基于概念分层的本体组合匹配策略研究

戴大蒙，慕德俊，吴佳斌

（西北工业大学 自动化学院，陕西 西安 710072）

摘 要：在开放式环境下本体一般由不同组织不同领域专家独立开发，为了让本体能够协同工作或者被重复应用，需要通过本体匹配来解决。文章提出的基于概念分层的组合策略的本体匹配策略利用本体概念的层次性特点，降低匹配过程的计算复杂度，通过组合匹配策略解决单一匹配策略的局限性，并提出了基于语义关系对匹配结果进行一致性判定。

关 键 词：本体匹配，组合策略，概念分层，相似性度量
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-2758(2013)01-0014-05

随着语义应用的急剧增长，这些应用会依赖更多的本体库，众多本体库之间的协作需求也越来越多。为了让这些本体发挥最大的作用，就需要使得开发出的本体能够被重复使用，更为理想的情况下，这些本体能够被组合使用。

然而，本体库是由不同组织乃至不同领域的不同专家共同构建而成。同时知识并不是静态的，任务的改变，领域的改变，甚至同一概念随时间变化而变化时都会使得知识发生变化，引发了本体的不一致性问题，这将阻碍本体的有效重用。本体匹配正是解决这种本体的异构性的关键技术。因此本体匹配作为语义关联分散本体的基础技术，在下一代的语义应用中不可或缺。

本体中的语义信息可以通过结点、边以及它们的关系表示出来，这为处理本体匹配问题提供了有效的途径[1,2]。匹配系统通过计算本体中概念之间的语义相似度来判断它们的对应关系，当两者的相似度较高时认为它们是匹配的，然而无法精确估算 2 个概念的相似程度，只能通过不断地改进匹配算法以求获得最优的结果[3,4]。

对 2 个本体库的概念进行匹配的最基本的方法是对所有的概念逐个进行比较，也就是对于第 1 个本体库的任意概念，都与第 2 个本体库的所有概念逐个比较。这种计算的时间复杂度为进行匹配的 2 个本体库本体概念个数的笛卡尔积。

1 本体匹配与相似性分析

本体匹配的方法一般有基于语法、实例和结构方法等，基于语法是根据概念的定义及其同义词进行匹配而没有考虑概念间的语义，基于实例是考虑概率间的相似性判定，而基于结构是考虑概念节点的层次关系[5]。

这些匹配技术的基础就是基于名称的技术，通过比较概念的名称、注释、标签等，找出相似本体，一般是计算这些描述性字符串的相似度或者基于词语的语言学描述进行相似度计算。基于名称的方法直观，缺点是术语描述方法不一致导致相似性计算结果完全相反。

基于结构的技术是通过比较本体的内部结构，包括本体的属性集，属性之间的关系，通过计算这些内部结构来计算概念的相似性，主要用于具有相同或者相似领域的概念，这是因为相似或相同本体其内部结构也基本相似或相同[6]。

基于语义技术的方法是采用本体概念的语义知识计算本体的相似性，根据本体的相似性度量值来判定本体的匹配性。

收稿日期：2012-03-12 **基金项目**：浙江省公益性技术应用研究计划(2012C23040)及 2010 年浙江省高校优秀青年教师资助项目资助

作者简介：戴大蒙(1975—)，女，西北工业大学博士研究生，主要从事智能信息处理、数据挖掘和粒计算等的研究。

随着本体的急剧增加,本体库包含的信息量急剧增加,特别是在开放式的平台上的本体,本体库的急剧增加导致单一匹配策略难以达到匹配的效率。需要一种组合的方式来完成本体匹配[7,8]。

