文章编号: 1001- 2486(2010) 04- 0141- 04

基于本体的信息与信息空间及其代数结构*

胡小荣[1], 李建平[1], 黄宏斌[2], 吴 强[1], 黄建华[1]
(1. 国防科技大学 理学院, 湖南 长沙 410073; 2. 国防科技大学 信息系统与管理学院, 湖南 长沙 410073)

摘 要: 实现特定信息环境下基于信息组织的信息资源服务, 需要对信息、信息空间有更合适的、形式化的描述。文章基于本体层次对信息、信息的运算、信息之间的关系以及信息空间给出了数学描述并对其代数结构进行了研究。从而为特定信息环境下的信息资源服务提供理论支撑。

关键词: 本体; 信息; 信息空间; 代数结构
中图分类号: TP391 文献标识码: A

The Algebraic Structure of Information and Information Space Based on Ontology

HU Xiao-rong[1], LI Jian-ping[1], HUANG Hong-bin[2], WU Qiang[1], HUANG Jian-hua[1]
(1. College of Science, National Univ. of Defense Technology, Changsha 410073, China;
2. College of Information System and Management, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: For achieving information resource service in the specific information field based on organizations of information, more appropriate and more formal description of information and information space is needed. The mathematic definition of information, information operations, relationship of information and information space based on ontology, and their algebraic structure were studied. So a theoretical support was provided for information resource service in the specific information field.

Key words: pntology; information; information space; algebraic structure

信息的接收、获取、组织、管理与分析是个人、部门、行业以及工程领域所必须时刻面对的任务。如何解决由海量的、无序的信息造成的"信息泛滥"和"信息缺乏"问题, 如何在信息组织基础上提供用户需求的信息服务等[1] 一直是信息科学研究的重要课题。

科学、合理地描述信息以及所处的信息空间[2-4], 是解决上述问题的理论基础。信息的定义[6] 多种多样, 有基于科学意义、哲学意义和工程意义等方面的类别。每种定义都基于不同的学科领域和研究层次, 无疑具有科学性与合理性。但也不能否认, 在具体应用时存在一定的局限性。

要在一般意义上给信息一个公认的定义是困难的, 也不是必需的。因为信息的组织、管理和应用都是在特定的信息环境以及信息系统中进行的。特别在工程领域的信息环境中, 人们面对的信息绝大部分都是经过处理与组织的结构化或半结构化的特定数据(至少在信息系统的高级层次里是这样)。此时, 对信息给出一个更为确定的、抽象的形式化定义是可能的, 对于建立特定信息环境下的信息资源聚焦服务[1] 更是必需的。

信息空间[8] 是描述信息环境的一个重要概念。关于信息空间的定义和结构亦有不少研究成果。例如, 文献[8] 基于人工智能研究提出了信息的全息空间拓扑结构; 文献[9] 提出了信息以及信息空间的一种数学描述; 文献[7] 基于信息获取科学, 根据信息的"差异性", 提出了信息的数学描述以及信息空间的公理化体系; 文献[1] 基于特定信息环境下的信息资源服务, 提出了信息资源聚焦的概念和基于本体的元数据模型, 并对信息及其信息空间的形式化描述进行了研究。但上述结论都是基于某一特定问题或特定领域得出的, 缺乏一般性。

* 收稿日期: 2009- 11- 18
基金项目: 国家部委资助项目
作者简介: 胡小荣(1965—), 男, 副教授, 在职博士生。

1 本体论与信息概念

1.1 本体

信息具有本体论和认识论两个层次[1]。本体(Ontology)是领域知识的概念化说明,它将特定领域有关的对象、概念及其关系以形式化的说明来严格规定。

概括地说,一个完整的本体应由概念、关系、函数、公理和实例等五类基本元素构成[10]。

定义1[1] 本体 $O=\{C, R, F, A, I\}$,其中:

(1) C:概念。本体中的概念是广义上的概念,它除了包括一般意义上的概念外,还包括任务、功能、行为、策略、推理过程等。本体中的这些概念通常按照一定的关系形成一个层次结构。

(2) $R \subseteq 2^{C\times C}$:概念之间的关系表示概念之间的一类关联。如概念之间的"Subclass-of"关系、"Part-of"关系等。

(3) $F \subseteq R$:一种特殊的关系。其中第 个元素 C 相对于前面 -1 个元素是唯一确定的。函数可用如下形式表示:$F: C_1 \times C_2 \times \ldots \times C_{-1} \rightarrow C$。

(4) A:概念或概念之间的关系所满足的公理是一些永真式。

(5) I:领域内概念实例的集合。在有些本体模型中概念的实例不被看成是本体的组成部分。

由于本体具有较强的数据描述能力、一定的推理能力以及面向语义的输出能力和相似度的计算能力,因此基于本体论来描述语义信息、提供面向语义的查询等,可以使用户更好地理解数据[11]。

