ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.3, March 2008, pp.565−583
DOI: 10.3724/SP.J.1001.2008.00565

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

粒度粗糙理论研究*

陈 波+, 周明天

(电子科技大学 计算机科学与工程学院,四川 成都 610054)

Granular Rough Theory Research

CHEN Bo+, ZHOU Ming-Tian

(College of Computer Science & Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China)

+ Corresponding author: Phn: +86-28-83202688, E-mail: bluesbeyond@vip.sina.com

Chen B, Zhou MT. Granular rough theory research. *Journal of Software*, 2008,19(3):565−583. http://www.jos.org.cn/1000-9825/19/565.htm

Abstract: This paper systematically clarifies granular rough theory from three aspects for its motivation, theory and implementation. Three expectations that motivate granular rough theory are analyzed as follows: 1) to emphasize representative semantics of roughness, with explicitly encoded semantic contexts in underlying representation model; 2) to extend applicability of roughness to a wider range of information sources, with representation model designed to accommodate semi-structured data; 3) to describe a variety of application contexts of information structure, to adapt roughness methodology to disciplines driven by mereology, and to exhibit potentials of combining mereology and computer science in the sense of developing innovative interdisciplinary methodologies, with a pure mereological approach to roughness. From theoretic perspective, granular representation calculus is defined, which plays the role of common representation model for both ordinary information sources and roughness methodology. In terms of this model, corresponding to the notion of lower approximation, border region and upper approximation for roughness, Kernel granule, hull granule and corpus granule are constructed respectively. From pragmatic perspective, upon open source implementation of "Entity-Attribute-Value" model, a rapid prototyping method for granular rough theory is described to provide a test-bed for verification purpose and to apply the roughness methodology for analyzing clinical data more naturally. Significance of granular rough theory, some open problems and further research are summarized.

Key words: granular representation calculus; granular rough theory; pure mereological roughness; semi-structured data representation; entity-attribute-value model

摘 要: 从动机、理论和实现三方面系统地阐述了粒度粗糙理论体系.分析了构建粒度粗糙理论的 3 点动机: 1) 通过显式编码语义上下文的信息表示模型,强调粗糙性的表示语义;2) 通过半结构化思想设计表示模型,扩展粗糙性方法适用的信息源;3) 通过构建纯粹总分学关系上的粗糙性,描述丰富的信息结构应用语境,扩展粗糙性方法到总分学推动的领域,并展示结合总分学和计算机科学创建新型跨学科方法学的潜力.理论上定义了粒度表示演算,使其兼具一般信息源和粗糙性方法底层表示系统的双重功能,在此基础上构造内核、外壳及主体信息颗粒,分别对应粗糙性的下界近似、边界区域及上界近似概念.实现上,提出了通过"实体-属性-值"模型开源系统进行粒度粗糙理

* Received 2006-11-17; Accepted 2007-08-03

论快速原型化的思路,从而提供实验平台验证理论的正确性,同时,更自然地对临床数据进行粗糙性分析.作为总结,阐述了粒度粗糙理论的意义、未解决问题及未来的研究方向.

关键词: 粒度表示演算;粒度粗糙理论;纯粹总分学粗糙性;半结构化数据表示;实体-属性-值模型

中图法分类号: TP391 **文献标识码:** A

粗糙集理论(rough set theory)[1−4]是波兰数学家 Zdzislaw Pawlak 创建的一门独立软计算方法论,实现了 Frege 关于含糊性(vagueness)的思想:通过集合的边界区域而非部分隶属关系(如模糊集)表达的非精确性.Polkowski 和 Skowron 在粗糙集理论基础上提出了粗糙总分学(rough mereology)[5−7],该理论将粗糙性概念引入波兰哲学家和逻辑学家 Stanislaw Lesniewski 的总分学理论(mereology)[8,9],扩展了原始总分学,应用于复杂对象系统的合成质量控制等领域.

根据对粗糙集和粗糙总分学理论内涵的分析[10−12],本文提出了一套新的面向粗糙性表示语义的粗糙性理论:粒度粗糙理论(granular rough theory)[13],其核心是将粗糙方法处理的信息系统通过一套半结构化的粒度表示演算(granular representation calculus)来表达,并在该演算系统的概念体系中,运用信息颗粒之间的 Lesniewski 总分学关系构造下界近似、上界近似、边界等相关粗糙性概念.此后,粒度粗糙理论被逐步扩展,包括适应多智能主体系统[14],用粒度表示演算表达 Internet 信息资源[15],以及以粒度表示演算为基础,构造本体驱动的 Web 信息系统框架,用于语义网格的知识层原型[16].

本文主要从动机(第 1~3 节)、理论(第 4~5 节)、实现(第 6 节)3 个方面着手,深入探讨粒度粗糙理论的构建动机,规范了粒度表示演算中的概念及运算的定义和标记,系统阐释了利用粒度表示演算,在纯粹总分学关系上构造粒度粗糙性的原理,提出了一种快速原型化的方法,并指出了粒度粗糙理论的开放问题及研究方向.

1 显式强调粗糙性方法的表示语义

粗糙集理论研究的对象是二维决策信息表,称为决策信息系统 $I=(U,C\cup\{d\})$,其中,U 代表系统研究实体的全集,C 代表所有条件属性,d 表示唯一的决策属性(对于有多个决策属性的系统,多个决策属性的组合可以定义为一个单独的决策属性).粗糙集理论的主要思想是,在决策表中按照不同个体在某些属性集合上取值的不可区分性(indiscernibility),将实体全集划分为若干关于条件属性的等价类,将这些等价类与决策属性取值上划分的等价类进行比较,按照两种等价类之间的集合包含关系,可以将等价类划分成 3 类:条件属性集合取值完全确定,部分确定,完全否定决策属性取特定值,由此构成了特定集合的下界近似、上界近似,两种近似之间的差异部分,即为表达 Frege 含糊性思想的集合边界区域.为便于本文后续举例说明各种概念和情况,表 1 给出了经典粗糙集理论处理的一个样例信息系统 I^*.为简便起见,假设不考虑 I^*属性间的相互依赖关系(属性依赖在给定信息系统中是客观确定的,因此需要进行属性约简,该问题是粒度粗糙理论亟待解决的开放问题.信息表 I^*用于后续各种定义和运算的用法示例,不考虑其属性依赖不影响理论本身的正确性),另假设表中数字代表对应属性的具体值,不同属性相同数字值具有不同含义.

Table 1 Sample information table I^*

表 1 样例信息表 I^*

	c_1	c_2	c_3	c_4	c_5	d
u_1	0	1	0	2	4	d_1
u_2	1	0	1	2	3	d_2
u_3	0	1	0	2	3	d_1
u_4	2	0	2	1	2	d_2
u_5	2	1	1	1	4	d_1
u_6	2	1	1	0	1	d_1
u_7	1	0	2	0	0	d_1
u_8	0	1	1	2	3	d_3

考察粗糙集理论和模糊集理论,单纯从元素对集合的隶属关系角度来看,位于粗糙集上下界近似之间的边界元素与集合之间,的确可以表征为隶属函数取值在(0,1)之间的模糊隶属关系.但在模糊集的定义过程中,先验

的隶属度概念是人为设定的,不受任何底层模型的约束,其语义仅仅限定了元素与集合之间的隶属关系.粗糙集与模糊集差异的核心体现在其定义中显式地引入了决策信息表,粗糙性的构造完全基于客观存在的信息表,信息表的模式元数据隐含了实体、属性、值之间的内在联系.粗糙集理论呈现了信息表中由条件属性所描述的概念,在信息表的水平方向上映射,通过上下界近似的方式表示决策属性所描述的概念.这种近似方法不强调特定元素与集合之间的关系,而是强调从不同角度观察实体类获取的概念相互表示.从我们的观点来看,粗糙性研究最重要的贡献是提供了一种近似表达概念的途径.在决策信息系统中,由决策属性取值所陈述的对实体相关事实的判断以及这些状态的不可区分性在实体上划分的决策等价类构成了要表达的目标概念(target conception),而表达目标概念的素材,是实体在条件属性上表现出来的客观状态以及由这些状态的不可区分性在实体上划分的条件等价类,称为源概念(source conception).粗糙性方法学的表示语义(representative semantics)是使用特定条件属性组合取值特征描述的若干实体类来近似表示决策属性取值特征描述的一个实体类.表示语义规定了目标概念只有构成一个实体类的总体(即等价类)才具有意义,否则就是意义不明确的平凡概念(trivial conception).经典粗糙集理论的不确定性(含糊性)在于用源概念表达目标概念时引入了不确定的边界,但是源概念及目标概念本身是确定的.例如,文献中最常出现的疾病诊断,生病与否这个决策属性本身的概念是确定的,条件属性中高热、咳嗽等症状也是确定的,只是在用体温等条件属性描述的源概念来近似决策属性划分的“患病”这个目标概念的时候,出现了不确定性.而扩展的模糊粗糙集的一种,就是允许源概念本身是一个具有模糊性的概念.例如,在源概念中描述“被诊断人一定程度地出现某种症状”,这里的“一定程度”指明了成员和集合之间隶属关系的不确定性,与在此基础上进行后续粗糙性分析引入的不确定性具有不同的含义.由此可见,表示语义是粗糙集区别于模糊集而独立成为一种软计算方法论的重要特征.

由于忽略了粗糙性方法的表示语义,存在一种较为典型的使用粗糙集的误区,也就是把粗糙集的“概念近似”简单地理解为“对实体名字标识符集合的近似”.例如,在 I^*中随机地指定几个实体名字标识符构成的集合 $X=\{u_1,u_2,u_4\}$,然后按照粗糙集的上下界近似方法,求得某些属性相关的实体名字标识符构成的两个集合 $\underline{X}=X_1$, $\overline{X}=X_2$,满足 $X_1\subseteq X\subseteq X_2$.该例子表面上看是无懈可击的,但却陷入了误区.上述例子最重要的错误,首先在于被近似表达的目标概念所对应的实体类应该是客观确定的,而不是随机指定的.从决策属性角度看,$X=\{u_1,u_2,u_4\}$这个实体集合在决策属性取值上不能代表一个以决策属性划分的等价类(既可以取值 d_1 也可以取值 d_2),也不能代表等价类的组合(取值 d_1 或 d_2 的实体还有其他的),它所表达的即为平凡概念,对平凡概念的近似是无意义的.由于这种无意义的概念近似最终获得了一个不能从表示语义上说明的“粗糙集”,该结果由两个实体名字标识符的集合分别充当下界与上界近似,使用者只能从传统集合论的观点来看这个“集合近似”.在这种观点下,粗糙性方法的结果仅仅是寻找原集合的一个子集和一个超集,以上下界逼近的方式来近似原集合,并比较不同的近似结果的优劣.但是,如果粗糙集主要是一种集合近似的方式,原集合 X 在粗糙性分析最初就是给定已知的,为什么还要进行近似呢?直接令 $X_1=X=X_2$,是否就获得了最佳的近似结果呢?粗糙集的动机,是要通过这个近似的过程来揭示隐藏在条件属性和决策属性所确定的概念之间的关系.在整个粗糙性分析过程中,以“更好地近似了 X”作为一种系统质量标准是不具有任何意义的.

Pawlak 粗糙集分析方法在理论体系上并不存在必然造成上述错误的缺陷,但是由于数据粗糙性分析的过程中,主要操作对象和结果都是在由不同属性限定的实体名字标识符构成的集合上,因此隐式的语义上下文一旦被忽略,就极易造成上述误用.对于结果而言,脱离了附加的决策规则,上下界对应的实体名字标识符集合也就不存在实际的意义.粒度粗糙理论研究的意图是寻求以 Pawlak 粗糙集理论为基石,适当调整粗糙性分析过程中源概念和目标概念的表示模型,以显式的方式将隐藏在信息表模式中的语义上下文编码在粗糙性构造过程中,并将粗糙性方法的表示语义明晰地反映在最终构建的粗糙性结果上.

