基于粒度计算理论的知识融合模型研究

蒋黎黎[1a,2]，梁　坤[1a]，叶　爽[1b,1c]

(1. 合肥工业大学 a. 管理学院；b. 建设监理有限公司；c. 土建学院，合肥 230009；2. 安徽三联学院 基础部，合肥 230601)

摘　要： 为了解决分布和多源知识的融合与创新问题，提出受控分众分类法，分类结果提高了知识资源的标注精度，降低了知识组织成本。此外，为了消除本体模块间的异构，构建语义一致的领域本体，将粒度计算理论引入知识融合领域，提出微商空间法，对本体模块进行分解与重构，结果使得构建的领域本体更加简洁和语义一致。最后，采用实证分析的研究方法，利用 Protégé 本体编辑工具对提出的方法进行验证，结论表明该方法是有效的。

关键词： 知识融合；粒度计算；受控分众分类法；微商空间理论；Web 2.0；本体

中图分类号： TP182　　**文献标志码：** A　　**文章编号：** 1001-3695(2012)10-3697-04

doi:10.3969/j.issn.1001-3695.2012.10.024

Knowledge fusion model research based on granular computing theory

JIANG Li-li[1a,2], LIANG Kun[1a], YE Shuang[1b,1c]

(1. *a. School of Management, b. Construction Supervision Co., Ltd., c. School of Civil Engineering, Hefei University of Technology, Hefei* 230009, *China*; 2. *Dpet. of Foundation, Anhui Sanlian University, Hefei* 230601, *China*)

Abstract: In order to solve the problem of knowledge fusion and innovation from multiple distributed sources, this paper proposed controlled folksonomy. The result improved tagging precision of knowledge resources and reduced the cost of know-ledge organization. Furthermore, in order to eliminate the heterogeneity of ontology modules and construct the domain ontology with consistent semantic, it proposed quotient-let space methods. It applied granular computing theory in the field of knowledge integration. Quotient-let space method decomposed and reconstructed the distributed ontology modules to form domain ontology with simplicity and semantic consistency. Finally, it adopted an empirical analysis to verify the proposed method by Protégé. The conclusion shows that themethod is effective.

Key words: knowledge fusion; granular computing; controlled folksonomy; quotient-let space theory; Web 2.0; ontology

知识融合是指在特定的环境下通过流程、算法和应用检验来实现分布、异构和多源知识的转换与创新过程[1,2]，以期消除实例知识的冗余和不一致性，提高融合知识的语义规范性和准确性，以便能够基于融合的知识作出正确有效的推理与判断。目前，知识融合的研究主要集中在知识融合体系结构的设计与优化[3,4]及知识融合算法[5]等方面，通过代理、网格计算和中间件等技术进行异构知识的转换和映射，在特定的知识环境中实现融合。

知识融合体系方面，Preece 等人[2]研究了网络环境下的知识融合，包括知识的表达、重用、不确定知识的推理和多 agent 决策。Smirnov 等人[6]针对融合环境和融合目的，提出了简单融合、选择融合、扩展融合、吸收融合和平面融合五种典型的知识融合框架。通过对知识管理任务和流程的划分，徐赐军等人[7]提出一个包含元知识集构建、知识测度指标确定、知识融合算法设计和融合知识后处理等功能模块的知识融合框架。知识融合算法主要有基于本体的融合算法、语义网的融合算法、知识网格的融合算法以及基于主题图的融合算法等[8,9]。基于本体的融合算法中，Kuo 等[10]利用图划分技术提出基于本体的知识融合框架；Laskey 等人[11]运用 OWL 表示概率本体，并提出一个 PROWL 概率本体的输入、存储和推理的系统，实现了多个噪声信息源的融合；Dietmar 等人[12]通过本体实例的自动构建来克服本体知识的冗余和冲突造成的低效率。随着互联网的不断发展，Web 2.0 环境为知识融合提供了新的机遇。Web 2.0 是以用户参与为主要特征的一系列网络应用，包括博客(blogs)、维基(wikis)、社会网络(social network)、资源聚合(RSS)和大众标注等。Web 2.0 注重与用户的交互，在 Web 2.0 环境下产生了丰富的用户生成内容(UGC)，海量的 UGC 为知识融合提供了广泛的知识源[13~15]。然而，用户在 Web 2.0 环境下获取的知识具有不同的粒度特征。因此，本文将粒度计算理论引入知识融合领域，设计基于粒度计算的知识融合算法，消除知识冗余和不一致性，实现知识融合与创新。

