《智能信息处理》课程考试

本体聚类与合作机制基于社会选择

刘海鹏

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2020 年 12 月 06 日

本体聚类与合作机制基于社会选择

刘海鹏

(大连海事大学 计算机科学与技术 辽宁省大连市 中国 116026)

摘 要：语义 web 为网页扩展了计算机可理解的、可处理的语义信息，然而由于本体数量激增导致的异构本体现象阻碍了语义的通信与融合．本体合并是解决本体异构的有效途径之一，旨将多个由 agent 构建的异构源本体通过本体合并机制形成一个共享的顶层本体，以期形成一个更大的语义共享空间．本文将本体合并看作是社会选择的一种应用，用于分析个体源本体与决策共享本体之间的关系．由于构建者的背景知识和推理能力不同会对合并结果产生影响，因此本文综合考虑源本体的可信度和一致赞同属性，设计了包含本体聚类器和本体聚集器的本体合并机制．首先，以社会选择和描述逻辑为基础构建本体合并框架和具体流程；在此基础上设计了基于距离的本体聚类算法，以减少不可信本体对合并结果的不利影响；接着对社会选择中的聚集函数进行总结和改进，并将其应用在本体合并中，介绍了积分聚集规则和阶梯性聚集规则．最后，本文对本体聚集规则的一致赞同属性做出分析，并通过对比实验验证了本体合并机制的有效性．。

关键词 语义 web；本体合并；社会选择；一致赞同性；可信度

中图法分类号 TP311.20 **DOI 号** 10.3969/j.issn.1001-3695.2014.01.030

Study on Ontology Clustering and Merging Mechanism Based on Social Choice

Liu Haipeng

(Computer science and technology，Dalian maritime university, LiaoningDalian,116026,China)

Abstract ：*The semantic web extends computer understandable and processing semantic information for web pages，however the phenomenon of heterogeneous ontology caused by the multiplication of ontologies in the same domain has hindered semantic communication and fusion． Ontology merging is an effective solution to ontology heterogeneity in semantic web which focuses on merging a group of individual local ontologies with distinct sources as a shared collective ontology to form a larger semantic shared space． We can view the ontology merging as the problem of social choice such as voting theory and judgement aggregation，analyzing the relationship between the individual local ontology and shared collective ontology． The collective ontology will be affected by diverse background knowledge and reasonable abilities that different ontology builders has． Therefore，in this paper we consider the reliability as well as unanimity，and propose the ontology merging mechanism that includes ontology clustering and ontology aggregation． We firstly describe the framework and steps of ontology merging based on description logic，and design the distance-based ontology clustering to reduce the negative impact of unreliable ontologies． Then we summarize and improve the aggregation functions，apply aggregation rules to ontology merging and present the scoring aggregation rule as well as staged-elimination rule． Finally，we study the unanimity of aggregation rules and verify the effectiveness of ontology merging mechanism by experimental comparison results．

Key words: the semantic web; ontology merging; social choice; unanimity; reliability

1 引言

随着 web 技术的不断发展，语义 web 为网页扩展了计算机可理解的、可处理的语义信息，从而解决了因信息形式异构、语义多重性带来的机器之间无法理解和交互的问题［1］. 知识图谱是大数据的结构化表示，作为互联网资源组织的基础，能够促进语义 web 的发展. 本体是对特定领域中概念及概念之间关系的形式化表达，可以作为知识图谱表示的概念模型和逻辑基础. 因此在语义 web 中，本体可以为 web 上的信息提供语义解释，有利于人机交互［2］. 然而随着特定领域的扩大和本体数量的剧增，不同用户或 agent 根据不同需求、应用来构建本体，他们往往采用不同的建模方式或本体描述语言，但这些本体所描述的内容在语义上有时会有重叠或者关联，这就造成了本体的异构. 由于分布式本体异构的特性阻碍了 web 服务的互操作性，使之成为了语义 web 中亟待解决的问题之一［3］. 本体映射(Ontology mapping) 和本体合并(Ontology merging) 都能有效的解决上述问题. 本体映射是将源本体通过映射规则转换成目标本体的过程，它侧重于发现两个本体中术语或概念之间的语义关联，本体映射不会改变源本体的结构; 而本体合并是指在给定一组源本体的基础上，通过预先定义好的合并规则生成一个新的共享本体，新本体能够为具体语义提供一个共享的词汇表，但这个新本体相对于源本体结构上发生了改变［4］. 具体来说，本体合并是将特定领域中 n(n≥2) 个异构的局部本体(Individual local ontology) 合并成一个新的共享决策本体(Shared collective ontology) 的过程，因此我们可以将此过程看成是社会选择理论(Social Choice Theory，SCT) 中的一种应用. 社会选择是经济学的一个重要分支主要研究和分析个体偏好和社会决策之间的关系，并为聚集个人偏好提供了多种聚集规则和评判准则. 近年来，众多学者将 SCT 与计算科学领域相结合，从而形成一门新的交叉学科“计算社会选择”，在涉及聚集个体偏好的领域有着广泛应用［5］. 例如，社会选择函数在多智能体系统(Multi-agent system) 和推荐系统(Recommender system) 等方面有着举足轻重的意义. 投票理论(Voting theory) 和判断聚集(Judgement aggregation)作为社会选择的经典应用为本体合并奠定了理论基础，其中常见的聚集方式以及一般性质，例如一致性、单调性、中立性等都为本体合并机制的设计提供了方法论. 基于以上，本文从 SCT 的角度出发，综合考虑本体构建者的可信度和本体间的相似度，针对异构源本体的合并问题做出研究. 文章的主要架构如下: 第二部分介绍国内外学者对本体合并的研究现状以及存在的不足之处;第三部分构建了以社会选择和描述逻辑为基础的本体合并框架，并给出了相关的定义和公式;第四部分具体阐述了本体合并机制中的两个关键技术，本体聚类和本体聚集(包括积分规则和阶梯规则) 的具体算法及其相关性质;第五部分采用应用案例和对比实验验证了本文所提出的合并机制的有效性;第六部分针对目前工作存在的问题作出总结并对规划了未来的工作方向.