2 通过表示模型扩展粗糙性方法学的适用范围

经典粗糙集理论研究的信息系统特指结构化的单一决策信息表 $I=(U,C\cup\{d\})$.但从实际应用来看,粗糙性分析作为一种知识发现方法,决策信息系统 $I=(U,C\cup\{d\})$可以是任意表示模型下的信息源,该信息源通过条件

属性 C 和决策属性 d 描述了 U 中的实体.在本文中,除非特别指明在经典粗糙集理论范畴,“信息系统”或“系统”均泛指任何可能成为粗糙性数据分析对象的信息源.而“系统的结构”则是指信息源通过基本信息单元组合成为有意义的整体时,在信息源内部基本信息单元之间呈现的相互关系.这里的“结构”与“(半)结构化”中的用法不同,后者主要是指众多数据是否呈现一种表达形式上的同质性,是否易于抽取出统一的元数据模式.

为了强调粗糙性方法的表示语义,粒度粗糙理论需要构造一种新的表示模型,该表示模型将经典粗糙集理论中抽象的实体名字标识符集合转换成显式包含语义上下文的单元,用于表达信息系统中的各种概念.如果这种新的表示模型能够更普遍地描述各种现实存在的信息源,同时又作为粗糙性构建的底层模型,就能自然地将粗糙性分析方法应用到更广泛的领域中.

现实信息源从可能的表示模型来看,主要分为结构化和半结构化两大类.

结构化的信息系统典型模型是实体-关系(entity-relation,简称 ER)模型.ER 模型构成了关系型数据库的基础,参见文献[20,21].通过将数据实例表达为具有固定结构的若干字段,形成关系数据库中表的一行记录.ER 模型中,表的字段结构以及字段类型是在定义数据库模式时通过元数据确定下来的,后续记录完全遵循已有数据库的模式定义,每一行记录都包含了实体不同属性相对应的若干事实,不同行的记录虽然包含有不同的属性值,但在结构和语义上是同质的.在主流的商务应用环境中,ER 模型简洁的二维表结构之上构筑的商用关系型数据库管理系统提供了管理数据生命周期的高性能平台.

即便在结构化信息系统范畴内,经典粗糙集理论在表示模型上由于采用的是单信息表,也不能直接匹配关系型数据库中多表存储的实际情况.为此,Milton 等人研究了多表信息系统中利用粗糙集进行关系型学习(relational learning)的问题[22].其主要解决方案是,首先对多个信息表进行组合,生成唯一的信息表视图,然后进行后续的分析工作.这也反映出经典粗糙集理论在表示模型上的确存在着一定的局限性,它要求非单信息表的信息系统经过预处理成为标准形式才能应用粗糙性分析,而对于下面将讨论的半结构化数据环境,则需要更大的预处理开销.

半结构化数据(semi-structured data)是近年来研究的一个热点方向,涉及到众多不同的应用领域.这里考察以半结构化数据为主的临床医疗数据管理领域和 Web 信息系统领域的现存表示模型,探讨粒度粗糙理论为获得更自然而广泛的适应性,其表示系统应采纳的设计思想.

在临床研究数据管理系统(clinical study data management system,简称 CSDMS)中,基于 ER 模型的主流商用关系型数据库管理系统面临着一些特殊困难:数千种检查项目类型,每种都要获取相应的生理参数,在 ER 建模中意味着一个病例表需要有数量众多的字段,而最大字段数在商用关系型数据库中是存在限制的;新的医学检查技术总是不停地出现,而一些现存检查手段也会出现具体细节的调整,每项检查都会涉及到数量不等的患者新属性,这些属性类型需要方便地加入到病例库中进行记录、检索,因此,临床病例数据具有普通应用所不具有的不稳定性.在 ER 体系中,频繁地修改表结构或者数据库模式是不可接受的;对于每个病例而言,该患者所接受的实际检查的数量相对于所有可用的检查类型仅仅是非常小的一部分,ER 表中每行数据都对应所有检查涉及的属性,因此该表中大量的属性值为空,表的稀疏度很高,存储及访问存在着很大的资源浪费.

对上述困难的典型解决方案是采纳实体-属性-值(entity-attribute-value,简称 EAV)模型.该模型是一种适应存在大量异质数据操作环境的特殊设计范例,也称为行建模(row modeling).EAV 模型的基本特点是,存储表的每行分为 3 部分,保存实体标识符(可能附加对应属性取值的时间戳)、属性名以及属性值.每行记录描述了实体一个特定属性相关的单一事实,保存在系统中的不同实体数据,不仅仅是属性值上的差别,还存在属性种类和数量上的差异,因此,这些数据是异质的.由于数据库模式定义对应的元数据全部作为数据行存储在表中,所以,需要记录的实体属性类型可以按照实际需求,作为一个新的行动态添加.在存储结构上,如果把传统的 ER 数据表看作一个二维的矩阵,那么 EAV 模型中的数据表可以看作稀疏矩阵的存储方式,实体标识符和属性名限定了属性值在矩阵中的位置.EAV 模型作为主流数据库设计依据具有很大的争议,因为这种思想事实上是将数据库模式元数据与实际业务数据的管理工作部分地融合起来,使数据库的设计与使用不能完全分开,并且其高度的灵活性导致了数据库查询的实现更为复杂,查询性能上也存在诸多问题需要额外处理.基于此,EAV 模型可以看作

是对成熟的 ER 建模技术的补充,大部分 EAV 系统的实现是在现有的 ER 模型的关系型数据库上扩展形成的,由于有效地利用了成熟 ER 系统提供的数据管理功能实现,EAV 系统中经常被质疑的“重新创造车轮”的问题可以获得较好的解决.目前最成功地使用 EAV 模型设计的开放源码 CSDMS 是耶鲁大学医学信息中心 Nadkarni 等人开发的 TrialDB[23−25].Nadkarni 等人指出,在医疗机构中部署的主流商用 CSDMS 产品 Oracle Clinical,ClinTrial 和 MetaTrial 均采用了 EAV 模型来表示临床数据[26].美国国家卫生研究所(NIH)的临床信息学管理系统项目组(Clinical Informatics Management System Project)发布的测试数据表明[27],在基于 EAV、结构化(ER)以及 CLOB(character large object)这 3 种模式存储的临床数据系统中,EAV 表现出了最佳的性能,这种模型最适合作为临床数据分析与报表所基于的数据仓库结构.

相对于临床数据管理领域,Web 上的半结构化信息资源表示模型具有更多的形态.从较早的 Web 信息访问实践来看,正如 Cohen 在文献[28]中所指出的,存在两类主要的系统模型,分别是信息检索(information retrieval,简称 IR)和知识整合(knowledge integration,简称 KI).前者主要是以关键字检索为代表的各种搜索引擎,后者是以数据库管理系统的思想来处理 Web 信息.在 KI 模型中,对半结构化数据的表示大都采用了特定的表示模型.较有影响力的 KI 系统是 Papakonstantinou 等人提出的 TSIMMIS[29].TSIMMIS 系统采用了对象交换模型(object exchange model,简称 OEM)作为半结构化数据表示模型.OEM 是一种轻量级表示方法,后来被用在重要的半结构化数据查询语言 Lorel[30]中.OEM 中的每个对象都具有以下结构:(label,type,value,object-id),其中,标签(label)是一个描述当前对象要表示的概念名称,类型(type)表示对象取值的数据类型,值(value)是当前概念的具体取值,而对象标识符(object-id)提供了对该语义单元的引用.当数据类型为字符串、整型等基本数据类型时,对象是原子对象;而当对象取值为(label,object-id)的对象引用集合时,该对象为复杂对象.例如,(city,string,“巴黎”,&13)是一个原子对象,而(address,set,{(city,&13),(zip code,&12),(country,&14)})是一个复杂对象.OEM 表达的数据可直接转换成图,其中,对象作为图的顶点,而标签作为图的边.这样一个信息源的描述即可建模为一个 OEM 图.在 OEM 图中,四元组中的“类型”作用被弱化,类比 EAV 模型,OEM“对象标识符”对应 EAV“实体标识符”,OEM“标签”对应 EAV“属性名”,OEM“取值”对应 EAV“取值”,两者虽然是各自独立发展的表示模型,且具体存在一些差异,但在基本的表示思想上具有一定的共性.

随着语义 Web(semantic Web)[31]相关概念的发展,为了使 Web 上的信息资源便于被计算机处理,资源描述框架(resource description framework,简称 RDF)[32]成为业界描述 Web 信息源的规范之一.从 RDF 的模型理论语义[33]来看,在用 RDF 描述资源时,最基本的表示单元是一个{谓语(predicate),主语(subject),宾语(object)}构成的三元组,表达了语言学角度陈述句“〈主语〉〈谓语〉〈宾语〉”所包含的语义.上述的三元组可描述为如图 1 所示的有向标记图.对应于信息源描述,三元组“主语”对应“资源(resource)”,“谓语”对应“性质(property)”,“宾语”对应“文本(literal)”,RDF 语句重新表述为“〈资源〉〈具有性质〉〈取值文本〉”.由此可见,同样是描述半结构化信息源的 RDF,在最基本语义单元表达上,与前面 EAV 和 OEM 模型是相似的.

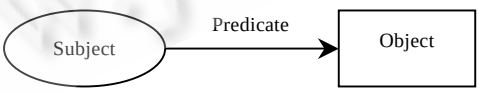

Fig.1 Directed diagram for RDF triple

图 1 RDF 三元组的有向标记图

再来看国内学者研究成果中的一种半结构化数据表示模型 SDOM[34].SDOM 是一种无模式限制的自描述对象模型,并在此基础上设计了一套查询语言.该模型的最基本语义单元,即每个对象的结构也采用了类似 OEM 的四元组方式(标号,类型,值,OID),其结构设计思想与上述各种模型也是一致的.

根据以上分析,设计粒度粗糙理论的数据表示模型时应充分考虑信息源的数据异质性,适应半结构化数据表示,而以“属性-值”为范本的元组模型是一种符合现有技术趋势的选择.粒度粗糙理论在选定了元组作为其最基本的信息单元后,只要适当地定义这些基本单元的结合方式,就能有效地描述上述各种复杂的信息源.再在这种表示系统之上等价地构造出粗糙性概念,就构建起了粗糙性方法学应用到各种复杂信息源的桥梁.

3 构建基于纯粹总分学关系的粗糙性方法

通过以元组为基本描述单元的表示系统,粒度粗糙理论要处理的信息源被表达为许多基本元组按照一定方式的组合,用于描述不同数量的实体在不同属性组合上呈现出来的事实.这些不同粒度的元组及其组合统称为信息颗粒(information granule),而在给定属性组合上呈现的事实称为示象(aspect).信息颗粒不再是传统粗糙集理论中的简单实体名称标识符,而是编码了语义上下文的独立概念单元.信息单元之间的相互关系表示了与现实信息系统结构相关的信息,需要一种比抽象集合论隶属或包含关系具有更丰富的应用语义内涵,但表达形式统一的新型关系.总分学关系作为部分到整体关系的统称,具有丰富的现实语义,Artale 等人在文献[35]中归纳了 Winston,Chaffin 及 Herrmann 的 WCH 分类[36],“部分到整体”关系可以适应以下 6 种不同类型的现实语境:

- 构件/集成对象(Component/Integral-Object):集成对象的特点是具有一定结构,其构件可以拆分,并各自具有一定的功能,如“车轮-汽车”.
- 成员/集体(Member/Collection):描述集体中的成员概念,这种概念中的成员并不具有集体所呈现的整体功能特性,但是成员可以从集体中分离.如“树-森林”.
- 分块/大块(Portion/Mass):作为整体的大块被认为是由同质的分块聚合而成,在分块成分上与大块相同,并能被分离.例如“盐粒-盐”.
- 成分/对象(Stuff/Object):表达了事物的组成成分,通常表达“由…制成”,例如“钢铁-自行车”.
- 阶段特性/活动(Feature/Activity):指示活动的一个阶段,与构件/集成对象相似,但阶段不能从活动中被分离出来,例如“付款-购物”.
- 地点/地区(Place/Area):在被不同对象占据的区域上的空间关系,例如“沙漠-绿洲”.