1　基于本体的知识组织模式

知识融合效果与知识组织的模式密切相关。与传统的知识元、分类表、叙词表等知识组织模式相比，本体具有更强的、更丰富的知识表达能力。自 1991 年 Neches 等人将其应用于人工智能领域以来，本体在信息检索、电子商务、知识集成等诸多领域得到了广泛的研究。国外研究机构、研究人员已提出多

收稿日期： 2012-02-18；**修回日期：** 2012-04-10

作者简介： 蒋黎黎(1983-)，女，浙江湖州人，博士研究生，主要研究方向为企业管理(lily_xiaqi@163.com)；梁坤(1985-)，男，安徽合肥人，博士研究生，主要研究方向为知识管理；叶爽(1983-)，男，安徽合肥人，监理工程师，硕士研究生.

种领域本体建模方法(如 KACTUSK、methodology)、多种本体构建工具(如 Protégé、Swoop),并构建了多个本体(如 OBO)。

定义 1 本体包含六个元素 $\{C,A^C,R,A^R,H,X\}$。其中,C 表示概念集,A^C 表示每个概念的属性集,R 表示关系集,A^R 表示每个关系的属性集,H 表示概念层次,X 表示公理集。

定义 2 概念属性集 $A^C(c_i)$,概念集 C 中的每个概念 c_i 用来表示相同种类的一组对象,并能用相同的属性集进行描述。

定义 3 关系 $r_i(c_p,c_q)$,关系集 R 中的每个关系 r_i 表示概念 c_p 和 c_q 间的二元关联,并且此关系的实例是一对概念对象 (c_p,c_q)。

定义 4 关系属性 $A^R(r_i)$,用于表示关系 r_i 的属性。

定义 5 概念层次 H。H 是概念集 C 的概念层次,且是 C 中概念之间的一组父子类关系。

定义 6 公理 X。X 中的每个公理是对概念的属性值和关系的属性值的约束,或是对概念对象之间关系的约束。

2 受控分众分类法

分众分类法是 Web 2.0 环境下重要的知识组织方法,其具有丰富的标签资源和用户数据。但是大众标注过程中标签使用的自由性、随意性、不准确性导致了标引词模糊、不一致、同义词激增等问题。受控词表具有词汇规范、严格以及语义关系丰富等优点。区别于分众分类法自下而上、以用户为中心的知识资源组织模式,受控词表采用自上而下、规范定义方式对资源进行组织。然而,随着网络用户数量的增加和问题空间的扩大,受控词表复杂度不断加大,成本高昂,可操作性欠佳。本文将分众分类法与受控词表相结合,使两者优势互补,提出受控分众分类法。受控分众分类法具有两个重要特征:结合了分众分类与受控词表的优势;对资源标注的方式与分众分类法不同。

受控分众分类法结合了分众分类与受控词表两者的优点,在具体操作过程中,采用多轮反复、逐渐优化的方式确定本体构建的标签资源。第一步是利用分众分类法,产生大量标签资源;第二步则运用受控词表对产生的标签资源进行规范化。在第一步产生标签的过程中,分众分类法充分利用了大众智慧,但是标签命名自由度较大,定义不严谨,且由于人们对概念的理解和表述方式的差异,容易产生语义不一致。第二步在第一步产生的标签资源的基础上对标签进行规范,节约了编制受控词表的成本和复杂度,且分众分类法对受控词表还可以进行更新、修改和完善。反复进行这两步,最终得到的标签资源既具有语义规范性又汲取了大众智慧。

