信息的需求量和准确度要求的提升，研究人员提出基于语义网的语义检索模型。语义网就是要给 Web 上的信息加上注释——本体。而异构本体的出现，则是由多方面原因造成的：网络的自治性可能会导致整个网络中存在多个节点本体；人们对客观世界的认识程度和认识方法不同，以及本体的表示形式和内在逻辑结构的多样性，也会造成描述同一领域本体的差异性。因此，要想实现不同网络节点之间基于语义的信息交流，就必须解决不同节点本体之间的异构问题，实现异构本体间的互操作。

本文的主要工作是研究如何有效的实现异构本体间的语义映射。本文提出的方法，综合考虑到了词汇字面（literal）的含义，以及语义结构方面包括节点深度、节点密度、边权重等多方面的因素，优化了传统的提取字面语义方法和映射算法。此外还充分分析了结构匹配时应考虑的若干问题，使算法更加完善，得到的匹配结果更加准确、高效。

*基金资助： “863”计划课题(2008AA01Z401)；国家自然基金课题(60903028)；教育部博士点基金课题(20070055054)；天津市科技发展计划课题（08JCYBJC13000）

2. 相关工作

根据本体映射所涉及的本体成分不同，本体映射的方法可以分为基于语法的方法、基于实例的方法、基于结构的方法等。基于语法的的方法，利用概念的名称、同义词和概念定义等进行匹配。概念相似度的计算是基于文本的比较；基于实例的方法，是指在进行本体映射时利用概念的实例作为计算概念间相似度的依据；基于结构的方法，在映射时参考了概念间的层次结构，如节点分类关系(父节点、子节点)、语义邻居关系(兄弟节点)等。由于节点的层次关系中蕴涵了大量的潜在的语义，因此在映射方法中利用了这一信息。目前，在本体映射方面的国内外的研究已有许多。其中 Agirre[1]在利用 WordNet 计算词语的语义相似度时，除了节点间的路径长度外,还考虑到概念层次树的深度,概念层次树的区域密度。文献[2][3]综合考虑了概念名称，属性，和包括父子关系，相离关系等在内的多种因素来计算两本体的相似度。Resnik[4]提出了一种根据两个词的公共祖先节点的最大信息量来衡量这两个词的语义相似度的方法。

3. 基于异构本体的相似度计算

针对异构本体映射的情况，本文提出一种综合了字面概念层和本体结构层的更全面、更高效的计算方法。计算公式为：

$$Sim(ID1,ID2) = k * Sim\_literal(ID1,ID2) + (1-k) * Sim\_struct(ID1,ID2) \quad (0<k<1) \qquad (1)$$

Sim_literal(ID1,ID2)为字面相似度计算结果值，Sim_struct(ID1,ID2)为结构相似度计算结果值，k 为调节因子。

3.1 基于字面概念的相似度计算

字面概念相似度计算是指仅从单词含义角度去比较两组本体概念的相似程度。本文借助于普林斯顿大学设计的 WordNet 词典[5]来计算概念的字面相似度。WordNet 是一种基于认知语言学的英语词典，它不是把单词以字母顺序排列，而是按照单词的意义组成一个“单词网络”。

3.1.1 借助 WordNet 将短语转换为语义集的优化算法

一个词汇在 WordNet 中会对应一个或多个同义词集合（Synset）。每个 Synset 表示一个概念或语义。在 WordNet 中每个 Synset 代表一个同义词集合，具有相同含义的词汇在 WordNet 会对应相同的 Synset。因此，可以利用 Synset 的标识符统一地表示本体概念。因为某个单词可能有多种词义，如果将单词的所有词义均加以考虑来进行匹配计算，势必会造成过多的冗余。所以，本文考虑使用一种方法实现提取最有可能描述该单词的词义，从而减小无效计算以提高效率。这种算法利用 WordNet，将其中短语转换成更为符合这个短语希望表达含义的语义集，如公式(2)(3)所示。

$$\text{set\_synset}(\text{w}) = \left\{ S_{\text{k}} \mid \left\{ \begin{array}{l} \dfrac{\left|context_w(S_k) \cap C\right|}{\left|C\right|} \geq \dfrac{\left|context_w(S_i) \cap C\right|}{\left|C\right|} \\ i \in \{1,2..k..n\}, 1 \leq k \leq n, n\text{为}w\text{的语义集个数} \end{array} \right\} \right\} \qquad (2)$$

其中：

$$Context_w(S) = \{S \cup meronym(S) \cup holonym(S) \cup hyponym(S) \cup hypernym(S)\} \qquad (3)$$

set_synset(w)为单词 w 的语义集合；$Context_w(S)$表示在 WordNet 中包含 w 的语义集合 S 产生的语义环境；C 表示本体概念名集合，即表示本体概念的词汇的集合；$|Context_w(S_i) \cap C|$表示与 C 的交集的势；|C|表示集合 C 的势。