2 相关工作

为了促进语义 web 的发展，国内外学者对本体映射、本体集成等相关问题做出了大量研究并取得了较为满意的成果. 其中，比较成熟的本体合并系统有 PROMPT［6］，FCAMerge［7］，Chimaera［8］和 SMART［9］等. PROMPT 是一种用于提供本体半自动化映射与合并的算法，能够对本体的类、属性及其关系进行合并; FCA-Merge 是一种自底向上的合并方式，运用自然语言处理和形式化概念分析，在人工交互的基础上完成本体合并; Chimaera 是支持本体合并和不一致性诊断的系统，它能识别不同源本体中具有相同语义的概念和概念之间的关系. SMART 是一种不受平台限制的基于通用知识模型的合并系统，在合并的基础上增加了对本体不一致性的检测，并给出了修复这些不一致本体的方法. 然而这些系统都需要人工参与才能完成，费时又费力，尤其是当源本体数量过大时，系统可能无法精确的对其进行合并. Wang［10］对现有的本体映射方法进行了分类与总结，认为基于机器学习的本体映射技术能较好的解决由于本体规模不断激增给本体映射带来的问题. 在文献［11］中，作者将概念代数相关理论应用于本体合并当中，并提出了 FCA-OntMerg 方法. 这种方法首先将本体规范化成统一的概念代数表示形式，再根据严格的概念格准则进行合并以生成一个新本体，实验结果验证了该方法在解决本体异构问题上的有效性. 唐杰等人［12］将映射问题转化为风险决策问题，以贝叶斯决策为理论设计了风险最小化的本体自动映射模型，实验证明了该方法具有更高的查准率和

查全率．在社会选择理论方面，投票理论的发展和判断聚集的进步都为本体合并奠定了基础［13，14］． 投票理论是将投票者的个体偏好聚集为集体偏好的过程，在投票过程中要平等对待每个候选者，做到“公平公正公开”，常见的投票聚集方式包括 Borda 规则、k-赞成投票规则、单记移让式投票等［15，16］． Borda 计数法和赞成投票制都是较为简单的聚集方法，每个候选项根据选票的数量来确定最终的赢者． 单记移让式投票较为复杂，每一轮中都会淘汰得票数最少的候选者，最后未被淘汰的候选者即为赢者，在此过程中所有的选票在每一轮的计票中都会被计算到． 文献［17］借助工具 OntoCmaps 提出了以图论为基础的投票机制，探讨了在本体学习中概念检测的问题．Dietrich F［18］和 Pigozzi G［19］致力于研究如何将一群个体对命题的判断聚集为一个集体决策，即判断聚集． 在判断聚集中既要保持聚集过程中的公平性，又要关注聚集结果的正确性，并由此提出了多种聚集方式，如基于前提/结论的方法、基于距离的聚集过程等． 然而这些聚集方式均要兼顾公平、公正的性质，需要平等的对待每个候选者并假设所有投票者拥有的背景知识相同，这点在本体合并中并不适用，因为每个本体构建者对领域知识的认知程度并不相同．