本文根据概念分层的特征和本体匹配组合策略 2 个方面提高本体匹配的效率和匹配精度。

首先,本体是一个形式化的知识表示形式,它具有严格的层次结构和逻辑结构。因此,在本体的概念层次结构上,将本体泛化程度大致相等的概念划分为若干个层。将概念粒度相差不大的层之间的概念进行匹配,可能蕴含着更多的本体相似的机会,而时间复杂度将大大降低。由于概念从上层开始匹配,如果父节点的概念不相似,则该节点相似的可能不大,于是将这些匹配对的选择舍弃。虽然这种方式引入使得查全率降低,但正是由于这一操作,使得匹配的复杂度大为降低。

其次,本文采用多种匹配策略进行组合,计算本体匹配过程的本体相似度,再通过匹配聚合的方式计算最终的相似度,进行匹配。

2　本体概念分层匹配策略

2.1　系统框架

本文建立基于本体的概念分层方式进行本体匹配,系统的模型如图 1 所示。

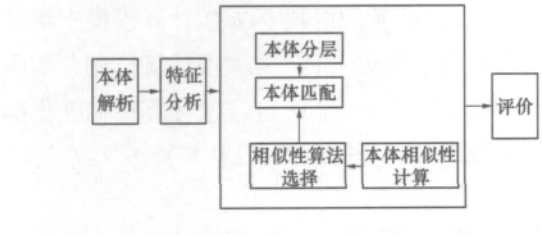


图 1　本体匹配的基本模型

在匹配系统框架中,本体库通过解析和特征分析,映射到分层体系后,根据合适的相似性选择算法进行匹配。

其中本体匹配的步骤如下所述。

1) 匹配概念选择,从进行匹配的本体中选取需要匹配的概念进行比较。

2) 相似度计算,根据不同的匹配策略标识给定 2 个本体的相似度。

3) 相似性聚合,对步骤 2) 产生的多个相似度进行聚合形成新的相似度。

4) 迭代以上步骤,产生合适的匹配结果。

2.2　本体概念分层

本体中每个概念都带有一定的层次性语义信息。在本体的层次体系中,语义信息粒度较大的本体处于层次体系的上层,粒度较小则处于较低层。从概念的语义上来看,语义内容范围越大的概念之间,其相似的机会越大。

根据本体中概念间的继承关系来实现集合之间的映射,再根据信息粒度的大小进行分层,如表 1 中的本体库 O_1 和 O_2,根据描述的概念层次性特性,将本体中的概念进行分层,结果如图 2 所示。

表 1　匹配本体库 O_1 和 O_2

Ontology O_1	Ontology O_2
Object	Thing
Automobile	Car
Motor	Bike
Motorbike	Ship
Nonmotor vehicle	

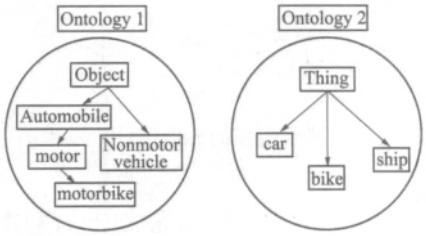


图 2　本体分层体系图

2.3　本体概念分层匹配效果分析

在本体匹配步骤 1) 的搜索步骤选择时,首先从最上层概念进行匹配,如果蕴含的概念高于比较本体概念,则选择其下层本体放入比较队列并依次比较,如果比较结果小于该概念设定的阈值,则舍弃其分支。

采用分层策略的优点是通过本体概念蕴含的层次性属性和本体概念的语义关系,大大减少了匹配次数,缺点是在舍弃分支时可能舍弃了部分正确结果,从而降低需要进行本体匹配的本体应用中的查全率。这种策略的目的就是通过急剧降低本体匹配复杂度,增加本体匹配的效率提高实用性。

3 本体匹配组合策略

本体概念的语义相似性度量是本体匹配的计算基础，根据不同的匹配策略，需要匹配的概念 A 和 B 其相似性计算方式分别阐述。

3.1 基于名称的相似性计算

基于名称的相似性计算如公式(1)所示，根据 Levenshtein 编辑距离进行计算，计算结果映射到区间 [0,1]。

$$sim_{\max}(A,B) = \max\left(0,\frac{\min(|A|,|B|) - ed(A,B)}{\min(|A|,|B|)}\right) \quad (1)$$