上述 6 种总分学关系的应用语境可以描述现实客观世界的事物,其表达能力足以表述信息系统的信息颗粒之间的相互关系,进而表达信息颗粒通过总分关系相互作用产生的结构信息.在本文后面详细讨论的粒度粗糙理论表示系统中,若仅限于对经典粗糙集理论的单信息表进行描述,信息颗粒之间的关系主要是呈现成员/集体的总分学关系语境.其他几种总分学关系语境对于描述更复杂信息系统将发挥更大的作用.例如,当信息系统需要描述 OEM 中复杂对象时,复杂对象的“值”与复杂对象之间即为构件/集成对象关系,同样的关系也适用于描述多表信息系统或 XML 树节点的某些结构;当信息系统具有时序特征时,就可以阐释为阶段特性/活动关系;当处理空间信息时,信息颗粒之间则需要地点/地区关系来建立联系.

为了阐明将总分学引入到粒度粗糙理论的表示系统最重要的动机,首先需要了解总分学的基本理论体系.

3.1 总分学理论体系基础

“部分与整体关系”的分析从原子学派开始,通过古代和中世纪的本体论者的著述延续至今,一直是西方哲学研究的焦点之一,它通过 Brentano 及 Husserl 的《逻辑调查》(logical investigation)进入当代哲学研究领域.Lesniewski 总分学与后期出现的 Leonard 及 Goodman 的个体演算(calculus of individuals)[37]是对该范畴的两种相似但存在重要差异的形式化理论.关于总分学的综述性介绍参见文献[38].Lesniewski 总分学相关理论体系是波兰逻辑学家 Stanislaw Lesniewski 为解决经典 Cantor 集合论的一些理论问题而提出来的[2].Lesniewski 的理论体系包括 3 个层次递进的公理演绎系统[8]:第一要义(protothetic)、本体论(ontology)和总分学(mereology),分别定义了其“核心逻辑观点”、“客观存在的理论”及“部分整体关系理论”.第一要义是 Lesniewski 关于命题、连接符、连接符构成算符以及更高阶由算符构成算符的无限可扩展逻辑;本体论是其关于名词、动词、名词性算符或动词性算符的无限可扩展逻辑.第一要义与本体论一起构成了可与 Russell 及 Whitehead 的《数学原理》(principia mathematica)[39]体系相当的独立数学逻辑体系,可作为数学及任何公理化理论如总分学的基础.在此逻辑基础上,Lesniewski 总分学可被称为一种公理化的关于“部分与整体最一般关系”的超逻辑(extralogical)理论.

Lesniewski 总分学的基本原理原始描述为:

(1) 只有 A 中的组成成分但不同于 A 的个体是 A 的部分(特指真部分),没有个体是自己的部分;如果 A 是

B 的部分,则 B 必不是 A 的部分;任何 A 的部分是 A 的部分.这里规定了部分关系的 3 个特性,即反自反、反对称以及传递性.设 PP(proper part)表示二元部分关系.

($P1$) $\neg PPxx$

($P2$) $PPxy\rightarrow\neg PPyx$

($P3$) $PPxy\wedge PPyz\rightarrow PPxz$

(2) 只有 A 中的任何个体是 A 的一个组成元素.任何个体在其自身中,并且任何 A 的组成成分也在 A 中.只有任何 A 的部分或者 A 本身在 A 中.用 INGR 表示二元组成关系(ingredient).

($I1$) $INGRxx$

($I2$) $INGRxy\wedge INGRyz\rightarrow INGRxz$

($I3$) $INGRxy\rightarrow PPxy\vee x=y$

(3) A 覆盖(overlap)B 定义为当且仅当至少个体 A 的一个组成成分在 B 中;即 A 和 B 至少拥有一个共同的组成成分.A 覆盖 B 当且仅当 B 覆盖 A.A 在 B 中当且仅当 A 的每个组成成分在 B 中,这些情况包括:A 的每个部分是 B 的部分,或者 A 的是一个个体,其每个组成成分都覆盖 B,或者每个覆盖 A 的个体也同样覆盖 B.

($O1$) $Oxy\equiv_{def}\exists z(INGRzx\wedge INGRzy)$

($O2$) $Oxy\leftrightarrow Oyx$

($O3$) $INGRxy\leftrightarrow(\forall z(PPzx\wedge PPzy))\vee(\forall z(PPzx\wedge Ozy))\vee(\forall z(Ozx\wedge Ozy))$

(4) 由于只有相互组成的个体是相同的,因此下面 5 个条件中的任何一个都构成了 A 和 B 相同的充分必要条件:

仅有 A 的组成成分在 B 中;

A 的每个部分在 B 中,并且 B 的每个部分在 A 中;

A 为一个个体,其所有组成成分覆盖 B,并且它都被 B 的所有组成成分所覆盖;

每个 A 的组成成分覆盖 B,并且每个 B 的组成成分覆盖 A;

只有 A 的每个覆盖者覆盖 B.

(5) A 是 b 的一个集体(collection)当且仅当 A 中每个组成成分都至少有 1 个组成成分在 A 中的一个 b 中;即只有所有组成成分都覆盖自身中一个 b 的个体是 b 的集体.A 是 b 的总体(totality)当且仅当:1) 每个 b 在 A 中并且 2) 每个 A 的组成成分的一些组成成分在至少 1 个 b 中.

(COL) $xCOLy\equiv_{def}\exists y[(y\in b)\wedge INGRyx\wedge\forall z[INGRzx\wedge\exists w[INGRwz\wedge INGRwy]]]$

(TTL) $xTTLy\equiv_{def}\forall y[((y\in b)\rightarrow INGRyx)\wedge xCOLy]$

Tarski 将 Lesniewski 总分学加以归纳[9],总结出如下定义和公设.在术语上,真部分(proper part)等价于原始定义的部分关系(PP),部分关系(P)等价于原始定义的组成关系(INGR),不相交(disjoint)关系等价于原始定义的覆盖(overlap)关系取非,总和(sum)等同于原始定义的总体(totality),对应的形式化描述由 Bennett 等人给出[40].

总分学定义 1. 个体 x 称为个体 y 的真部分,若 x 是 y 的部分且 x 与 y 不等同.

$$PP(x,y)\equiv_{def}(P(x,y)\wedge\neg(x=y)).$$

总分学定义 2. 个体 x 称为与个体 y 不相交,若不存在个体 z 同时是 x 和 y 的一部分.

$$DR(x,y)\equiv_{def}\neg\exists z[P(z,x)\wedge P(z,y)].$$

总分学定义 3. 个体 x 称为个体类 α 所有元素的总和,若 α 的每个元素是 x 的一部分并且不存在 x 的组成部分与 α 的所有元素不相交.

$$SUM(\alpha,x)\equiv_{def}\forall y[y\in\alpha\rightarrow P(y,x)]\wedge\neg\exists z[P(z,x)\wedge\forall y[y\in\alpha\rightarrow DR(y,z)]].$$

总分学公设 1. 若 x 是 y 的组成部分并且 y 是 z 的组成部分,则 x 是 z 的组成部分.

$$\forall x\forall y\forall z[P(x,y)\wedge P(y,z)\rightarrow P(x,z)].$$

总分学公设 2. 对每个非空个体类α,有且仅有一个个体 x 是α所有元素的总和.

$$\forall\alpha[\exists x[x\in\alpha]\rightarrow\exists!x[SUM(\alpha,x)]].$$

附加定义. 为方便描述,附加定义覆盖关系为不相交关系取非.

$$O(x,y)\equiv_{def}\neg DR(x,y).$$

3.2 结合总分学与粗糙性方法的动机

经典 Cantor 集合论理论上的问题如最大超集悖论等仅仅局限于纯理论,集合论作为经典数学体系的基础地位并未受到影响.Lesniewski 总分学在数学界并未因为其避免了集合论的某些理论问题而得到广泛采纳.尽管如此,Lesniewski 总分学与其底层的逻辑系统一起,为拓扑学中的时空理论(spatiotemporal theories of topology)、Taski 公理化立体几何(Tarski's axiomatic geometry of solids)[9]、科学描述现实的规范语言 L(canonic language L)[8]等提供了数学基础.其中,Taski 立体几何以及近年来出现的 Taski 立体几何扩展,如"区域连接演算(region connection calculus)"了[41]、"基于区域的定性几何(region-based qualitative geometry)"[40]等理论,表明总分学在空间信息处理等领域发挥的作用.总分学关系在应用语境上的多样性很适合在语义 Web 中构造本体(ontology)内部元素之间的关系,例如类型与子类型关系的关系、对象与性质的关系、包/序列与其元素的关系.事实上,在本体计算的一个研究分支中,意大利学者 Guarino 的形式本体论(formal ontology)[42]体系中,总分学和拓扑学构成了其理论的两个基本要素,该体系给出了本体驱动信息系统(ontology driven information system)的思想,为构建基于知识的信息系统提出了一套新的范例.

鉴于总分学的上述实用性价值,引入总分学关系以便今后面向空间信息处理及本体计算环境应用粗糙性方法.但是,本文的最底层动机并非局限于利用总分学的已有应用优势,更重要的是通过建立纯粹总分学概念之上的粗糙性,展示总分学作为与经典集合论相当的体系所具有的能力与特性.以此抛砖引玉,使总分学这一不为人熟知的庞大体系获得必要的重视,从而发展出更多的适用于计算机科学领域的跨学科理论.

作为最早将总分学中部分到整体的关系引入到粗糙性理论的实践,Polkowski 及 Skowron 提出了粗糙总分学(rough mereology)及其上的适应性演算系统(adaptive calculi)[5−7].简言之,粗糙总分学可以看作是对 Lesniewski 总分学的扩展,最核心的思想是:该理论将 Lesniewski 总分学的"部分"关系替换为表示"以一定程度作为一部分"的粗糙包含(rough inclusion)关系.基于总分学基本原理和 Taski 立体几何概念,本文对粗糙包含的内涵进行了如下分析[10]:

对象全集 U 上一个取值在[0,1]之间的实函数$\mu(x,y)$若满足下列条件,则称为粗糙包含:

(1) 任何 x 有$\mu(x,x)=1$.该条件对应于 Lesniewski 总分学中组成关系(INGR)所具有的自反特性,即 Taski 立体几何空间中球域自身的包含关系.这里描述了粗糙包含的自反性.

(2) 若$\mu(x,y)=1$,则对任何三元组(x,y,z)有$\mu(z,y)\geq\mu(z,x)$.该条件表示在 Taski 立体几何空间中,球域 A 是球域 B 的组成部分,则任意球域 C 覆盖球域 B 的程度大于覆盖球域 A 的程度,这里描述了粗糙包含数值的单调性.

(3) 对任意 x,存在 n,使得$\mu(n,x)=1$.对象 n 满足该条件,则称为一个*μ-null* 对象.这是粗糙总分学对空对象存在的假设,而在 Lesniewski 总分学中被排除在系统之外.若$\mu(x,y)=1=\mu(y,x)$,则令 $x=_{\mu}y$.这里引入了粗糙总分学特有的程度等价意义.