3 本体合并框架

本节主要阐述基于社会选择的本体合并框架，首先在 3.1 节介绍描述逻辑及问题模型所需的定义和概念，在 3.2 节中描述合并框架的组成部分和具体流程．

3.1 问题描述与基本定义

语义 web 中的本体是一组公式的集合，而描述逻辑(Description Logic，DL) 是一种用于描述概念和概念间层次关系的 知识表示语言［20］，具有很强的表达能力和推理能力，可以根据 构造算子在原子概念上构造出复杂的术语和关系． 因此，基于 描述逻辑的本体更适用于语义 web 环境下的知识融合．其中最 基本的描述逻辑为 ALC，由原子概念和角色名组成， DL 知识库可以表示为 K = ＜ Tbox，Abox ＞， Tbox 包含描述概念一般性质的内涵知识，Abox 包含描述论域 中特定个体的外延知识．那么基于 DL 的本体可以被描述成两个部分 O = ＜ OT，OA ＞，其中 OT 是术语集 Tbox，OA 是断言集 Abox．术语集 Tbox 用于描述概念及概念之间的关系，形如 C1 ≡C2，C1 C3 (C1，C2，C3 均为原子概念) ；断言集 Abox 用于描 述论域中特定个体，形如 C1 (a) ，R(a，b) (a，b 为论域中的具体 对象，R 是二元关系) ． 解释 I 由解释域和解释函数构成，它能 1154 小型微型计算机系统 2019 年将概念映射到该解释域的子集，将角色名映射为解释域上的二元关系．对于概念 C，若存在一个解释使得则称概念 C 是可满足的(Satisfiability) ． 如果存在一个解释 I 使得本体中 Tbox 和 Abox 所有概念和断言都为真，则称该本体具有一致性 (Consistent) ．

现给定一组有限的候选公式集合 Φ 和专家(本体构建 者) 集合 Agent N = { 1，2…n} ，其中每个 agent i∈N 都可以根 据自身的知识或经验构建一个具有一致性的本体，记为 Oi Φ． 用 CO 表示所有具有一致性的本体集合，此集合中不仅包含了 O1，O2…On，还包含了其他未被 agent 所构建但具有一 致性的本体．本文将 O = (O1，O2，… On) ∈CON 称为一个本 体组合． 由于每个构建者的背景知识和对领域认知程度不同， 本文创新性地为每个构建者都设置了可信度属性，可信度用符号 ri∈［0，1］表示，由 agent i 所构建的本体 Oi 也具有相同 的可信度，则 R = (r1，r2，…，rn) 表示本体组合的可信度向量． 基于上述形式化表示，本文分别定义本体聚类和本体聚集两 个概念

定义 1．本体聚类本体聚类是指根据本体的相似度将 可信度相近的本体划分到一起． 已知类别集合 C 和本体组合 O，分类规则 G: O→C 能使任意 Oi∈O 经过映射后被划分到 一个类 Cj∈C 中．

定义 2．本体聚集定义本体聚集规则为将源本体组合 O 映射到一个共享的决策本体的过程，即 F: CON→2Φ，经过聚集规则生成的决策本体 F(O) 未必具有一致性．

本文的主要研究思路是通过本体聚类算法舍去可信度较 低的本体，从而降低其对合并精度的影响; 再利用聚集规则对 可信度高的源本体进行合并以生成一个共享的决策本体．精准的聚类算法能为本体聚集奠定良好基础，而对于本体聚集 规则来说，本文希望能够得到一个具有一致性的共享本体．

3.2 本体合并流程

从社会选择的角度出发，考虑每个本体构建者

的可信度和本体间的相似度，本文构建了如图 1 所示的面向可信度的异构本体合并框架，主要由三个部分组成，分别是本体聚类器和本体聚集器和决胜规则．本体聚类器是本体聚集的前提，舍弃低可信度本体从而为合并本体的精度提供了保障，它可以根据每个本体的可信度和本体间的相似度对其进行聚类 ;而本体聚集器则是通过某种聚集规则对源本体进行合并以期形成一个共享的决策本体 ;若经过聚集后有多个符合要求的本体则需要借助决胜规则来确定唯一 F(O) ，本文不详细介绍此规则．

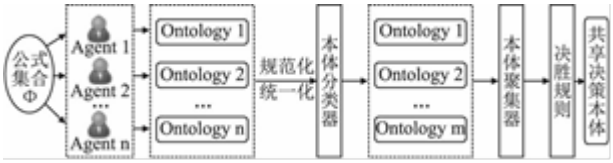

图 1 基于社会选择的本体合并框架

Fig． 1 Framew ork for ontology merging based on social choice

根据图 1 所示的本体合并框架，本文所提出的本体合并机制具体的工作流程有如下几个步骤:

1) 每位 agent 根据自己的经验和知识背景构建出特定领域的本体;

2) 基于描述逻辑，将本体规范成统一的形式 O = ＜ Tbox，Abox ＞ 以便于后续工作的开展;

3) 本体分类器依据设定好的聚类算法和本体的 可信度向量 R 对本体组合 O 中的源本体进行聚类，形成 k 个集合．为了便于计算，本文令 k = 2，设定了 2 个可信度集合，分别为高可信度集合 O^H 和低可信度集合 O^L;

4) 删除可信度低的本体集合 O^L，保留高可信度本体，令本体组合 $O = O^H$;

5) 最后对高可信度本体组合 O 使用本体聚集器和决胜规则以生成唯一的共享决策本体 F(O) ．

4 本体合并过程

对本体进行聚类有助于获得更准确的本体聚集结果，本小节将详细介绍本体合并框架中两个重要的组成部分，本体聚类器和本体聚集器．

4.1 本体聚类器

本体聚类器一方面要根据每个本体的可信度对其进行分类，目标是形成两类本体集合 O^H 和 O^L，分别表示高可信度 本体集合和低可信度本体集合; 另一方面也要考虑本体之间 的相似度，目标是将相似度高的本体划分为一类． 为了衡量两 个本体之间的相似度，现定义距离函数 d: 2 Φ × 2 Φ R + ∪ { 0} 表示两个本体之间的距离，距离越小则表示两个本体越 相似． 常用的距离函数有海明距离、欧式距离和曼哈顿距 离 ［21，22］ ，他们都满足

1) 非负性: d(O i ， O j) ≥0;

2) 同一性: d(O i ， O j) = 0 若 O i = O j ;

3) 对称性: d(O i ， O j) = d(O j ， O i)

本文采用海明距离，即将本体 O_i 转化成 O_j 需要改变最少的公式个数． 理想情况下，我们希望集合 O^H 中的本体尽可 能的相似，而异类本体尽可能不同． 兼顾本体可信度和本体相 似度，本文设计了异构本体聚类器，聚类算法具体流程如 Al- gorithm 1 所示．

算法 1． Distance-based Ontology Clustering for reliability

Input: O， R

Output: O H ， O L

1 Order ontology profile O in the ascending order according to reliability vector R

2 λ L = { the first n [ ] 2 ontologies of the ascending order}

3 λ H = O \ λ L

4 repeat

5 O H = { ontology O h with highest reliability}

6 O L = { ontology O h with highest reliability}

7 for i: 1→n do

8 compute the distance betw een O l and O i d li ;

9 compute the distance betw een O h and O i d hi ;

10 if (d li ≥d hi) then

11 O H ←O H ∪{ O i }

12 else O L ←O L ∪{ O i }

13 end if

14 end for

15 if | { O j ∈O H ∧O j ∈ λ L } | O H ≥ 1 2 then

16 O h ←O h － 1 with the second highest reliability

17 end if

18 until | { O j ∈O H ∧O j ∈ λ L } | O H ＜ 1 2

其中第 1 行是对本体集合进行初始化，将本体

按照可信 度进行升序排序．第 2-3 行根据可信度高低程度粗略的将本体分为两类，显然这样的分类是不够精确的因为没有考虑本 体间的相似度．在第 7-14 行计算每个本体 O i 与最高可信度 本体 O h 和最低可信度本体 O l 的距离，并根据距离远近确定 其所在的类．第 15-17 行表示，若存在大量可信度低的本体与 O h 相似，则说明 O h 可能不准确，因此选取可信度次高的本体 作为类中心并进行迭代更新，直到 O H 中包含的本体可信度较高同时类内本体较为相似时则停止更新(第 18 行) ．

定理 1．　异构本体聚类算法的时间复杂度为 O(n* log 2 n + nt) ，其中 t 是循环次数．

证明: 根据本体可信度向量 R，我们采用堆排序对本体集 合中的 n 个源本体进行升序排序，时间复杂度为 O(n* log 2 n) ．内循环(第 7-14 行) 中分别计算本体 O i 与类中心 O h 和 O l 的距离并确定其所在类别，共执行了 3n 次; 在外层循环 中，由于设置迭代次数为 t，那么执行的频率为 t* (3 + 3n) ; 综上所述，算法的时间复杂度 T(n) = O(n* log 2 n) + t (3 + 3n) = O(n* log 2 n) + O(nt) = O(n* log 2 n + nt) ．