基于名称的相似性计算从概念描述的字符串进行比较，出现同义词和多义词的现象，本文通过基于 WordNet 进行提供的 API 接口进行转化，计算流程图如图3所示。

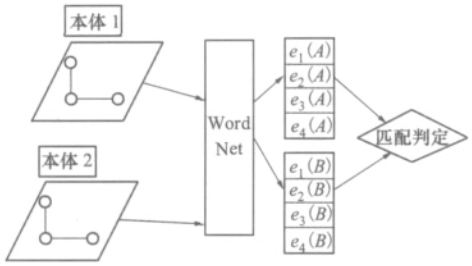


图3 基于名称的相似性计算流程图

在基于名称的本体匹配算法中，针对本体库 O_1 和 O_2，其中1条匹配的结果如图4所示。

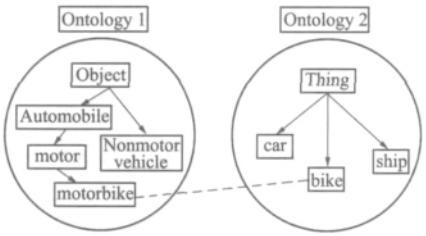


图4 基于名称的1条匹配记录

3.2 基于结构的相似性计算

基于结构的相似性是计算2个概念的内部结构的集合相似度，内部结构相似度通过计算内部结构的属性集合[9]。如公式(2)所示:

$$sim_{ztr}(A,B) = \frac{\sum_{a\in A}\bar{a}}{|\sum_{a\in A}\bar{a}|}\cdot\frac{\sum_{b\in B}\bar{b}}{|\sum_{b\in B}\bar{b}|} \quad (2)$$

3.3 基于语义的相似性计算

本文的基于语义相似性计算方法以心理学领域概念相似性度量模型为基础，结合人工智能领域概念相似性度量模型，建立概念相似性度量模型，具体过程另文阐述。

在分布式系统的本体匹配中，参与匹配的多个本体的描述方式可能不一致，在本文中假设参与匹配的本体均由 DL 描述语言描述。一个基于 DL 描述的本体 $O=(T,R)$，由概念集合 T(TBOX) 和关系集合 R(RBOX) 组成。

设 TBOX T 中的概念 A 和 B 的属性集合 Ψ_A 和 Ψ_B，则 Ψ_A 和 Ψ_B 两属性集合的模糊差集为 $\widetilde{\Psi}_{A-B}$，它是一个论域为 Ψ_A，隶属函数为 $\mu\widetilde{\Psi}_{A-B}$ 的模糊集。且满足公式(3):

$$\mu\widetilde{\Psi}_{A-B}(x) = 1 - \max(SR(x,y)),\ \forall y \in \Psi_B \quad (3)$$

TBOX T 中不互斥(disjointed) 的2个概念 A 和 B 的相似性度量函数为公式(4)所示。

$$\mathrm{sim}^k(A,B) = \frac{f((\widetilde{\Psi}^k_{A\cap B})_\lambda)}{f((\widetilde{\Psi}^k_{A\cap B})_\lambda) + \alpha f((\widetilde{\Psi}^k_{A-B})_\lambda) + (1-\alpha) f((\widetilde{\Psi}^k_{B-A})_\lambda)} \quad (4)$$

式中，$(\widetilde{\Psi}^k_{A\cap B})_\lambda$ 是 Ψ_A 和 Ψ_B 第 k 次计算模糊交集的 λ 截集；$(\widetilde{\Psi}^k_{A-B})_\lambda$ 是 Ψ_A 和 Ψ_B 第 k 次计算模糊差集的 λ 截集；$(\widetilde{\Psi}^k_{B-A})_\lambda$ 是 Ψ_B 和 Ψ_A 第 k 次计算模糊差集的 λ 截集；α 为相对显著性系数，且 $0 \le \alpha \le 1$。