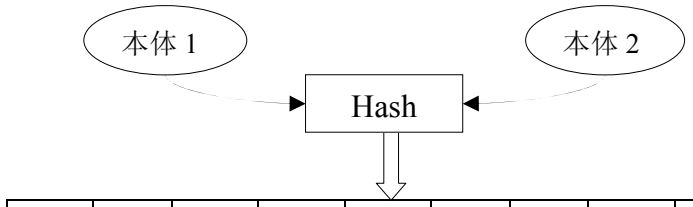
3.1.2 字面概念映射方法

求两个本体中字面概念相似的匹配对，一般的方法是先求取所有可能的映射，再找出符合条件的映射。假设两个本体含有的元素 ID 个数分别是 N1 和 N2，则时间复杂度为 O(N1*N2)。由于实际可能的匹配对远小于 N1*N2 个，必然会存在很多无用计算。为了提高效率，本文采用以语义集 Si 为关键字，进行 Hash 映

射的方法，如图 3 所示。
相似度计算公式为：

$$Sim\_Literal(ID1, ID2) = \max\{2 * N / (\left|D_i^1\right| + \left|D_j^2\right|)\} \quad 1 \le i \le n1, 1 \le j \le n2 \tag{4}$$

其中 N 表示 O1 中 ID1 的一组描述短语 D^1_i 与 O2 中 ID2 的一组描述短语 D^2_j 中相同的 word 数，$|D^1_i|$为 D^1_i 的 word 总数，$|D^2_j|$为 D^2_j 的 word 总数，n1 为 ID1 的描述短语的数目，n2 为 ID2 的描述短语的数目。



S1	S2	S3	S4	S5	S6	S7	S8	S9
O1-1 O2-3		O1-4	O2-1	O1-3 O2-2			O1-8 O2-9	

图 1：两个本体的字面概念的 Hash 映射过程

3.2 基于本体结构的相似度计算

本体结构的相似度计算指的是从本体图的结构角度，包括概念节点的语义距离、边的权重、概念的深度、密度以及信息量等角度出发，对之前经过字面相似度计算初步处理的结果集进行的映射计算模式。

3.2.1 一些相关定义

定义 1(语义距离[6]) 指在本体图中连接两个概念节点的通路中最短路径的边数，为两个概念构成最短距离的有向边数量。当然也可以为边赋予权重，这样距离的计算不是简单的有向边数，而是各条边的权重之和。两个概念的语义距离越大其相似度越低；反之，两个概念的语义距离越小，其相似度越高。

定义 2(概念深度[7]) 在本体图中，从概念节点到根节点 Root 的语义距离为该节点的深度，其中根节点的深度为 1。在本体的层次树中，越靠近底层，表示两个概念代表的术语越具体，则两个概念越相近。同样语义距离的两个概念，相似度随着它们深度总和的增加而增加，随着它们之间深度差的增加而减小。

定义 3(概念密度[7]) 概念节点与其孩子节点所连接的边数为该概念的密度。 认为如果在层次网络中某一局部的节点分布较密集，说明在该处对概念的分类越具体，进行匹配时相似度可能越大。

定义 4(概念结点的信息量[8]) 根据感兴趣度的定义，设信息量 IC 为

$$IC(c) = -\log p(c) \tag{5}$$

$$p(c) = \text{概念节点}c\text{后代节点数目} / \text{本体中节点总数} \tag{6}$$

3.2.2 结构相似度计算方法

结合以上各因素，结构相似度计算采用计算带权值的语义距离并综合考虑概念深度和密度的方法。同时，为边赋权值时主要考虑节点信息量的因素。为了更加直观方便地处理两个独立的本体，为两个本体增加一个虚拟的公共父节点，这样就可以将两个独立本体的映射问题转化为一个本体中的相似度计算问题。计算公式为：

$$\begin{aligned} Sim\_struct(c1, c2) = {} & \frac{\alpha k}{k + Dis(c1, c2)} \times sim\left(Comparent(c1), Comparent(c2)\right) + \\ & \beta\left(b + (1-b) \times \frac{e(c1) + e(c2)}{2}\right) + \frac{\gamma}{2}\left(\frac{d(c1)}{d(c1)+1} + \frac{d(c2)}{d(c2)+1}\right) \end{aligned} \tag{7}$$