面向可信度的本体聚类算法简单、直观，综合考虑了异构本 体的可信度以及本体间的相似程度，减少了类中心本体选择不当 对聚类结果造成的不利影响; 但对于本体规模较大的情况，其时 间和空间复杂度较高，还需要在后续工作中进一步优化．

4．2 本体聚集

借鉴社会选择理论中的聚集规则，本文对其进行总结与 分类，并在原有的聚集规则之上兼顾本体可信度和本体相似度，形成了积分聚集规则和阶段性淘汰规则两种方式．

4.2.1 积分聚集规则

积分聚集规则通过函数 score: CO->R，将具有一致性的本体集合映射到一个实数集． 每个一致性本体根据积分函数都可以获得某个得分，积分聚集规则的最终目标是将得分最 高的本体作为共享本体，即

$$F(O) = \text{agrmax}_{co \in CO}\ score(CO)$$

在积分聚集规则中，最终生成的共享决策本 F(O) 体必然具有一致性．本文介绍两种积分聚集规则，分别是基于距离 的得分规则和基于可信度的得分规则．

1) 基于距离的得分规则

基于距离的得分规则将每个本体的得分定义为与本体集 合 O 距离的相反数，即

$$score(CO) = -d(CO, O) = -\sum_{i \in N} d(CO, O_i)$$

由上述公式可知，若 score(CO) 越大，则它与本体集合 O 的距离越小．换言之，CO 与其他本体相似越高，更容易得到 agent 的认可; 反之 CO 得分越低，与其他本体相似程度越低．

2) 基于可信度的得分规则

最简单的基于可信度的得分规则定义每个本体的得分即 为自身的可信度，将可信度最高的本体作为共享本体． 显然得分规则过于简单且没有考虑 CO 中的其他本体． 因为可信度高的本体构建者对领域认知也可能会出错，而可信度低构建者也有可能构建出比较完备的一致性本体．

6 总 结

本体合并是促进语义 web 发展的关键技术之一，为了解决目前存在的语义异构问题，以社会选择为理论依据，将不同 agent 构建的异构本体(个体偏好) 通过本体合并机制生成一 个顶层本体(集体决策) 以期形成一个更大的语义共享空间． 由于不同 agent 的背景知识不同，所构建的本体也具有不同 的可靠性． 因此，本文针对异构本体的可信度，设计了本体聚 类器和本体聚集器，本体聚类器为本体聚集器缩减了本体聚 集规模，从而提高了本体合并精度． 然而还存在着一些不足之 处使得这种合并机制无法在真实场景中应用． 一方面，这种基 于距离的本体分类器的时间和空间复杂度都会随着本体数量 的扩大而剧增; 另一方面，本体聚集器的设计中还未考虑社会 选择中的其他重要性质，例如中立性、单调性和独立性等． 未 来的主要工作将致力于设计一种决胜规则和完善这两方面的 不足以不断提升本体合并机制，并将此机制用于描述语义 w eb 服务，使具有计算机可理解的语义．

参 考 文 献

[1]徐伟华,杨蕾,张晓燕.模糊三支形式概念分析与概念认知学习[J].西北大学学报(自然科学版),2020,50(04):516-528.

[2]陈希邦. 基于粒计算与形式概念分析的面向对象（属性）概念的粒描述[D].江西理工大学,2020.

[3]刘萍,彭小芳.基于形式概念分析的词汇相似度计算[J].数据分析与知识发现,2020,4(05):66-74.

[4]李金海,李玉斐,米允龙,吴伟志.多粒度形式概念分析的介粒度标记方法[J].计算机研究与发展,2020,57(02):447-458.

[5]折延宏,胡梦婷,贺晓丽,曾望林.两种多粒度形式概念分析模型的比较研究[J].计算机工程与应用,2020,56(10):51-55.

[6]曾望林. 基于属性粒化的多尺度形式概念分析研究[D].西安石油大学,2019.

[7]张榕宁.国外网络信息检索研究现状[J]. 图书馆论坛,2006,(8):188—190.

[8]杨青,王瑞菊. 浅析网络住处检索中的问题与对策[J].图书馆学研究,2004,(6):82—83.

[9]李爱红.网络搜索引擎的比较研究[J].中国信息导报.1999(1):25 一 26.

[10]徐亨南.人工智能与智能信息检索.信息检索(江西图书馆学刊),2005,35(l):53 一 54.

[11]金海,袁平鹏.语义网数据管理技术及应用[M].北京：科学出版社,2010.

[12]黄果,周竹荣,周亭.基于语义网的信息检索研究.西南大学学报(自然科学版),2007,29(l):77 一 80.