(4) 若对象 x,y 有属性:如果 $z\neq_{\mu}n$ 且$\mu(z,x)=1$,则存在使得$\mu(t,z)=1=\mu(t,y)$,那么有$\mu(x,y)=1$.表示在 Lesniewski 总分学中,如果在 B 的任意非空组成成分 A 中存在某个部分是 C 的组成部分,则 B 是 C 的组成成分.即当 B 的任意组成成分 A 都被 C 所覆盖时,B 构成 C 的组成成分.该条件描述了粗糙包含关系上,从"作为子部分"到"作为部分"关系的推理.

(5) 对任意对象集体Γ,存在一个对象 x,具有以下性质:

(A) 如果 $z\neq_{\mu}n$ 且$\mu(z,x)=1$,则存在 $t\neq_{\mu}n,w\in\Gamma$,使得$\mu(t,z)=\mu(t,w)=\mu(w,x)=1$;表示对于 A 的每个组成部分,B 都存在某个组成部分与 x 中的Γ类对象覆盖,对应 Lesniewski 总分学中关于对象集体的定义(COL);

(B) 如果 $w\in\Gamma$,则$\mu(w,x)=1$;该条件表示所有的Γ类对象都是 x 的组成成分;

(C) 如果 y 满足上面两个条件,则$\mu(x,y)=1$.该条件对应于 Lesniewski 总分学中的对象总体定义(TTL).

任何满足 5(A)的 x 称为Γ中的对象集合(set),对应于 Lesniewski 总分学的集体(collection);如果此外还满足5(B)、5(C),则 x 称为Γ中的对象类(class),对应于 Lesniewski 总分学的总体/总和.

在经典粗糙集理论的信息系统中,对象 x 关于特定属性集合 B 属于 X 的程度,在数值上通过粗糙隶属函数(rough membership function) $\mu_{X,B}(x)=\frac{\|X\cap[x]_B\|}{\|[x]_B\|}$ 来表示[3],其中,$[x]_B$ 是 x 在属性集 B 上不可区分关系确定的等价类,$\|\cdot\|$表示集合的势(cardinality).从粗糙集的概念看,粗糙隶属函数取值在[0,1]上,当取值为 1 时,表示 x 在 X 的下界近似区域中;为 0 时,表示 x 位于 X 的上界近似区域之外;而取值(0,1)时,x 位于边界区域中.基于粗糙隶属函数,粗糙总分学定义了标准粗糙包含(standard rough inclusion) $\mu_U(X,Y)=\frac{\|X\cap Y\|}{\|X\|}$,$X$ 为空时,取值为 1.

粗糙包含$\mu(x,y)=r$ 可以解释为"x 以程度 r 作为 y 的部分",按照标准粗糙包含及其参照的粗糙隶属函数的粗糙集意义可以认为,在结合粗糙性方法与总分学方法的着眼点上,粗糙总分学是将在集合论上定义的粗糙性应用到描述总分学关系,用于表达"部分到整体关系的含糊性"或"粗糙的部分到整体关系",从而达到用粗糙性概念扩展 Lesniewski 总分学的目的.而本文的着眼点则是以 Lesniewski 总分学的精确部分到整体关系来描述信息系统表示模型中信息颗粒之间的关系,并通过这种关系来最终定义基于纯粹总分学关系的粗糙性.

4 粒度表示演算(granular representation calculus,简称 GRC)

4.1 原子信息颗粒与复合信息颗粒

粒度表示演算定义 1. 原子信息颗粒(atomic granule)是从信息系统中析取出来的粒度最小且包含完整有意义信息的语义单元,形式化为一个三元组(u,c,v),三元分别对应实体名字标识符 u、属性名 c 以及属性取值 v,陈述了"u 具有属性 c 取值 v"这样一个单一事实,符号表示为ξ或$\xi(u,c,v)$.原子颗粒相等定义为构成原子颗粒三元组的每个元均相同.原子信息颗粒代表着粒度表示演算系统中的独立包含完整概念的最简单个体.除了来源于总分学的概念以外,原子信息颗粒是 GRC 中定义的唯一原语,更复杂的结构可以通过原子信息颗粒构造.例如在 I^*中的原子信息颗粒$\xi(u_1,c_1,0)$,$\xi(u_7,c_3,2)$等.

这里有一个问题需要阐释,如果原子信息颗粒的三元组(u,c,v)是最基本的个体,那么 u 是什么呢?从语言学角度类比,三元组的三元分别对应了陈述句的主语、谓语和宾语.在词汇空间中,主语、谓语和宾语都是个体,但是上升到句子空间,主语、谓语和宾语都不能成为一个完整的、有意义的句子,句子空间个体最简单的即为该陈述句而不是其成分.简言之,u 仅仅是实体名字标识符,是被描述实体的引用,单纯从"u"这个符号上无法获取名字之外更多的信息.在经典粗糙集理论中,由于单纯处理实体名字标识符,对信息表引入的语义上下文存在隐式依赖,从而造成粗糙性方法的误用.而在粒度粗糙理论中,三元组的属性名 c 和属性值 v 组合起来限定了实体名字标识符 u 的语义上下文.

遵循 Lesniewski 总分学的习惯,在系统中不包含空(NULL)信息颗粒.一旦三元组中任意元没有取值,该信息颗粒就不存在.对于特定的场合需要表示没有符合条件的信息颗粒存在.操作无意义时,使用符号∅表示作为结果的占位符,类似于在算术运算中被 0 除得到的"非数 NaN(not a number)",运算没有产生任何新东西.需要强调,∅不具有任何本体论意义上的存在性,它不是一个空信息颗粒,因此它不能进一步参与后续运算.

粒度表示演算定义 2. 原子颗粒的聚合运算(atomic aggregation).该操作用于多个原子颗粒来合成更复杂的信息颗粒,是指参与聚合的颗粒表达的事实同时成立.聚合运算把多个信息颗粒放在一起,符号化为使用括号将多个单独的信息颗粒标记为一个整体,信息颗粒之间以冒号分开.例如,两个原子颗粒ξ_1,ξ_2 聚合的结果形如$(\xi_1:\xi_2)$.聚合运算符是⊙,二元时,$\xi_1\odot\xi_2\equiv_{def}(\xi_1:\xi_2)$;多元时,$\odot(\xi_1,\xi_2,\ldots,\xi_n)\equiv_{def}(\xi_1:\xi_2:\ldots:\xi_n)$.聚合运算的结果称为复合信息颗粒(compound granule),记作Θ.

聚合运算是整个粒度表示演算的最基本运算,是众多信息颗粒各司其职、构成复杂信息系统的基本途径.在上述定义中,并没有预设聚合运算的具体应用语义,没有规定参与聚合运算的原子颗粒必须具有怎样的合法性,也没有规定聚合之后的复合信息颗粒因为聚合而反映了怎样的总分学关系语境.该运算可以看作具有参数化应用语义,具体应用系统设计者若需要利用粒度表示演算来表达对应系统,则必须首先显式规定其聚合运算的应用语义,主要是聚合运算所描述的信息颗粒之间可能的总分学关系语境(前文所述 WCH 分类中的情况).采用此种定义方式,是为了粒度表示演算具有更高的灵活性和可扩展性,以便适应各种不同的信息系统.

粒度表示演算定义 3. 复合颗粒的聚合(aggregation)运算与融合(fusion)运算.聚合运算在复合颗粒上与在原子颗粒上的功能相同,算符也是⊙,都是把独立的信息颗粒绑定为一个整体,构成新的复合颗粒,即$\Theta_1\odot\Theta_2=_{def}(\Theta_1:\Theta_2)$.融合运算(⊕)提取参与运算的信息颗粒所包含的原子颗粒,然后将这些原子颗粒重新聚合成新的复合颗粒,对$\Theta_a=(\xi_{a_1}:\xi_{a_2}:...:\xi_{a_n})$和$\Theta_b=(\xi_{b_1}:\xi_{b_2}:...:\xi_{b_m})$,$\Theta_a\oplus\Theta_b=(\xi_{a_1}:\xi_{a_2}:...:\xi_{a_n}:\xi_{b_1}:\xi_{b_2}:...:\xi_{b_m})$.参照总分学的习惯,参与聚合成一个复合颗粒的结构称为复合颗粒的成分(ingredient),成分与该复合颗粒之间存在着部分到整体的总分学关系 P,这种关系记为$P(\xi,\Theta)$或$P(\Theta_1,\Theta_2)$.

粒度表示演算定义 4. 信息颗粒的自聚合(self aggregation)是指信息颗粒与自身聚合(或相等的信息颗粒进行聚合),如$\xi_1\odot\xi_1=(\xi_1:\xi_1)$,对此有两点说明:(1) 一个信息颗粒与自身聚合并不能添加任何其他信息,因此,聚合运算中任意数量的重复操作数(参与运算的颗粒)仅在结果颗粒中出现 1 次,即$\xi_1\odot\xi_1=(\xi_1)$;(2) 复合信息颗粒(ξ_1)不能表述比ξ_1更多的信息,因此,这是$P(\xi,\Theta)$定义中部分与整体之间的“等同”关系,则$\xi_1=(\xi_1)$.同样的说明对复合颗粒的聚合也成立.需要指出的是,上述说明中的第 2 点与经典集合论存在本质上的差异,在集合论中,一个元素与仅包含该元素的单元素集合不相等.

粒度表示演算定义 5. 平凡复合信息颗粒(trivial compound granule)是指成分仅包含单一原子颗粒的复合信息颗粒,此时,该复合颗粒与其成分在总分学关系上是相等的.基于此,原子信息颗粒因而也可以参与复合信息颗粒的运算.

粒度表示演算定义 6. 非平凡信息颗粒结构分类.从不同类型成分应用不同运算获得的复合颗粒之间存在着差别.目前为止,非平凡复合颗粒仅仅简单地区分为单纯复合信息颗粒(plain compound granule)和高阶复合信息颗粒(higher order compound granule),前者所有成分都是原子颗粒,后者成分中含有复合颗粒.定义这种差别的原因是,高阶复合颗粒中编码了其组成部分的结构信息.以复合颗粒的聚合和融合运算为例,前者产生的结果为高阶复合颗粒,后者为单纯复合颗粒.由于聚合运算的具体应用语义因系统而异,该运算带入的结构信息对整体语义的影响也不同,进而区分两类复合颗粒的必要性和重要性也不尽相同.例如,两个复合信息颗粒分别表示为一张简单 ER 表和一棵 XML 树,后者就具有比前者复杂得多的层次相关结构信息,从总分学语境 WCH 分类来说,前者的信息颗粒之间一般是成员/集体关系,后者则体现出除此之外的构件/集成对象关系.在具有构件/集成对象关系的信息源中,聚合与融合的差异体现出了重要意义.

作为对原子信息颗粒多元聚合运算的补充,可用上述复合颗粒的结构观点阐明原子信息颗粒上的 n 元聚合运算和 n–1 元聚合运算的关系.前 n–1 个原子颗粒通过聚合运算生成一个单纯复合颗粒,对第 n 个原子颗粒,如果直接按照$\odot_n(\xi_1:\xi_2:\ldots:\xi_n)=\odot_{n-1}(\xi_1:\xi_2:\ldots:\xi_{n-1}):\xi_n$进行聚合运算,则产生一个高阶复合颗粒$(\Theta_{n-1}:\xi_n)$,但原子颗粒的聚合结果应该是单纯复合颗粒,故而正确的关系应该是$\odot_n(\xi_1:\xi_2:\ldots:\xi_n)=\odot_{n-1}(\xi_1:\xi_2:\ldots:\xi_{n-1})\oplus\xi_n$.