其中α+β+γ=1，Dis(c1,c2)为语义距离，e(c1)为概念节点密度，d(c1)为深度(深度和密度的计算方法见定义 2、3) ，k、b、α、β和γ为调节因子。

$$Dis(c1,c2)=\sum_{\substack{c1到公共父节点\\路径上所有节点x}} wt(x,p(x))+\sum_{\substack{c2到公共父节点\\路径上所有节点x}} wt(p(x),x) \tag{8}$$

其中：

$$wt(c,x)=Ls(c,x)*T(c,x) \tag{9}$$

$$Ls(c,x)=-\log(P(c\mid x))=-\log(P(c\cap x)/P(x))=IC(c)-IC(x) \tag{10}$$

Ls(c,x)为祖先后代节点间的信息量差值。当两节点为祖先后代关系时 T(c,x)=1，否则（为 part of 等其他关系）T(c,x)=k，k>1。

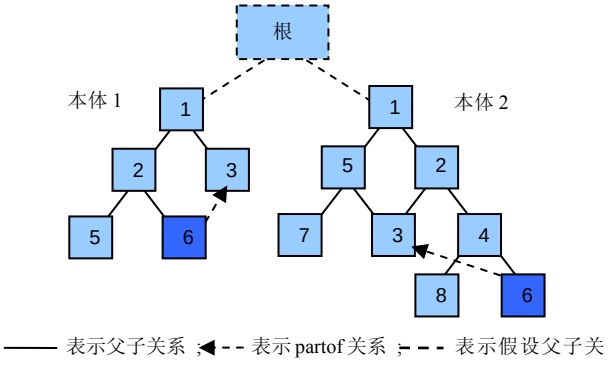


图 2：结构映射关系图

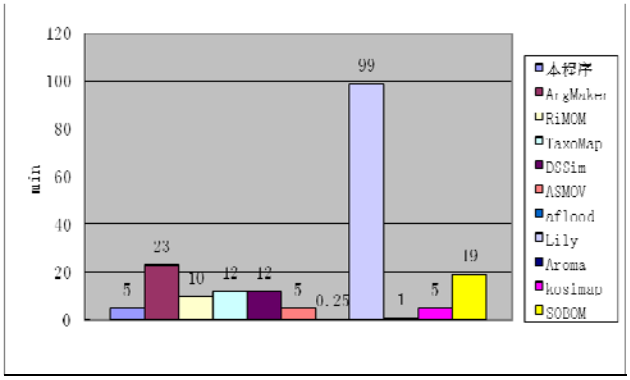


图 3：系统运行时间比较

4. 实验测试

4.1 实验环境设置

实验测试在一台 1.8GHz AMD Athlon 双核 CPU，2GB DDR2 RAM 内存的计算机上进行，开发平台为 Eclipse。实验使用了来自 OAEI 2009（Ontology Alignment Evaluation Initiative 2009）[9] 的本体文件 nci_anatomy 和 mouse_anatomy，它们分别是关于人类解剖学和鼠类解剖学的生物医学本体，主要由“is-a”、“part-of”等关系连接各个概念 ID。每个本体具有不同的概念分类层次，分别包含 3304 个和 2744 个概念。

4.2 对本文提出的概念映射算法进行时间性能上的测试

时间复杂性是评价算法好坏的一个重要指标，如何能够更有效地得到匹配结果是算法必须要考虑的因素。我们对算法的运行时间进行了实际测量，以期得到更接近实际应用的性能结论。2009 年的 OAEI 的本体匹配比赛共有 10 组参加者，图 3 显示了本文提出的算法与参赛程序的运行时间比较。可以看出，本文提出的算法运行时间是 5 分钟，在所有算法中排在并列第 3 位。这主要得益于在字面匹配时的优化算法如借助 WordNet 转化短语为语义集时的优化算法，字面概念映射时采用的 Hash 一次扫描等策略提高了系统的运算效率。

4.3 对算法的查准率和查全率进行比较

我们通过对算法的查准率和查全率进行测试，验证了算法搜索匹配结果的有效性。由于实验所用本体过于庞大，求取专家匹配集比较困难。为了便于实验，本文决定首先从 nci_anatomy 和 mouse_anatomy 两本体图中分别提取等量本体概念得到两个子本体，同时尽量保持本体的整体结构性并且保证两子本体之间存在相似性。然后由领域专家对子本体的概念相似度作出主观评价，得到最佳的映射对集合。最后比较本文提出的算法与传统的单独考虑字面语义匹配算法和结构相似匹配算法的查准率和查全率。

图 4 显示了当抽取的本体概念个数分别为 20，30，40，50 时，本系统和单独考虑字面语义匹配算法和结构相似匹配算法的查准率比较效果。可以看出，两个平凡算法的曲线呈下降趋势。说明随着本体概念数量的增多，得到的映射结果集比较大，同时找到的正确映射结果并没有增多，导致查准率曲线逐渐下降；