在某个特定信息环境,信息就是数据[9]或数据之间的关联(映射)关系。基于本体的思想可以给信息及其相关概念一个形式化描述,进而对其数学结构予以研究。

1.2 基于本体的信息概念

借鉴本体的概念化思想,给出一种基于特定信息环境的信息以及信息空间的数学描述,以便运用数学理论来研究特定信息环境下的信息资源服务。

在特定信息环境下,信息可以分解为描述的对象(实体对象)、对象的属性、属性值、反映实体(类、对象)之间的关联关系以及属性分配与属性赋值的映射关系的函数等等。当然,孤立地看,这些数据,不能被称为信息,只是构成信息的要素。当它们按照一定的方式关联成一个整体时就形成了信息。所以,可以将上述基础数据称为信息元素。

定义2(信息元素) 实体对象 e、属性 i 以及映射关系 r 称为信息元素。

定义3(简单信息) 给定实体对象 $e \in E$,设其属性集为 I_e,属性值集为 L_e,映射集为 R_e。若存在映射 $R_j \in R_e (j=1,2)$, $R_{e1}: I_e \rightarrow \{e\}$,即 $\forall i \in I_e$,有 $R_{e1}(i)=i(e)$(将属性 $i \in I_e$ 赋给 e);$R_{e_2}: I_e \rightarrow L$,即 $\forall l \in L_e$,有 $R_{e2}(l)=i$(将属性值 l 赋给属性 i)。则四元组 $If_e=<\{e\}, I_e, R_e, L_e>$ 称为关于实体对象 e 的简单信息。

定义4 设有简单信息 $If_{e1}=<\{e_1\}, I_{e1}, R_{e1}, L_{e1}>$, $If_{e2}=<\{e_1\}, I_{e2}, R_{e2}, L_{e2}>$。称 $If_+ = <E_+, I_+, R_+, L_+>$ 为 If_{e1} 与 If_{e2} 的和。其中:

(1) $E_+ = \{e_1, e_2\}$:其语义解释为"实体对象 e_1 或 e_2"。

(2) $I_+ = I_{e1} \cup I_{e2}$:含义为集合的并。

(3) $R_+ = \{R_{e+} \mid R_{e+}: Dom(R_{e1}) \cup Dom(R_{e2}) \rightarrow E_+\}$:例如存在 $R_{ej}^I: I_{ej} \rightarrow \{e_j\} (j=1,2)$,则 $R_+: I_{e1} \cup I_{e2} \rightarrow E_+$。

(4) $L_+ = L_{e1} \cup L_{e2}$:含义为集合的并。

定义5 设有简单信息 $If_{e1}=<\{e_1\}, I_{e1}, R_{e1}, L_{e1}>$, $If_{e2}=<\{e_1\}, I_{e2}, R_{e2}, L_{e2}>$。称 $If_{e1} \cdot If_{e2} = <E\cdot, I\cdot, R\cdot, L\cdot>$ 为 If_{e1} 与 If_{e2} 的积。其中:

(1) $E\cdot = \{e_1, e_2\}$:其语义解释为"实体对象 e_1 与 e_2"。特别当 e_1 与 e_2 具有包含关系时 $E\cdot = e_1 \cap e_2$。

(2) $I\cdot = I_{e1} \cap I_{e2}$:含义为集合的交。

(3) $R\cdot = \{R_{e\cdot} \mid R_{e\cdot}: Dom(R_{e1}) \cap Dom(R_{e2}) \rightarrow E_+\}$:例如存在 $R_{ej}^I: I_{ej} \rightarrow \{e_j\} (j=1,2)$,则 $R_+: I_{e1} \cap I_{e2} \rightarrow E_+$。

(4) $L\cdot = L_{e1} \cap L_{e2}$: 含义为集合的交。

若 $E\cdot, I\cdot, R\cdot, L\cdot$ 任一为 Φ, 则称 $If_{e1} \bullet If_{e2}$ 为伪信息。

信息的和与积称为信息的运算。易证以下性质:

性质 1　信息的和与积仍为信息。

性质 2　和、积运算可推广至有限个或可数个信息的运算。

1.3　信息的一般定义

简单信息经过运算得到的信息称为复合信息。经过运算后实体对象可能扩展成为实体对象集、实体类或实体类集 E; 属性分配映射和属性取值映射集可能扩展成为包含实体对象之间、实体对象与实体类以及实体类之间的关系或映射的集合 R。