粒度表示演算定义 7. 自融合(self-fusion).针对高阶复合颗粒的单目运算,去掉其编码的成分结构信息,将原有的复合成分用最基本的原子成分代替,使其转换成单纯复合颗粒.假设$\Theta=(\Theta_a:\Theta_b:\xi_c)$,其中,$\Theta_a=(\xi_{a_1}:\xi_{a_2}:...:\xi_{a_n})$,$\Theta_b=(\xi_{b_1}:\xi_{b_2}:...:\xi_{b_m})$,则$\nabla(\Theta)=(\xi_{a_1}:\xi_{a_2}:...:\xi_{a_n}:\xi_{b_1}:\xi_{b_2}:...:\xi_{b_m}:\xi_c)$,∇为该单目运算符.

粒度表示演算定义 8. 信息颗粒粒度(granularity).当需要研究信息颗粒相关的一些量化指标时,信息颗粒的粒度(granularity)可以从两方面考量,信息颗粒包含基本信息的数量以及信息颗粒中信息的复杂度.信息颗粒Θ的粒度势(granular cardinality)定义为:其成分所包含的全部原子颗粒数量,记为$\aleph(\Theta)$.对任意的原子颗粒ξ的势$\aleph(\xi)=1$,而高阶的复合信息颗粒的势需要首先自融合其复合成分为原子颗粒才能求得.对于无意义的占位符∅,其势可以定义为$\aleph(\varnothing)=0$,但这仅仅表示在不存在任何信息颗粒的情况下,势的取值为 0.信息颗粒Θ的粒度阶

(granular order)的定义为:平凡复合信息颗粒(等价于原子颗粒)的阶为 0,其他非平凡复合信息颗粒的阶等于其成分的最高阶加 1,记为$\Im(\Theta)$.信息颗粒的阶在最初描述的 GRC 版本中没有定义,引入该定义,从信息复杂度方面描述信息颗粒粒度,以适应在本体计算环境中对层次描述的可能需求.

4.2 几种重要的复合信息颗粒

本节将定义几种对粗糙性构造有用的复合信息颗粒.

粒度表示演算定义 9. 集簇信息颗粒(cluster granule).$\odot(\xi_1(u_{i_1},c_j,v_t),\xi_2(u_{i_2},c_j,v_t),\ldots,\xi_n(u_{i_n},c_j,v_t))$表示所有在相关属性 c_j 上取相同值 v_t 的原子信息颗粒$\xi_k(k=1,2,\ldots,n)$(信息表意义上的纵向)的聚合结果,简记为$((u_{i_1},u_{i_2},\ldots,u_{i_n}),c_j,v_t)$,标记为$\Xi(c_j,v_t)$.例如在 I^*中,$\Xi(c_3,1)=((u_2,u_5,u_6,u_8),c_3,1)$,$\Xi(c_5,0)=((u_7),c_5,0)$.根据总分学公设 2,一个特定集簇颗粒可以看作是一类原子颗粒的总体,这些原子颗粒属于个体类"描述'当前实体具有属性 c_j 取值为 v_t'的原子颗粒".通过信息颗粒到信息颗粒所描述实体上的语义映射,上述理解等价于:集簇颗粒是"一个描述所有的具有属性 c_j 取值为 v_t 之实体总体的信息颗粒".

粒度表示演算定义 10. 示象信息颗粒(aspect granule).$\odot(\xi_1(u_i,c_{j_1},v_1),\xi_2(u_i,c_{j_2},v_2),\ldots,\xi_n(u_i,c_{j_n},v_n))$表示同一个实体 u_i 任意数目的原子信息颗粒$\xi_k(k=1,2,\ldots,n)$(信息表意义上的横向)聚合结果,简记为$(u_i,(c_{j_1},c_{j_2},\ldots,c_{j_n}),(v_1,v_2,\ldots,v_n))$,标记为$\Psi(u_i,(c_{j_1},c_{j_2},\ldots,c_{j_n}))$.示象颗粒描述了一个给定实体多方面的属性,例如在 I^*中,$\Psi(u_1,(c_1,c_3,c_4))=(u_1,(c_1,c_3,c_4),(0,0,2))$,表明实体 u_1 具有属性 $c_1=0,c_3=0,c_4=2$.

粒度表示演算定义 11. 示象集簇信息颗粒(aspect cluster granule).

$\odot(\Psi_1(u_{i_1},(c_{j_1},c_{j_2},\ldots,c_{j_m})),\Psi_2(u_{i_2},(c_{j_1},c_{j_2},\ldots,c_{j_m})),\ldots,\Psi_n(u_{i_n},(c_{j_1},c_{j_2},\ldots,c_{j_m})))$表示所有在相关属性集合$(c_{j_1},c_{j_2},\ldots,c_{j_m})$上具有相同赋值的示象颗粒$\Psi_k(k=1,2,\ldots,n)$(信息表意义上的纵向)聚合结果,简记为$((u_{i_1},u_{i_2},\ldots,u_{i_n}),(c_{j_1},c_{j_2},\ldots,c_{j_m}),(v_1,v_2,\ldots,v_m))$,标记为$\Gamma((c_{j_1},c_{j_2},\ldots,c_{j_m}),(v_1,v_2,\ldots,v_m))$.分别称示象集簇颗粒中 $E=(u_{i_1},u_{i_2},\ldots,u_{i_n})$,$A=(c_{j_1},c_{j_2},\ldots,c_{j_m})$,$V=(v_1,v_2,\ldots,v_m)$为实体段(entity segment)、属性段(attribute segment)和取值段(value segment).这是一种简写方式,一个示象集簇颗粒代表了描述"具有属性段 A,其对应取值段为 V 的实体"的所有示象信息颗粒总体.例如在 I^*中,$\Gamma((c_1,c_2),(0,1))=((u_1,u_3,u_8),(c_1,c_2),(0,1))$,其中,实体段为$(u_1,u_3,u_8)$,属性段为$(c_1,c_2)$,取值段为(0,1).若一个示象集簇信息颗粒的实体段仅有唯一的实体,该示象集簇颗粒退化成普通的示象颗粒;若其属性段仅有唯一属性,则退化为普通的集簇颗粒.这两种情况下,允许采用任意可用的标记方式.

4.3 特殊复合信息颗粒相关的更多运算

对于特殊复合信息颗粒,定义运算参与者或结果限定在上述特殊颗粒范围内的额外运算.

粒度表示演算定义 12. 归并(merge)运算.该运算在集簇信息颗粒及示象集簇信息颗粒之上进行操作,用于产生新示象集簇信息颗粒,可能的形式有$\Xi(c_j,v_j)\otimes\Xi(c_k,v_k)\equiv_{def}\Gamma((c_j,c_k),(v_j,v_k))$,$\Xi(c_j,v_j)\otimes\Gamma(A_k,V_k)\equiv_{def}\Gamma(A_l,V_l)$以及$\Gamma(A_j,V_j)\otimes\Gamma(A_k,V_k)\equiv_{def}\Gamma(A_l,V_l)$,其中,$\otimes$为归并运算符.归并运算实际上是简单运算复合而成,分 3 步简单运算来完成,以$\Xi(c_j,v_j)\otimes\Xi(c_k,v_k)$为例:

1) 两个集簇颗粒首先进行融合运算,获得一个新的单纯复合颗粒,$\Xi(c_j,v_j)\oplus\Xi(c_k,v_k)=\Theta$.
2) 对Θ中所有的原子颗粒按照实体进行聚合,产生每个实体对应最大可能粒度势($\aleph(\Theta)$)的示象颗粒.
3) 保留所有含有属性段(c_j,c_k)和取值段(v_j,v_k)的示象颗粒作为结果信息颗粒的成分.在某些情况下,归并操作不能获得结果,此时使用占位符$\varnothing$.

在 I^*中,要归并$\Xi(c_3,1)=((u_2,u_5,u_6,u_8),c_3,1)$和$\Gamma((c_1,c_2),(0,1))=((u_1,u_3,u_8),(c_1,c_2),(0,1))$,相应的步骤如下:

1) 融合运算

$$\Xi(c_3,1)\oplus\Gamma((c_1,c_2),(0,1))=((u_2,c_3,1):(u_5,c_3,1):(u_6,c_3,1):(u_8,c_3,1):(u_1,c_1,0):(u_3,c_1,0):(u_8,c_1,0):(u_1,c_2,1):(u_3,c_2,1):(u_8,c_2,1)).$$

2) 合并成分中所有最大粒度势示象颗粒

$$((u_2,c_3,1):(u_5,c_3,1):(u_6,c_3,1):(u_1,(c_1,c_2),(0,1)):(u_3,(c_1,c_2),(0,1)):(u_8,(c_1,c_2,c_3),(0,1,1))).$$

3) 删除不包含属性段(c_1,c_2,c_3)和取值段(0,1,1)的成分

$$((u_8,(c_1,c_2,c_3),(0,1,1)))=(\Psi(u_8,(c_1,c_2,c_3)))=\Psi(u_8,(c_1,c_2,c_3)).$$

这是一个退化的示象集簇颗粒,即普通示象颗粒.

为了便于描述一些在复合颗粒成分上进行的常见操作,除了前面定义的自融合运算以外,这里定义 3 个与特殊复合颗粒有关的单目内部运算(unary internal operation):横向汇聚(horizontal convergence)、纵向汇聚(vertical convergence)和完全汇聚(full convergence),实现单纯复合颗粒到高阶复合颗粒之间的转换.

粒度表示演算定义 13. 横向汇聚(horizontal convergence).用于单纯复合信息颗粒,将其中的原子颗粒成分聚合成针对每个实体具有最大粒度势的示象颗粒作为结果颗粒的成分.与自融合相对应,该操作将单纯颗粒转换成高阶颗粒.从原始信息表的角度考虑,相当于把复合颗粒中所有同一行的单元格组合起来.

粒度表示演算定义 14. 纵向汇聚(vertical convergence).用于单纯复合信息颗粒,将其中的原子颗粒成分聚合成针对每个属性具有最大粒度势的集簇颗粒作为结果颗粒的成分,该操作也将一个单纯信息颗粒转换成高阶颗粒.从传统粗糙集理论中原始信息表的角度考虑,相当于把复合颗粒中所有同一列的单元格组合成多个等价类.

粒度表示演算定义 15. 完全汇聚(full convergence).用于单纯复合信息颗粒,将其中的原子颗粒成分聚合成具有最大粒度势的示象集簇颗粒作为结果颗粒的成分,该操作也将一个单纯信息颗粒转换成高阶颗粒.这是上述两种方向的汇聚综合.

粒度表示演算定义 16. 示象转换(aspect shift)运算.

给定一个示象集簇颗粒(E,A_i,V_i),其对应的实体段 E,从 A_i 属性段的角度看,E 构成了具有当前 A_i 取值 V_i 的实体(总分学意义上的)总体,但是,当换一个角度 A_j 来观察这些实体时,原本同一类实体会按照属性段 A_j 的取值发生怎样的分化呢?该运算表示当前示象集簇颗粒随着相关属性段变化,实体段重新与新属性段组合,按照新属性段不同的取值类型,生成一个或多个新的示象集簇颗粒作为结果颗粒成分,运算符为 δ_{A_j} .其执行过程为,首先,聚合 E 中所有实体对应于 A_j 中所有属性的原子颗粒,然后,再对该复合颗粒进行完全汇聚,生成所有针对 A_j 取值的示象集簇颗粒作为结果颗粒的成分.