而本算法充分考虑了字面和结构等多方面因素后，排除掉了一些干扰映射对，使结果更加准确，致使本系统的查准率明显高于其他两者且更加稳定。图 5 的则显示了三者的查全率比较，发现纯按字面匹配算法与本算法的查全率比较相当或略高一些，结构匹配算法查全率较低。这是因为本算法是在字面匹配算法的得到的映射集基础上，充分考虑结构性的因素对映射集进行筛检，所以得到的正确结果集肯定包含在之前的字面匹配得到的映射集之中，所以查全率不会高于纯字面匹配的算法得到的结果。但两者非常接近，表明我们的算法由于考虑了多方面的因素，并未遗漏更多的匹配。而纯按结构匹配的算法由于没有考虑字面语义因素导致映射不准确，较为明显地影响到了查全率。

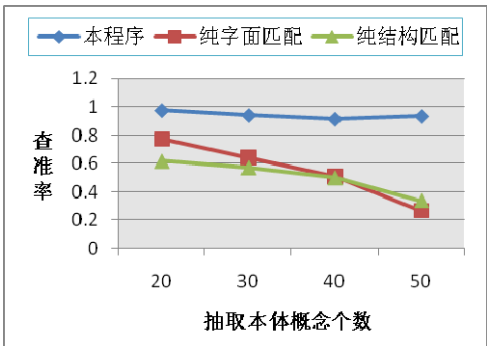


图 4：查准率比较

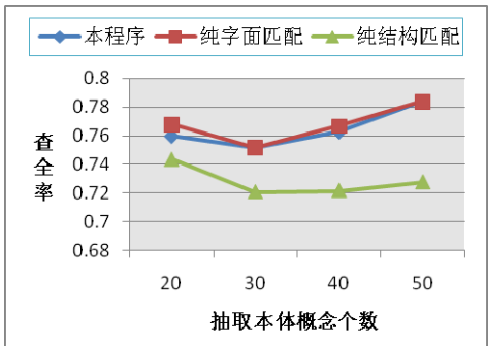


图 5：查全率比较

5 结束语

本文提出了一种针对异构本体的映射算法：综合了字面概念层和本体结构层等多方面的计算方法，使得得出的相似度计算结果更加准确全面，同时更注重了计算效率的问题。通过实验表明，本文提出的基于异构本体的映射算法，比传统的单独考虑字面语义匹配算法和结构相似匹配算法有较高的查全率和查准率。在后续工作中，我们将对算法继续改进，针对复杂本体情况，增加对属性，实例等因素的考虑，完善系统的同时寻求提升计算效率的新方法。

参考文献

[1] Agirre E, Rigau G. A Proposal for Word Sense Disambiguation Using Conceptual Distance.In:Proc. of the 1st International Conference on Recent Advances in NLP. Tzigov Chark, Bulgaria, [s.n.], 1995. The Gnutella Homepage. http://gnutella.wego.com,2002.

[2] Muhammad Fahad, Muhammad Abdul Qadir. Similarity Computation by Ontology Merging System: DKP-OM. In: Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference . 17-18 Feb. 2009.

[3] Yves R. Jean-Marya, E. Patrick Shironoshitaa, Mansur R. Kabukaa,b,*.Ontology matching with semantic verification. In: Web Semantics: Science, Services and Agents on theWorldWideWeb 7 , 235–251.2009.

[4] Resnik P. Using Information Content to Evaluate Semantic Similarity.In:Proc. of the 14th IJCAI. Montereal, Canada: [s.n.], 1995: 448-453.

[5] Kruse P M, Naujoks A, Roesner D, et al. Clever search: A wordnet based wrapper for internet search engines. In: Proceedings of the 2nd GermaNet Workshop, 2005.

[6] MONGE A E, ELKAN C P. The field-matching problem: algorithm and applications[A]. Proceedings of the Second Internet Conference on Knowledge Discovery and Data Mining[C]. Oregon, Portland,1996. 267-270.

[7] Jose L. Sevilla,Victor Segura, Adam Podhorski, Elizabeth Guruceaga,etc . Correlation bet ween Gene Exp ression and G O Semantic Si milarity[J]. IEEE /ACM Transcti ons on ComputationalBi ology and Bi onforma2tics, 2005, 2 (4) : 3302 338.df

[8] Resnik P. Using Information Content to Evaluate Semantic Similarity.In:Proc. of the 14th IJCAI. Montereal, Canada: [s.n.], 1995: 448-453.

[9] http://webrum.uni-mannheim.de/math/lski/anatomy09/