因此, 实体类(集)、属性集、关系(映射)集以及属性取值集构成了描述信息的要素, 称为信息的维[7]。更一般地, 信息 ξ 可表示成为一个四维向量 $\xi = (\zeta_1, \zeta_2, \zeta_3, \zeta_4)$。其中, $\zeta_i (i = 1, 2, 3, 4)$ 分别表示实体类(集) E、属性集 I、关系(映射)集 R 以及属性取值集 L。它们可以是有限集、可数集或不可数集。这样, 信息的维也就是对应信息元素的集合。由于简单信息只含一个实体对象, 故也可用一个三维向量表示。

定义 6(信息的定义)　称四维向量 $\xi = (\zeta_1, \zeta_2, \zeta_3, \zeta_4)$ 为信息。其中 $\zeta_i (i = 1, 2, 3, 4)$ 分别表示实体类(集) E、属性集 I、关系(映射)集 R 以及属性取值集 L。

若 $\zeta_i (i = 1, 2, 3, 4)$ 至少有一为 Φ, 这时 $\xi = (\zeta_1, \zeta_2, \zeta_3, \zeta_4)$ 实际上不是信息, 称 ξ 为伪信息。

2　信息之间的关系

实现信息资源服务(比如进行信息资源聚焦[1])时, 需要对信息进行一系列操作, 如聚类、分类、查询和比较等, 因此, 需要定义信息之间操作(或关系)。

定义 7(信息的等价)　设 I_A 为所有信息的集合。$\forall \xi_i = (\zeta_{i1}, \zeta_{i2}, \zeta_{i3}, \zeta_{i4}) \in I_A (i = 1, 2)$, 如果 $\zeta_{1j} = \zeta_{2j} (j = 1, 2, 3, 4)$, 则称 ξ_1, ξ_2 为等价信息(相同信息)。记为 $\xi_1 = \xi_2$。

定理 1　信息的等价关系"$=$"是 I_A 上的等价关系。

证明: 显然, 信息的等价关系"$=$"满足自反、对称、传递性。

有时, 信息聚焦服务关心的是信息的某一维或某几维是否相同。因此定义信息的部分等价如下。

定义 8(信息的部分等价)　$\forall \xi = (\zeta_{i1}, \zeta_{i2}, \zeta_{i3}, \zeta_{i4}) \in I_A (i = 1, 2)$, 如果存在正整数 $k_i (1 \leqslant k_i \leqslant 4, 1 \leqslant i < 4)$, 使 $\zeta_{1k_i} = \zeta_{2k_i}$, 且对 $\forall m \neq k_i (1 \leqslant m \leqslant 4)$, 有 $\zeta_{1m_j} \neq \zeta_{2m_j}$, 则称 ξ_1, ξ_2 为部分等价信息。记为 $\xi_1 \overset{k_i}{\sim} \xi_2$(依第 k_i 维等价)。

定义 9(信息的蕴含)　若 $\forall \xi = (\zeta_{i1}, \zeta_{i2}, \zeta_{i3}, \zeta_{i4}) \in I_A (i = 1, 2)$, 有 $\zeta_{1j} \subseteq \zeta_{2j} (j = 1, 2, 3, 4)$, 则称信息 ξ_1 蕴含信息 ξ_2。记为 $\xi_1 \prec \xi_2$。

定理 2　信息的蕴含关系"$\prec$"是 I_A 上的偏序关系。

证明: 易证, 信息的蕴含关系"$\prec$"满足自反性、反对称性和传递性。

3　信息空间及其代数结构

3.1　信息空间

定义 10　所有信息构成的非空集合 I_A 称为全信息空间。即 $I_A = \{\xi \mid \xi = (\zeta_1, \zeta_2, \zeta_3, \zeta_4)\}$。

定义 11　全信息空间 I_A 的非空子集称为信息空间。

显然, 全信息空间 I_A 亦为信息空间(在不引起混淆的情形下统称为信息空间)。

定义 12　设有信息空间 I_1, I_2 以及信息 ξ, 若 $\forall$ 信息 $\xi \in I_1$, 有 $\xi \in I_2$。称信息空间 I_1 为 I_2 的一个子信息空间(简称子空间)。记为 $I_1 \subseteq I_2$。

显然, 任一信息空间都是自己的子空间。

3.2 信息空间的代数结构

定义13 若信息空间 $I=\bigcup_{i=1}X_i$, 其中 X_i 为 I 的子空间。则称 $\{X_1, X_2, \cdots, X\}$ 为 I 的一个分割。

定义14 若 $\{X_1, X_2, \cdots, X\}$ 为信息空间 I 的一个分割, 且 $\bigcap_{i=1}X_i=\Phi$, 则称 $\{X_1, X_2, \cdots, X\}$ 为 I 的一个划分。

定义16 $\forall$ 信息 $\xi\in I$, I_ξ 为 I 中与 ξ 具有某种相关关系 R 的所有信息构成的信息空间, 称 I_ξ 为信息 ξ 关于 R 张成的一个子空间。