从传统粗糙集理论角度理解,示象转换运算是给出了在属性集合 A_i 上具有不可区分关系的一个等价类,按照 A_j 上的不可区分关系来重新划分为一个或多个等价类.该运算定义了从描述实体一个方面特征映射到描述另一方面特征的映射方法.在 I^*中,给定示象集簇颗粒$\Gamma((c_1,c_2),(0,1))=((u_1,u_3,u_8),(c_1,c_2),(0,1))$,一个最常见的需求是求实体段$(u_1,u_3,u_8)$从当前属性段$(c_1,c_2)$转换视角到决策属性 d 时如何分化,通过执行示象转换运算,有$\delta_d(\Gamma((c_1,c_2),(0,1)))=(((u_1,u_3),d,d_1):(u_8,d,d_3))$.

4.4 信息颗粒之间的总分关系判定

基于原子信息颗粒的原子性,所有的原子颗粒之间都是不相交的(disjoint),即$\forall\xi_i,\xi_j(i\neq j),DR(\xi_i,\xi_j)$.当两个复合颗粒具有公共的成分时,两者之间关系是覆盖关系,$O(\Theta_a,\Theta_b)$.

粒度表示演算定义 17. 重叠(wrap)运算.该运算用于获取两个复合信息颗粒公共组成部分,运算符记为◊,当$\nabla(\Theta_a)=(\xi_{a_1}:\xi_{a_2}:...:\xi_{a_n})$ 和$\nabla(\Theta_b)=(\xi_{b_1}:\xi_{b_2}:...:\xi_{b_m})$ 时,$\Theta_a\Diamond\Theta_b\equiv_{def}(\xi_{c_1}:\xi_{c_2}:...:\xi_{c_l})$ 其中,$\xi_{c_1}:\xi_{c_2}:...:\xi_{c_l}$ 是Θ_a 和Θ_b 共有的成分.重叠运算首先需要对参与运算的复合颗粒进行自融合运算,获得的结果没有保留内部复合信息颗粒的结构,是一个单纯复合颗粒.参与运算的复合颗粒很可能不相交,$DR(\Theta_a,\Theta_b)$,由于已经排除了空信息颗粒的存在,此时覆盖运算结果无意义,用占位符∅代替.

重叠运算对于判定两个信息颗粒之间的总分学关系至关重要,若结果存在,则覆盖关系成立;反之,若结果无意义,则信息颗粒不相交.当覆盖关系成立时,两个信息颗粒之间是否存在部分到整体的关系,以及两个复合信息颗粒是否相等还需要进一步判断.按照总分学的基本原理,两个信息颗粒Θ_a 和Θ_b 相等,等价于$P(\Theta_a,\Theta_b)\wedge P(\Theta_b,\Theta_a)$,即两者互为组成部分,因此,最关键的是如何确定信息颗粒之间的部分到整体关系 P.

如前所述,直接参与聚合运算的信息颗粒构成的结果信息颗粒的成分,即部分到整体关系成立,而按照部分到整体关系的传递性(总分学公设 1),这些信息颗粒的成分也构成了结果颗粒的成分.例如,若$\Theta_r=\Theta_1\odot\xi_0$,其中,$\Theta_1=\xi_1\odot\xi_2$,则 $P(\Theta_1,\Theta_r),P(\xi_0,\Theta_r),P(\xi_1,\Theta_r)$和 $P(\xi_2,\Theta_r)$均成立.现在的问题是,若$\Theta_2=\xi_0\odot\xi_1$,那么 $P(\Theta_2,\Theta_r)$是否成

立?这个问题再次涉及到聚合运算带来的结构信息.前面提到,聚合运算具有参数化的语义,从不同的系统中带来的结构信息的重要性也不同.为避免出现二义性,同时又不限制粒度表示演算可能面对的具有复杂结构信息的表示语义,特将复合信息颗粒之间的部分到整体关系区分为规范总分关系(canonical part-to-whole relation)和广义总分关系(generalized part-to-whole relation).规范部分关系要求作为部分的信息颗粒是整体本身、参与聚合颗粒或其成分,其结构必须与整体内部结构在具体总分关系语境上相一致.在规范总分关系下,$P(\Theta_2,\Theta_r)$一般不成立.广义部分关系仅要求作为部分的信息颗粒的原子成分全部是整体颗粒的成分,不关心成分内部聚合结构带来的差异,此时,$P(\Theta_2,\Theta_r)$成立.这两种总分关系的差异源于聚合运算在具体系统中总分学语境的不同,在需要比较规范总分学关系的系统中,必须提供相应的比较机制.由于该机制在不同的信息源上实现具有很强的特殊性,具体实现方法是未来研究粒度粗糙理论扩展到特定信息源时需要特别解决的重点和难点.在总分学语境较单纯的单表信息系统中,判断广义总分关系即可满足构建粒度粗糙性的要求.对于任意给定的两个复合信息颗粒Θ_a和Θ_b,要判断两者之间的广义总分关系,可按如下步骤进行:

1) 计算$\Theta_r=\Theta_a\Diamond\Theta_b$;

2) 计算原子成分数量:$\aleph(\Theta_a),\aleph(\Theta_b),\aleph(\Theta_r)$;

3) 比较 3 个粒度势,若结果粒度势$\aleph(\Theta_r)$等于操作数粒度势中较小的一个,则广义总分关系成立.

5 粒度粗糙性的构造

5.1 粒度粗糙性构造方法

粗糙性的表示语义,其粒度表示演算解释为使用条件属性对应的若干示象集簇信息颗粒来近似一个决策属性确定的示象集簇信息颗粒.本节以下内容将集中讨论在粒度表示演算表达的信息系统之上构造粗糙性的问题,为陈述方便,属性段是条件属性的信息颗粒称为“条件信息颗粒”,对应决策属性的则称为“决策信息颗粒”.

对条件属性段 B 中所有属性的全部赋值组合,分别聚合产生信息系统中的条件示象集簇信息颗粒.对决策属性 d,也按照 d 的不同赋值,聚合成多个决策集簇颗粒.令$\Gamma_k(B,V_k)$代表该条件颗粒系列中 B 赋值为 V_k 的示象集簇颗粒,令$\Xi(d,d_i)$代表 d 赋值为 d_i 的决策集簇颗粒.通过示象转换运算$\delta_d(\Gamma_k(B,V_k))$,获得$\Theta_k=(\Xi_1(d,v_1):\Xi_1(d,v_1):\ldots:\Xi_t(d,v_t))$,其中,$v_i(i=1,2,\ldots,t)$为决策属性 d 的可能取值.Θ_k 表示了$\Gamma_k(B,V_k)$实体段在决策属性上体现出来的取值多样性.对于不强调聚合运算结构信息的系统,使用重叠运算$\Theta_k\Diamond\Xi(d,d_i)$判断二者间的广义总分关系;对于需要强调结构信息的系统,应根据该系统聚合运算反映的总分学语境的具体情况判定二者间的规范总分学关系.在此基础上,可将当前示象集簇颗粒$\Gamma_k(B,V_k)$分为 3 类:

粒度粗糙性定义 1. 若Θ_k是$\Xi(d,d_i)$的部分,即有 $P(\Theta_k,\Xi(d,d_i))$,则$\Gamma_k(B,V_k)$称为关于 d_i 的正则 B-颗粒(regular B-granule with respect to d_i),记为$\mathfrak{R}_B(d_i)$.

粒度粗糙性定义 2. 若Θ_k覆盖$\Xi(d,d_i)$但不是它的部分,即有 $O(\Theta_k,\Xi(d,d_i))\wedge\neg P(\Theta_k,\Xi(d,d_i))$,则$\Gamma_k(B,V_k)$称为关于 d_i 的非正则 B-颗粒(irregular B-granule with respect to d_i),记为 $\hat{\mathfrak{R}}_B(d_i)$.

粒度粗糙性定义 3. 若Θ_k与$\Xi(d,d_i)$不相交,即有 $DR(\Theta_k,\Xi(d,d_i))$,则$\Gamma_k(B,V_k)$称为关于 d_i 不相干(irrelevant with respect to d_i).

基于上述分类,仿照粗糙集理论的集合上界近似与下界近似,可以定义信息颗粒的近似概念如下:

粒度粗糙性定义 4. 所有关于 d_i 的正则 B-颗粒聚合成的复合信息颗粒$\mathfrak{R}_B(d_i)$称为关于$\Xi(d,d_i)$的内核信息颗粒(kernel granule with respect to $\Xi(d,d_i)$),记为$\Lambda_B(d_i)$,该信息颗粒是决策信息颗粒$\Xi(d,d_i)$的下界近似.

粒度粗糙性定义 5. 所有关于 d_i 的非正则 B-颗粒聚合成的复合信息颗粒 $\hat{\mathfrak{R}}_B(d_i)$ 称为关于$\Xi(d,d_i)$的外壳信息颗粒(hull granule with respect to $\Xi(d,d_i)$),记为$\Delta_B(d_i)$,代表着粗糙性概念表达的 Frege 所阐述的含糊性中的边界区域.

粒度粗糙性定义 6. 关于$\Xi(d,d_i)$的内核信息颗粒与外壳信息颗粒聚合成$\Xi(d,d_i)$的上界近似,称为关于$\Xi(d,d_i)$的主体信息颗粒(corpus granule with respect to $\Xi(d,d_i)$),记为$\Omega_B(d_i)$.

以 I^*为例,令 $B=(c_1,c_2),d_i=d_1$.

1) 通过聚合运算,求得当前决策颗粒:$\Xi(d,d_1)=((u_1,u_3,u_5,u_6,u_7),d,d_1)$.

2) 通过聚合运算,求得与(c_1,c_2)相关的示象集簇颗粒:

$\Gamma_1(B,(0,1))=((u_1,u_3,u_8),B,(0,1))$;

$\Gamma_2(B,(1,0))=((u_2,u_7),B,(1,0))$;

$\Gamma_3(B,(2,0))=((u_4),B,(2,0))$;

$\Gamma_4(B,(2,1))=((u_5,u_6),B,(2,1))$.

3) 对 2)中每个示象集簇颗粒作从条件属性段 B 到决策属性 d 示象转换运算,结果决策颗粒:

$\Theta_1=\delta_d(\Gamma_1)=(((u_1,u_3),d,d_1):(u_8,d,d_3))$;

$\Theta_2=\delta_d(\Gamma_2)=((u_2,d,d_2):(u_7,d,d_1))$;

$\Theta_3=\delta_d(\Gamma_3)=(u_4,d,d_2)$;

$\Theta_4=\delta_d(\Gamma_4)=((u_5,u_6),d,d_1)$.

4) 利用重叠运算比较 3)的结果与 1)中给出的决策颗粒$\Xi(d,d_1)$的广义总分学关系,有:

$O(\Theta_1,\Xi(d,d_1))\wedge\neg P(\Theta_1,\Xi(d,d_1))$;

$O(\Theta_2,\Xi(d,d_1))\wedge\neg P(\Theta_2,\Xi(d,d_1))$;

$DR(\Theta_3,\Xi(d,d_1))$;

$P(\Theta_4,\Xi(d,d_1))$.

5) 依据 4)的结果对条件颗粒分类:

关于 d_1 的正则 B-颗粒有:$\Re_B(d_1)=\Gamma_4(B,(2,1))$;

关于 d_1 的非正则 B-颗粒有:$\hat{\Re}_{B,1}(d_1)=\Gamma_1(B,(0,1))$和 $\hat{\Re}_{B,2}(d_1)=\Gamma_2(B,(1,0))$;

关于 d_1 不相干的颗粒有:$\Gamma_3(B,(2,0))$.