受控分众分类法在资源的标签认定上与分众分类法不同。传统的分众分类法对资源所属标签的认定采用取最高频的方法,而受控分众分类法采用综合评价法。例如对于本体 X,大众分类过程中产生了三个标签 A、B、C,若统计结果显示将 X 归入标签 A 的频率最高,那么分众分类法就将本体 X 标引为 A。但是大众观点不一定是正确的,真理往往掌握在少数人手里。随着分众分类法的深入应用,用户数量不断上升,一些稳定的网络虚拟社区将会形成,并出现不同的用户群体。受控分众分类法对不同的网络虚拟社区赋予不同的权重,并对最终分类结果进行综合评价。例如对于电子消费类产品,大学生虚拟社区一般比老年人虚拟社区所赋权重大。

3 问题空间的粒度分解

本文将本体的模块化构建与本体粒度分解可以看做是相反的过程,但是本体构建中的类层次结构与本体分解过程中粒度的粗细可以不同,即详细的类层次结构可用较粗的粒度进行分解,反之粗略的类层次结构在粒度分解中还可以利用较细的粒度丰富。本文利用粒度理论为模块化本体构建中的模块划分提供了依据。

词计算理论、粗糙集理论和商空间理论是粒度计算的三种经典理论。三者都将所讨论的对象的集合看做论域,通过子集来描述粒度。本文在对三种经典粒度计算理论研究的基础上独立提出了微商空间理论。

微商空间理论将宏观粒度计算与微观粒度计算有效地融合在一起。宏观粒度计算方面,根据粒属性,在所有可能的拓扑空间中找出最合适的商空间,从不同商空间观察同一问题,以便得到对问题不同角度的理解。微观粒度计算方面,在确定的拓扑结构中,利用对象属性作为知识基(X,R),讨论任一概念 L 如何用知识基进行表示。对那些无法用(X,R)中的集合的并来表示的集合,借用拓扑中的内核和闭包的概念,引入 $R_$下近似 $R\_(X)$(相当于 X 的内核)和 $R_$上近似 $R^-(X)$(相当于 X 的闭包)对概念范围进行确定。由于对象属性具有极细的粒度,大部分概念都不必确定上下近似值,除非是对象属性本身界定存在模糊性,此时可以将模糊隶属函数引入进来,利用最大隶属度原则确定概念的边界。

下面给出微商空间理论的定义及相关定理。

定义 7 对于给定问题空间 Q,有四元组(X,f,g,T)。其中,X 为论域,f 为粒属性,g 为对象属性,T 为拓扑结构。R 是论域上的等价关系。

定义 8 $[X]$为论域 X 上对应于 R 的商集,将$[X]$作为新的论域,则四元组$([X],[f],[g],[T])$称为原问题空间的微商空间。

定义 9 一个微商空间对应一个信息粒。

定义 10 对于概念 $K(K\subset X)$,若 K 可以表示为若干商集的并,则称 K 是可定义的,否则称 K 是不可定义的或是粗糙的。对于粗糙的概念利用上下近似集描述,并进行知识约简。

相关推论如下:

推论 1 若某命题在粒属性空间中是假的,则该命题在对象属性空间中也是假的。

推论 2 若某命题在两个或多个对象属性空间中是真的,则该命题在这些对象属性空间合成的商空间中也是真的。

4 基于粒度计算的模块化本体知识融合模型

针对领域本体构建过程复杂、难度大以及构建成本高等问题,将模块化的思想引入到本体工程领域,提出模块化本体构建。模块化本体能有效解决异构性、支持多人协同构建和粒度计算,不仅方便了本体的构建,也有利于本体的维护、共享和重用。目前许多学者对本体模块作了定义,Doran 等人[10]基于本体重用目的定义了模块概念,Stuckenschmidt 等人[17]定义了本体模块的通用结构,但却不能明确体现模块特点。本文从粒度计算的角度对本体模块重新定义如下:

定义 11 本体 O 为$\{C,A^C,R,A^R,H,X\}$,$i\subset\{A^C,A^R\}$,则 $O_R=$

$\frac{U}{i}=\{[X],[f],[g],[T]\}$ 为本体 O 的本体模块。这样定义出的本体模块是一系列具有相同等价关系的本体的集合。

基于粒度理论的模块化本体知识融合如图 1 所示。

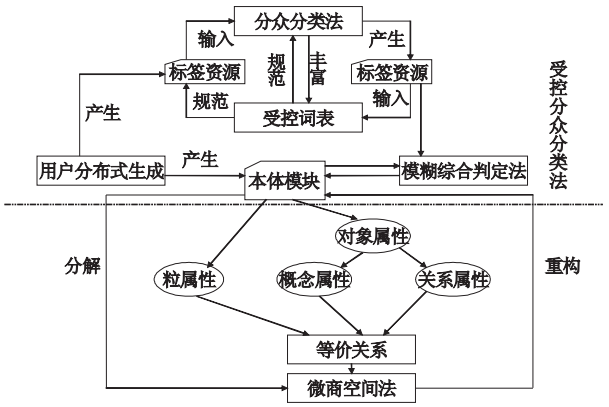

图1　知识融合模型示意图

在图 1 所示的知识融合模型中,本体模块由网络大众分布式构建,并对构建的模块赋予相应的标签进行标志。领域专家对网络大众上传的本体模块和标签作分别处理。利用受控分众分类法,经过多轮反复,形成稳定的标签资源。由于不同用户知识水平上的差异,对上传的本体模块所属的标签认定不同。领域专家对其上传的本体模块具有较深理解,他们对本体模块所属的标签的认定应赋予较高权重。一般用户群体对本体模块所属标签的认定则应赋予较低权重,利用综合评价法对本体模块所属的标签进行综合加权并最终认定。确定本体模块所属的标签之后,若将标签看做一个粒度级别,同一标签下的所有本体模块具有的共同属性又可称为粒属性;然而每个本体模块又具有自己的属性,称为对象属性(包括概念属性和关系属性)。这里的标签相当于一个类,而标签下的每个本体模块相当于类的对象。从属性集(所有的粒属性和对象属性的集合)中选择不同的属性作为等价关系,利用微商空间法,从不同的粒度对分布式构建的领域本体进行商空间重构,并寻求能使得商空间最为简洁和语义一致的等价关系,形成融合后的领域本体。微商空间法具体包含三个步骤:a)收集受控分众分类法产生的标签资源,依据形成的标签对领域本体自顶向下作粒属性一级的商空间重构;b)以对象属性作为等价关系,构建领域本体对象属性一级的商空间;c)依据简洁性和一致性等原则,调整并重构不同粒度下的商空间。简言之,微商空间法对原有的领域本体分类结构重新调整,形成更加合理的大类和细类结构。

微商空间法对分布式产生的本体模块进行分解并重构,克服了不同本体模块的异构性,使最终形成的领域本体具有语义一致性。

5　实证分析

本实验选用 Protégé 来开发本体模块。在魅族社区上,统计发帖者上传的标签,结合相应的产品技术专业术语表(可作为领域受控词表),依据受控分众分类法生成八个标签,即 Andriod、MOTOROLA、HTC、SANSUNG、LG、Smartphones、Microsoft VM、Operating System。分别以 Smartphones 和 Operating System 标签为对象构建两个本体模块,依据本体模块的粒属性和对象属性对本体模块进行粒度分解,并重新组合,使得融合出的领域本体具有语义一致性,如图 2、3 所示。

图 2 中的 Operating_System 本体模块中包含两个对象 Android和 Microsoft_VM,而 Smartphones 本体模块中每个子类的实例均有对象属性 Android 和 Microsoft_VM,且以这两个对象属性作为等价关系形成的商空间较为简洁。所以,在领域本体中将 Android 和 Microsoft_VM 作为粒属性形成新的商空间,每个商空间中的实例对象是原 Smartphones 本体模块的子类的实例对象。在新的领域本体中将原 Smartphones 本体模块的子类作为底层实例对象的对象属性。简言之,在微商空间法处理过程中将对象属性 Android 和 Microsoft_VM 作为粒属性,而将粒属性 HTC、SANSUNG、MOTOROLA、LG 作为对象属性。