4 多策略匹配结果的组合技术

在采用多种匹配策略对本体进行匹配，需要通过聚合技术对每个匹配产生的匹配结果进行聚合，进而提供最优的本体概念相似性匹配。目前有多种常用的聚合技术，第1种是模糊聚合，将权值和维度数值的每个位置关联，从而给最高值更多的重要性，这在本体匹配组合策略匹配结果选择明显得非常重要，它只允许保留具有最高匹配的结果。第2种是全局相似计算，这种相似计算把本体作为一个整体，而不是模糊计算，主要计算相邻本体，因而这种策略最终结果取决于本体库中的所有概念。第3种是基

于机器学习的方式，通过许多正面例子和反面例子来学习如何对匹配关系分类，常用的学习器有贝叶斯学习、WHIRL 学习、神经网络学习、决策树学习。本文的聚合函数如公式(5)所述。

$$sim_{agg}(A,B) = \frac{\sum w_i * trl_i(sim_i(A,B))}{\sum_{i=1,2,n} w_i} \quad (5)$$

trl 是转换函数，将相似性转换到 [0,1] 区间。w_i 是匹配策略中各个匹配策略计算的相似性权值，权值的不同，各匹配策略的影响因子不一，可以通过人工干预的方法选取最为合适的权值组合，如当 w_i 为 1 时，相似性匹配结果为各匹配策略相似性计算结果的平均值。

根据多策略组合匹配，针对本体库 O_1 和 O_2，其中部分匹配的结果如图 5 所示。

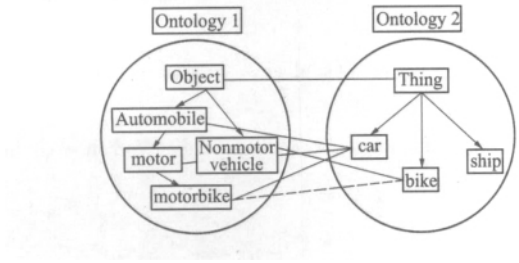


图 5　多策略组合匹配部分匹配示意图

5　本体匹配的一致性判定测试

任何一种知识表示系统都有可能存在冲突和错误，特别是在实际应用环境下基于描述逻辑的知识库存在不一致的可能性非常大，如本体建模的复杂性，本体集成和移植等多种原因均有可能造成不一致。

通过组合匹配策略计算的较高相似性的概念是否存在一致关系需要进一步的评估测试，如进行匹配的 2 个本体库 O_1 和 O_2 中的概念 A，B 是 1 对候选的相似概念，还需要满足对于本体库 O_1 和 O_2 均有 $A \cap B \neq \phi$ 的条件才具备一致关系。

对于组合匹配策略匹配结果的判定性测试，需要根据检测概念所有数据来解决概念对应本体可满足判定问题和 2 个概念之间的关系判定，这就要求知识专家加入大量的规则，因此严格按照描述逻辑定义的概念之间的关系对匹配方案进行评估的效率很低。为了在归并数据之前对匹配方案进行初步评估，本文提出基于语义的概念关系判定方法，判定过程的有以下主要几个步骤。

1) 根据匹配方案，创建概念之间的语义关系。如果 $A \subset B$ 测试概念间是否存在语义包含关系。如果 $A = B$，测试概念间是否存在语义等价关系。如果 $A \approx B$ 测试概念间是否存在相容关系，反之亦然。