6) 构造关于$\Xi(d,d_1)$的信息颗粒近似:

内核信息颗粒:$\Lambda_B(d_i)=(\Re_B(d_1))=\Gamma_4(B,(2,1))$;

外壳信息颗粒:$\Delta_B(d_1)=(\hat{\Re}_{B,1}(d_1):\hat{\Re}_{B,2}(d_1))=(\Gamma_1(B,(0,1)):\Gamma_2(B,(1,0)))$;

主体信息颗粒:$\Omega_B(d_1)=(\Lambda_B(d_1):\Delta_B(d_1))$.

6 粒度粗糙理论的“实体-属性-值”原型实现

经典粗糙集理论应用的一个重要方面是临床医疗信息系统中的知识发现与数据挖掘[43,44],而半结构化 EAV 模型是表达临床医疗系统异质数据的主流.在应用经典粗糙集方法分析 EAV 系统时,必须先变换成规整的结构化信息表.粒度表示演算 GRC 为了表达半结构化数据所代表的信息源,其设计思想遵循了几种主流半结构化数据模型,如 EAV,OEM,RDF 以及 SDOM 的共性,即采纳以“属性-值”为范本的元组模型.采用基于 GRC 的粒度粗糙理论进行临床医疗系统的粗糙性数据分析,显得更加自然和直接.同时,EAV 模型的开放源码实现,如 TrialDB,为粒度粗糙理论的快速原型化提供了基础.

要在相应的半结构化信息系统上实现粒度粗糙理论,关键有两方面的工作,包括从 GRC 信息颗粒到目标系统的信息表示单元的映射,以及从 GRC 信息颗粒上的运算到目标系统上操作的映射.GRC 最基本的原语是原子信息颗粒,即一个表示“实体 u 具有属性 c 取值 v”的三元组,对应了决策表的一个单元格,作为独立的个体参与后续的 GRC 定义运算.这正是 EAV 模型最基本的思想,将二维的 ER 表分解成离散的独立单元格,每个单元格的值连同实体标识符和属性名,映射为 EAV 表中的一行.

Anhøj 简单总结了 EAV 模型在临床数据库中的演化过程[45],最简单的 EAV 实现方式如图 2 所示.在简单 EAV 模型中,传统 ER 结构表 Patient 保存了病人的基本信息,如病人的姓名、性别以及分配的病人标识符;在属性表 Attribute 中定义了对属性定义的元数据,如属性的标识符、属性的名字、属性对应的数据类型等.体现 EAV 思想的关键在于数据表 Data,该表在关系型数据库的意义上是一个多对多的关系表,体现了病人表中病人与属

性表中属性之间的任意组合获得的数据.需要注意的是,在数据表中考虑了病历数据的时序性,即每个病人可能在不同的时间作多次检查,这就使得“病人 ID+属性 ID”不是唯一确定的.为此,特别引入了一个日期时间类型的字段 date 用于表示一个病人在特定日期检查获得的数据.

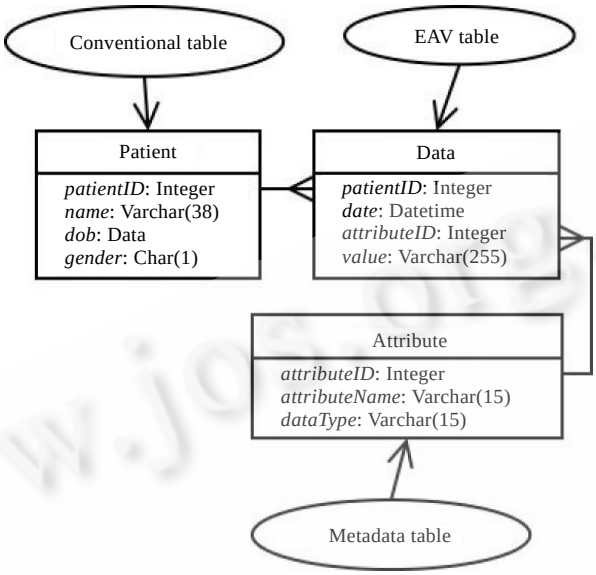

Fig.2 Database schema for simple EAV model[45]

图 2 简单 EAV 模型对应的数据库模式[45]

粒度表示演算 GRC 需要描述图 2 中的 Data 表,首先面临的问题是,如何用 GRC 的原子信息颗粒三元组(*u*,*c*,*v*)表示具有 4 个字段(*patientID*,*date*,*attributeID*,*value*)的 Data 表记录.这需要考虑临床医学数据挖掘对关联规则的具体需求才能确定.当关联规则需要从病人在各种生理检查项目中的测试数据,归纳出这些具体测试数据将如何确定病人罹患的疾病类型,此时,不必区分不同病理表现是否出现在同一患者的不同阶段,可以将数据表中的 patientID 与 date 组合成一个唯一的病人标识符构成 *u*.在目前的粒度粗糙理论的 EAV 实现原型中,采用的是这种简单的映射方式.另外一种映射方法可参考在文献[14]中针对多个智能主体视角,从原子信息颗粒构成上,对粒度表示演算的扩展方式,将原子信息颗粒的三元组扩展成包含一个时序维度 *t* 的四元组(*u*,*t*,*c*,*v*).需要指出的是,上述第 2 种维度扩展的方法需要涉及到较大的底层 GRC 扩展,这种扩展虽然增加了表示系统的复杂度,但却可以引入许多有趣的数据分析场景,例如,对疾病发展过程中针对同一病人呈现的时序特征进行分析等.这种对 GRC 信息颗粒的维度扩展问题,是粒度粗糙理论未来研究的一个重要方面,这样可以进一步扩展粗糙性分析方法适用的范围.

在确定了原子信息颗粒在 EAV 实现中的映射方式之后,可将粒度粗糙理论所涉及到的粒度表示演算中的几种运算翻译成 EAV 数据库操作,从而在 EAV 上实现粗糙性抽取工作.一个完整的粒度粗糙理论实现,需要实现粒度表示演算中定义的所有操作和概念表示,以便适应各种复杂的信息系统表示要求,但作为原型系统,可以仅考虑在粒度粗糙性构造所涉及到的几种运算.参见前面的实例,这些必须实现的运算包括:聚合运算、自融合运算、内部全汇聚运算、示象转换运算、重叠运算.如前所述,聚合运算的实际操作语义因特定聚合要求而有所变化,针对 3 种特殊复合颗粒的聚合运算,可以映射成数据库上的查询操作,分别查询指定实体在某些属性上的事实(示象颗粒)、在某个属性取特定值的所有实体(集簇颗粒)以及在某些属性集上取特定值组合的全部实体(示象集簇颗粒).从 SQL 语句的构造上看,示象聚合给定了 WHERE 子句的实体条件并限定了结果中需要返回的属性类型;集簇聚合给定了 WHERE 子句的单一属性取值条件,返回所有符合条件的 EAV 行;示象集簇聚合则给定了 WHERE 子句的多个属性的取值组合条件,返回所有符合条件的 EAV 行.全汇聚运算等价于在 SQL 语句

中加入了 GROUP BY 子句,限定对结果按照指定的属性段取值组合进行分组,而自融合运算对应于取消分组信息.示象转换运算对应了一个复杂的查询操作,将初始示象集簇颗粒对应的实体 ID 集作为后续查询的实体条件,对其中每一个实体,查询其在目标属性段上的取值.对于重叠运算,只需要进行两个示象集簇颗粒对应临时表的内联操作(INNER JOIN)即可.上述操作描述采用的都是标准关系型数据库术语,因为以 TrialDB 为代表的 EAV 系统构建在标准的商用关系型数据库系统之上,上述运算只需转换成具体 EAV 系统中的查询,可参考 TrialDB 中即席查询(ad hoc query)在标准关系型数据库管理系统上的实现[24].

在增强型的 EAV 模型及面向对象的 EAV 建模 EAV/CR 中,都在数据库模式中加入了较为复杂的元数据管理表,包括属性的类型、分组定义等,这给内建域本体(domain ontology)信息和应用本体(application ontology)信息提供了便利,进而可以为实现本体驱动信息系统提供支撑.对于这些更为复杂的 EAV 模型实现,均可采用上述方法将粒度表示演算的信息颗粒、信息颗粒运算及粗糙性抽取映射到其中.尽管 TrialDB 的源代码对于实现粒度粗糙理论的 EAV 原型是有效的,由于 TrialDB 基于的 WebEAV 以及 SQLGEN 目前的实现封装性较差、跨平台性低、大量即席查询与可视化的界面相绑定等弱点,更完善的原型系统需要移植上述功能到更通用的语言上,或尝试采用主流商用 CSDMS 作为实现基础.

特别说明,标准的关系型数据的库各种操作都是严格基于传统集合论的,在关系型数据库管理系统之上实现的 EAV 数据库也是基于集合论的.粒度粗糙理论试图构建一种基于纯粹总分学的粗糙性表示系统,EAV 数据库上的信息颗粒及运算映射只是粒度粗糙理论并非唯一的一种实现方式,可以看作是把抽象的总分学问题映射到底层集合论实现上,以便对理论进行快速原型化实现,从而通过实际系统来检验理论存在的问题,探索对理论的扩展.由于在设计 GRC 的时候充分考虑了主流的半结构化数据表示在设计思想上的共性,未来的工作中,可以探讨以 GRC 为桥梁,跨具体表示模型的实现异构信息源的一体化粗糙性数据分析机制.

7 结束语

本文从动机、理论和实现 3 个方面系统地整理并完善了粒度粗糙理论的体系,反映了粒度粗糙理论研究的进展.粒度粗糙理论的研究是探索跨领域方法学的一种尝试.归纳本文的思想,粒度粗糙理论及其实践意义包括:

1) 从总分学发展的观点看,由于基于纯粹总分学关系构造粗糙性,可以自然地把粗糙性方法学应用到总分学推动的领域,包括空间信息处理、本体驱动的计算环境等.更重要的是,展示总分学这一体系与集合论相当的描述能力,以期使该理论在计算机科学领域获得更多重视,从而衍生出新的跨学科方法论.

2) 从粗糙性的表示语义看,利用条件颗粒描述的实体事实,来带边界近似决策颗粒描述的实体事实,将底层元数据信息包含在粗糙性构造过程中,消除了隐式数据库模式带来的表示语义混乱,进而避免了粗糙性方法的滥用与误用,避免了为表示而表示的平凡概念近似.

3) 从半结构化表示模型看,扩展的表示模型具有更广泛的信息源描述能力,有效地建立起粗糙性方法学与各种现实信息源的联系.例如,通过 GRC 到 EAV 的映射,粒度粗糙理论获得了在临床医疗研究数据管理系统中主流的数据模型支持,并且现有的 EAV 开源系统又为建立粒度粗糙理论提供了便利.这对进一步研究粒度粗糙理论提供了实验基础,同时又为粗糙性方法学在临床医疗系统的应用提供了桥梁.

还需看到,粒度粗糙理论仍然处于萌芽状态,还有许多经典粗糙集理论中的理论问题必须严谨、缜密地对待,如非常重要的属性约简与求核的问题等[46,47].而为了不断完善粒度粗糙理论,可能的研究方向包括:

1) 关于原子信息颗粒从三元组扩展到更高维度的问题.例如,加入智能主体维度,表达不同观察者多重主观视图的协调问题[14];在 EAV 原型设计中提出加入时间维度对应同一实体不同时间属性信息的时序问题.

2) 利用粒度表示演算中高阶复合颗粒,表达取值段中包含信息颗粒(多表之间通过外关键字体现的构件/集成对象语境)的多表信息系统问题[15,22].通过元组语义联系,完善 EAV 上的粒度粗糙理论实现原型,并建立完整的 RDF,OEM 或其他半结构化表示模型与粒度表示演算的映射问题.其中,特定系统需要设计有效的规范总分学关系判定机制,可能会出现一些不同数据结构上的算法问题.