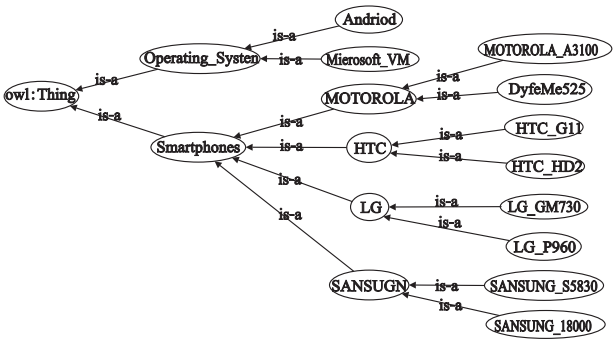

图2　融合前本体模块

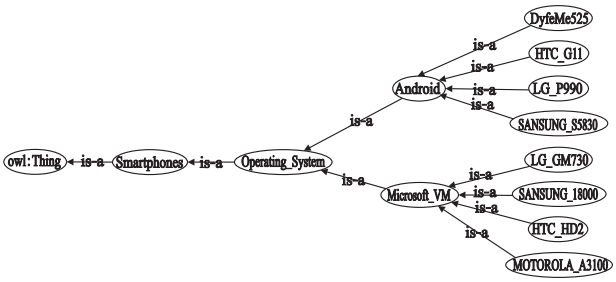

图3　融合后领域本体

在融合过程中对象属性和粒属性的变更是一个重点,其关键代码如下:

```
〈owl:ObjectProperty rdf:ID = "LG"〉
〈rdfs:domain〉
  〈owl:Class〉
    〈owl:unionOf rdf:parseType = "Collection"〉
      〈owl:Class rdf:about = "#LG_GM730"/〉
      〈owl:Class rdf:about = "#LG_P990"/〉
    〈/owl:unionOf〉
  〈/owl:Class〉
〈/rdfs:domain〉
〈/owl:ObjectProperty〉
〈owl:ObjectProperty rdf:ID = "MOTOROLA"〉
〈rdfs:domain〉
  〈owl:Class〉
    〈owl:unionOf rdf:parseType = "Collection"〉
      〈owl:Class rdf:about = "#MOTOROLA_A3100"/〉
      〈owl:Class rdf:about = "#DyfeMe525"/〉
    〈/owl:unionOf〉
  〈/owl:Class〉
〈/rdfs:domain〉
〈/owl:ObjectProperty〉
〈owl:ObjectProperty rdf:about = "#Android2.2"〉
〈owl:inverseOf rdf:resource = "#Android2.3"/〉
〈/owl:ObjectProperty〉
〈owl:ObjectProperty rdf:ID = "SANSUNG"〉
〈rdfs:domain〉
  〈owl:Class〉
    〈owl:unionOf rdf:parseType = "Collection"〉
      〈owl:Class rdf:about = "#SANSUNG_S5830"/〉
```

```
        〈owl:Class rdf:about = "#SANSUNG_I8000"/〉
      〈/owl:unionOf〉
      〈/owl:Class〉
  〈/rdfs:domain〉
  〈/owl:ObjectProperty〉
```

最后,将融合后的领域本体统一转换为 OWL 格式存入数据库便于日后分析挖掘。

在本实例中,由于构建的领域本体规模较小,采用 MySQL 数据库。在 MySQL 数据库中建立 Smartphones_db 数据表,在 Protégé 中,选择 File 菜单下"Convert Project to Format…",在弹出的对话框中选择 OWL Database,填入相关信息,并用 JDBC 将 Protégé 与 MySQL 相连接。

6 结束语

本文结合受控词表和分众分类法的优势,生成本体模块的标签资源;系统地研究了粒度理论并独立提出微商空间理论,应用于模块化本体知识融合领域,消除本体模块之间的异构性。融合后的领域本体具有简洁性和语义一致性。然而目前的研究工作还有待完善,基于微商空间的本体模块一致性检验、本体模块间的推理等将是下一阶段的研究工作。

参考文献:

[1] PREECE A D, HUI K, GRAY W A, *et al*. The KRAFT architecture for knowledge fusion and transformation [J]. Knowledge-based Systems,2000,13(2-3):113-120.