2) 在匹配方案中选取概念层次中的子结点，依次判定这些概念间是否存在等价关系、包含关系和相容关系，如果不满足，则根据结果标记该子结点不满足相应的语义关系。

3) 迭代步骤 2)，直到概念层次中概念所有子结点为空。

4) 根据步骤 2) 和步骤 3) 获得的语义关系建立概念的匹配结果判定。

根据判定，针对多组合策略匹配的结果进行一致性判定，判定结果如图 6 所示。

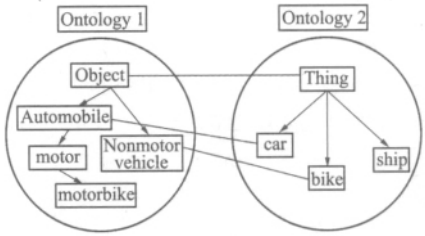


图 6　一致性判定结果示意图

6　结　论

目前提出的许多技术用以解决多本体应用中的本体匹配问题，目的是提高本体匹配结果精确性和一致性，本文提出的基于概念分层的组合策略匹配是根据本体概念的层次性特点，将概念匹配的过程限定在特定的层次上减少匹配可选集，提高本体应用的实用性。通过组合匹配策略并将匹配结果进行聚合从而解决单一匹配方式的局限性，并提出基于语义的方式对匹配结果进行语义一致性判定分析，在下一步的工作中，需要通过定量分析语义的一致性。

参考文献:

[1] Mao M, Peng Y, Spring M. An Adaptive Ontology Mapping Approach with Neural Network Based Constraint Satisfaction. Journal of Web Semantics, 2010

[2] 杨 宁,郭 雷,方 俊等. 语义 Web 作为背景知识的本体匹配. 西安工业大学学报, 2011, 31(2): 174-178
Yang Ning, Guo Lei, Fang Jun, et al. Ontology Matching Based on Semantic Web as Knowledge Background. Journal of Xi'an Technological University, 2011, 31(2): 174-178 (in Chinese)

[3] Euzenat J, Meilicke C, Stuckenschmidt H, Shvaiko P, Trojahn C. Ontology Alignment Evaluation Initiative: Six Years of Experience. Journal on Data Semantics, XV, LNCS 6720, 2011, 158-192

[4] Peukert E, Eberius J, Rahm E. AMC A Framework for Modelling and Comparing Matching Systems as Matching Processes. Proceedings of ICDE, 2011

[5] Ferrara A, Montanelli S, Noessner J, Stuckenschmidt H. Benchmarking Matching Applications on the Semantic Web. Proceedings of ESWC, 2011

[6] Elmeleegy H, Elmagarmid A, Lee J. Leveraging Query Logs for Schema Mapping Generation in UMAP. Proceedings of SIGMOD, 2011

[7] Hedeler C, Belhajjame K, Paton N W, Fernandes A A A, Embury S M, Mao L, Guo C. Pay-As-You-Go Mapping Selection in Dataspaces. Proceedings of SIGMOD, 2011

[8] Farooq A, Arshad M J, Shah A. A Layered Approach for Similarity Measurement between Ontologies. Journal of American Science, 2010, 6(12): 69-77

[9] Marc Ehrig, Steffen Staab, York Sure. Framework for Ontology Alignment and Mapping. Proceedings of the Workshop on Integrating Ontologies, 2005

Composition Strategy of Ontology Matching Based on Hierarchic Concept

Dai Dameng, Mu Dejun, Wu Jiabin

(Department of Automatic Control, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: In an open environment, ontologies are built by the experts from different fields and different independent development organizations. In order to allow the ontologies to work together or be repeated in different applications, we deem that ontology matching can meet the requirements. Sections 1 through 5 of the full paper explain our composition strategy mentioned in the title; their core consists of: the algorithms are based on hierarchic concepts; ontology matching strategies using the concept of the level of ontology can reduce the computational complexity of the matching process; the algorithm uses a combination of matching strategies to solve the limitations of a single matching strategy, and proposes a relationship based on semantic matching to detect the consistency of the matched results.

Key words: algorithms, calculations, computational complexity, flowcharting, hierarchical systems, mathematical models, ontology, schematic diagrams, semantics; composition strategy, hierarchic concept, ontology matching, similarity measurement