3) 信息颗粒的空间表示与可视化问题.

4) 本体驱动 Web 信息系统实现问题[16].

致谢 特此感谢本文评审专家给出的详细而中肯的修改意见以及编辑老师们的辛勤劳动.

References:

[1] Pawlak Z. Rough sets. Int'l Journal of Computer and Information Sciences, 1982,11(5):341−356.

[2] Pawlak Z. A treatise on rough sets. In: Peters JF, Skowron A, eds. Proc. of the Trans. on Rough Sets IV. LNCS 3700, Berlin, Heiderberg: Springer-Verlag, 2005. 1−17.

[3] Komorowski J, Pawlak Z, Polkowski L, Skowron A. Rough sets: A tutorial. In: Pal SK, Skowron A, eds. Proc. of the Rough Fuzzy Hybridization: A New Trend in Decision-Making. Singapore: Springer-Verlag Singapore Ltd., 1999. 3−98.

[4] Liu Q. Rough sets and Rough Reasoning. 2nd ed., Beijing: Science Press, 2003 (in Chinese).

[5] Polkowski L, Skowron A. Rough mereology. In: Ras ZW, Zemankova M, eds. Proc. of the 8th Int'l Symp. on Methodologies for Intelligent Systems. LNCS 869, Berlin, Heiderberg: Springer-Verlag, 1994. 85−94.

[6] Polkowski L, Skowron A. Rough mereology: A new paradigm for approximate reasoning. Int'l Journal of Approximate Reasoning, 1996,15(4):333−365.

[7] Skowron A, Polkowski L. Rough mereological foundations for design, analysis, synthesis, and control in distributed systems. Information Sciences, 1998,104(1-2):129−156.

[8] Luschei EC. The Logical Systems of Lesniewski. Amsterdam: North-Holland Publishing Company, 1962.

[9] Tarski A. Logic, Semantics, Meta-mathematics: Papers from 1923 to 1938. Translated by Woodger JH. Oxford: Clarendon Press, 1956. 24−29.

[10] Chen B, Zhou MT. A naïve exploration on intensions of rough mereology. Computer Science, 2002,29(9):7−10 (in Chinese with English abstract).

[11] Chen B, Zhou MT. Re-Examine semantics of rough theory. Computer Science, 2002,29(9):106−109.

[12] Chen B, Zhou MT. A lesniewski mereological analysis on roughness theory. Computer Science, 2006,33(7):171−175 (in Chinese with English abstract).

[13] Chen B, Zhou MT. A pure mereological approach to roughness. In: Wang GY, Liu Q, Yao YY, Skowron A, eds. Proc. of the 9th Int'l Conf. on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC 2003). LNCS 2639, Berlin, Heiderberg: Springer-Verlag, 2003. 425−429.

[14] Chen B, Zhou MT. Adapting granular rough theory to multi-agent context. In: Wang GY, Liu Q, Yao YY, Skowron A, eds. Proc. of the 9th Int'l Conf. on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC 2003). LNCS 2639, Berlin, Heiderberg: Springer-Verlag, 2003. 701−705.

[15] Chen B, Zhou MT. Extending granular representation calculus for internet media resources. In: Li JP, Daugman J, Wickerhauser V, Torresani B, Yen J, Zhong N, Pal SK, Tang YY, Li J, eds. Wavelet Analysis and Its Applications, and Active Media Technology, Proc. of the Int'l Computer Congress 2004. Vol.2. Singapore: World Scientific, 2004. 571−576.

[16] Chen B, Zhou MT. ODWIS as a prototype of knowledge service layer in semantic grid. In: Liew KM, Shen H, See S, Cai W, Fan P, Horiguchi S, eds. Proc. of the 5th Int'l Conf. on Parallel and Distributed Computing, Applications and Technologies (PDCAT 2004). LNCS 3320, Berlin, Heiderberg: Springer-Verlag, 2004. 772−776.

[17] Radzikowska AM, Kerre EE. A comparative study of fuzzy rough sets. Fuzzy Sets and Systems, 2002,126(2):137−155.

[18] Chakrabarty K, Biswas R, Nanda S. Fuzziness in rough sets. Fuzzy Sets and Systems, 2000,110(2):247−251.

[19] Banerjee M, Pal SK. Roughness of a fuzzy set. Information.Sciences, 1996,93(1):235−246.

[20] Codd EF. A relational model of data for large shared data banks. Communications of the ACM, 1970,13(6):377−387.

[21] Chen PP. The entity-relationship model—Toward a unified view of data. ACM Trans. on Database System, 1976,1(1):9−36.

[22] Milton RS, Maheswari VU, Siromoney A. Studies on rough Sets in multiple tables. In: Slezak D, Wang GY, Szczuka MS, Düntsch I, Yao YY, eds. Proc. of the 10th Int'l Conf. on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC 2005), Part I. LNCS 3641, Berlin, Heiderberg: Springer-Verlag, 2005. 265−274.

[23] Nadkarni PM, Brandt CA, Frawley S, Sayward FG, Einbinder R, Zelterman D, Schacter L, Miller PL. Managing attribute-value clinical trials data using the ACT/DB client-server database system. Journal of the American Medical Informatics Association, 1998,5(2):139−151.

[24] Nadkarni PM, Brandt CA. Data extraction and ad hoc query of an entity-attribute-value database. Journal of the American Medical Informatics Association, 1998,5(6):511−527.

[25] Marenco L, Tosches N, Crasto C, Shepherd G, Miller PL, Nadkarni PM. Achieving evolvable Web-database bioscience applications using the EAV/CR framework: Recent advances. Journal of the American Medical Informatics Association, 2003,10(5):444−453.

[26] Brandt CA, Deshpande AM, Lu C, Ananth G, Sun K, Gadagkar R, Morse R, Rodriguez C, Miller PL, Nadkarni PM. TrialDB: A Web-based clinical study data management system, AMIA 2003 open source expo. In: Musen M, ed. Proc. of the 2003 AMIA (American Medical Informatics Association) Annual Symp. Washington: AMIA Press, 2003. 794.

[27] Wang SA, Fann Y, Cheung H, Pecjak F, Upender B, Frazin A, Lingam R, Chintala S, Wang G, Kellogg M, Martino RL, Johnson CA. Performance of using oracle XMLDB in the evaluation of CDISC ODM for a clinical study informatics system. In: Long R, Antani S, Lee DJ, Nutter B, Zhang M, eds. Proc. of the 17th IEEE Symp. on Computer-Based Medical Systems. Washington: IEEE Computer Society, 2004. 594−599.

[28] Cohen W. A Web-based information system that reasons with the structured collections of text. In: Sycara KP, Wooldridge M, eds. Proc. of the 2nd Int'l ACM Conf. on Autonomous Agents. New York: ACM Press, 1998. 400−407.

[29] Papakonstantinou Y, Garcia-Molina H, Widom J. Object exchange across heterogeneous information sources. In: Yu PS, Chen ALP, eds. Proc. of the 11th Int'l Conf. on Data Engineering. Washington: IEEE Computer Society, 1995. 251−260.

[30] Abiteboul S, Quass D, McHugh J, Widom J, Wiener JL. The lorel query language for semistructured data. Int'l Journal on Digital Libraries, 1997,1(1):68−88.

[31] Berners-Lee T, Hendler J, Lassila O. The semantic web. Scientific American, 2001,284(5):34−43.

[32] Manola F, Miller E. RDF primer. W3C Recommendation, 2004. http://www.w3.org/TR/rdf-primer

[33] Hayes P, ed. RDF semantics. W3C Recommendation, 2004. http://www.w3.org/TR/rdf-mt/

[34] Xu XB, Gu N, Shi BL. Data model and query language of semi-structured information. Journal of Computer Research and Development, 1998,35(10):898−901 (in Chinese with English abstract).

[35] Artale A, Franconi E, Guarino N, Pazzi L. Part-Whole relations in object centered systems: An overview. Data and Knowledge Engineering, 1996,20(3):347−383.

[36] Winston M, Chaffin R, Herrmann D. A taxonomy of part-whole relations. Cognitive Science, 1987,11(4):417−444.

[37] Leonard HS, Goodman N. The calculus of individuals and its uses. Journal of Symbolic Logic, 1940,5(2):45−55.

[38] Varzi AC. Parts, wholes, and part-whole relations: The prospects of mereotopology. Data and Knowledge Engineering, 1996,20(3):259−286.

[39] Whitehead AN, Russell B. Principia Mathematica. 2nd ed., Cambridge: Cambridge University Press, 1957.

[40] Bennett B, Cohn AG, Torrini P, Hazarika SM. A foundation for region-based qualitative geometry. In: Horn W, ed. Proc. of the 14th European Conf. on Artificial Intelligence (ECAI 2000). Amsterdam: IOS Press, 2000. 204−208.

[41] Randell D, Cui Z, Cohn A. A spatial logic based on regions and connection. In: Nebel B, Rich C, Swartout W, eds. Proc. of the Knowledge Representation and Reasoning. San Mateo: Morgan Kaufmann Publishers, 1992. 165−176.

[42] Guarino N. Formal ontology, conceptual analysis and knowledge representation. Int'l Journal of Human and Computer Studies, 1995,43(5-6):625−640.

[43] Tsumoto S, Tanaka H. Induction of expert system rules from databases based on rough set theory and resampling methods. In: Ras ZW, Michalewicz M, eds. In: Foundations of Intelligent Systems, Proc. of the 9th Int'l Symp. on Methodologies for Intelligent Systems (ISMIS'96). LNCS 1079, Berlin, Heidelberg: Springer-Verlag, 1996. 128−138.

[44] Farion K, Michalowski W, Slowinski R, Wilk S, Rubin S. Rough set methodology in clinical practice: Controlled hospital trial of the MET system. In: Tsumoto S, Slowinski R, Komorowski J, Grzymala-Busse JW, eds. Proc. of the 4th Int'l Conf. on Rough Sets and Current Trends in Computing (RSCTC 2004). LNCS 3066, Berlin, Heidelberg: Springer-Verlag, 2004. 805−814.

[45] Anhøj J. Generic design of Web-based clinical databases. Journal of Medical Internet Research, 2003,5(4):e27. http://www.jmir.org/2003/4/e27/

[46] Wang GY. Calculation methods for core attributes of decision table. Chinese Journal of Computers, 2003,26(5):611−615 (in Chinese with English abstract).

[47] Wang J, Wang R, Miao DQ, Guo M, Ruan YS, Yuan XH Zhao K. Data enriching based on Rough Set theory. Chinese Journal of Computers, 1998,21(5):393−400 (in Chinese with English abstract).

附中文参考文献:

[4] 刘清.Rough 集及 Rough 推理.第 2 版.北京:科学出版社,2003.

[10] 陈波,周明天.Rough Mereology 内涵浅析.计算机科学,2002,29(9):7−10.

[12] 陈波,周明天.粗糙性理论的列氏总分学分析.计算机科学,2006,33(7):171−175.

[34] 许学标,顾宁,施伯乐.半结构化数据模型及查询语言.计算机研究与发展,1998,35(10):898−901.

[46] 王国胤.决策表核属性的计算方法.计算机学报,2003,26(5):611−615.

[47] 王珏,王任,苗夺谦,郭萌,阮永韶,袁小红,赵凯.基于 Rough Set 理论的“数据浓缩”.计算机学报,1998,21(5):393−400.

陈波(1977－),男,四川德阳人,博士生,主要研究领域为粗糙集理论,Web 智能,基于知识系统,推荐系统,中间件技术.

周明天(1939－),男,教授,博士生导师,CCF 高级会员,主要研究领域为计算机网络,中间件技术, Web 智能,网络安全.