[2] PREECE A D, HUI K, GRAY W A, *et al*. KRAFT: an agent architecture for knowledge fusion[J]. International Journal of Cooperative Information Systems,2001,10(1/2):171-195.

[3] XIE Neng-fu, CAO Cun-gen, GUO Hong-yu. A knowledge fusion model for Web information[C]//Proc of IEEE/WIC/ACM International Conference on Web Intelligence. Washington DC: IEEE Computer Society,2005:67-72.

[4] ANDRES F, NAITO M. Dynamic topic mapping using latent semantic indexing[C]//Proc of the 3rd International Conference on Information Technology and Applications. Washington DC: IEEE Computer Society,2005:220-225.

[5] CHAN P T, RAD A B, TSANG K M. Optimization of fused fuzzy systems via genetic algorithms[J]. IEEE Trans on Industrial Electronics,2002,49(3):685-692.

[6] SMIRNOV A, PASHKIN M, CHILOV N, *et al*. Knowledge logistics in information grid environment[J]. Feature Generation Computer Systems,2004,20(1): 61-79.

[7] 徐赐军, 李爱平, 刘雪梅. 基于本体的知识融合框架[J]. 计算机辅助设计与图形学学报,2010,22(7):1230-1236.

[8] 鲁慧民, 冯博琴, 李旭. 面向多源知识融合的扩展主题图相似性算法[J]. 西安交通大学学报,2010,44(2): 20-24.

[9] ADOLFO G A, CUEVAS A D. Knowledge accumulation through automatic merging of ontology [J]. Expert Systems with Applications,2010,37(3): 1991-2005.

[10] KUO T T, SENG S S, LIN Y T. Ontology-based knowledge fusion framework using graph partitioning[C]//Lecture Notes in Artificial Intelligence. Heidelberg: Springer,2003: 11-20.

[11] LASKEY K B, COST A P C G, JANSSEN T. Probabilistic ontology for knowledge fusion[C]//Proc of the 11th International Conference on Information Fusion. 2008: 1-8.

[12] DIETMAR J, SHCHEKOTYKHIN K, FRIEDRICH G. Automated ontology instantiation from tabular Web sources-the AllRight system [J]. Web Semantics: Science, Services and Agents on the World Wide Web,2009,7(3):136-153.

[13] LIU Wei, YAN Hua-liang, XIAO Jian-guo. Automatically mining review records from forum Web sites[C]//Proc of the 7th International Conference on Fuzzy Systems and Knowledge Discovery. 2010: 2450-2455.

[14] YANG Jiang-ming, CAI Rui, WANG Yi-da, *et al*. Incorporating site-level knowledge to extract structured data from Web forums[C]//Proc of the 18th International World Wide Web Conference. New York: ACM Press,2009:181-190.

[15] AGICHTEIN E, CASTILLO C, DONATO D, *et al*. Finding high-quality content in social media[C]//Proc of International Conference on Web Search and Web Data Mining. New York: ACM Press,2008: 183-193.

[16] DORAN P, TAMMA V, IANNONE L. Ontology module extraction for ontology reuse: an ontology engineering perspective[C]//Proc of the 16th ACM Conference on Information and Knowledge Management. New York: ACM Press,2007:61-70.

[17] STUCKENSCHMIDT H, KLEIN M. Integrity and change in modular ontologies[C]//Proc of the 18th International Joint Conference on Artificial Intellgence. San Francisco: Morgan Kaufmann Publishers, 2003:900-908.