在信息空间 I 中, 所有具有某种等价关系 R 的信息构成等价类。所有等价类的集合称为信息商集, 记为 I/R。

定义16 信息空间 I 按照等价关系 R_i 得到的商集族 $X_i=I/R_i$ ($i=1, 2, \cdots,$) 称为 I 上的一个商空间。

定理3 商集族 $X_i=I/R_i$ ($i=1, 2, \cdots,$) 为 I 的一个划分。

证明: (1) $\forall\xi\in I$, 则 $\exists\zeta\in I$ $\exists R_k$ ($1\leqslant k\leqslant$) 使 $\xi R_k\zeta$ (ξ 与 ζ 具有等价关系 R_k)。从而有 $\xi\in I/R_k=X_k$。故 $I=\bigcup_{k=1}X_k$。

(2) 设 $\xi\in X_k$ ($1\leqslant k\leqslant$), 若 $\exists X_l$ ($1\leqslant l\leqslant$, $l\neq k$), 使得 $\xi\in X_l$, 即 $\exists\eta, \zeta$ 使 $\xi R_k\eta$ 且 $\xi R_l\zeta$。由等价关系的传递性知, $R_k=R_l$, 也就有, $k=l$, 即 $\bigcap_{i=1}X_i=\Phi$。

定理4 信息空间 I 关于信息的积构成一个半群。

证明: (1) $\forall\xi_1, \xi_2\in I$, 有 $\xi_1\cdot\xi_2\in I$ (性质1)。

(2) $\forall\xi_i=(\zeta_{i1}, \zeta_{i2}, \zeta_{i3}, \zeta_{i4})\in I$ ($i=1, 2, 3$)

设 $\xi_1\cdot\xi_2=(\zeta'_1, \zeta'_2, \zeta'_3, \zeta'_4)$, 则 $\zeta'_i=\zeta_{1i}\cap\zeta_{2i}$ ($i=1, 2, 3, 4$)。设 $(\xi_1\cdot\xi_2)\cdot\xi_3=(\zeta''_1, \zeta''_2, \zeta''_3, \zeta''_4)$, 则 $\zeta''_i=(\zeta_{1i}\cap\zeta_{2i})\cap\zeta_{3i}$。因为 $(\zeta_{1i}\cap\zeta_{2i})\cap\zeta_{3i}=\zeta_{1i}\cap(\zeta_{2i}\cap\zeta_{3i})$, 记 $\zeta_i=\zeta_{1i}\cap(\zeta_{2i}\cap\zeta_{3i})$ ($i=1, 2, 3, 4$), 所以, $\zeta_i=\zeta''_i$。故 $\xi_1\cdot(\xi_2\cdot\xi_3)=(\zeta_1, \zeta_2, \zeta_3, \zeta_4)=(\xi_1\cdot\xi_2)\cdot\xi_3$。

定理5 信息空间 I 关于信息的和构成一个半群。

4 结束语

信息及信息空间的数学定义及其代数结构的研究, 为实现特定信息环境下信息资源聚焦服务提供了更为可靠的理论支撑。文章的研究结果是对信息相关概念的抽象描述, 所以, 在结合工程实现时适合各种建模方法。有利于对广域分布环境下的信息资源进行组织、索引, 建立信息资源统一视图, 并利用信息资源和用户信息需求的语义关系, 形成以用户关注的信息点为核心的, 满足用户需求的信息资源集合, 以便为用户提供有效的信息服务。

进一步研究方向: 扩充和完善信息的运算, 使信息空间具有更为良好的代数结构; 扩充和完善信息之间的关系, 使之更便于应用到信息资源聚焦服务; 根据信息空间的代数结构, 构造出符合信息资源服务目的的特征子空间, 使之具有更好的代数结构; 与工程实现相结合的研究。

参考文献:

[1] 黄宏斌. 基于语义关系的××信息资源聚焦服务方法及关键技术研究[D]. 长沙: 国防科技大学, 2007.

[2] Liu Z Z, Hang H B. HOM: An Approach to Calculating Semantic Similarity Utilizing Relations Between Ontologies[C]// Proceedings of the Asia Information Retrieval Symposium, 2008.

[3] Papazoglou M P, Proper H A, Yang J. Landscaping the Information Space of Large Multi-database Networks[J]. Data and Knowledge Engineering, 2001, 36(3): 251-281.

[4] 李毕祥. 试论信息空间与信息空间的结构[J]. 情报理论与实践, 1996, (6): 23-25.

[5] Zou H B, Jin H, et al. HRIC: Hybrid Resource Information Service Architecture Bsaed on GMA[C]// Proceedings of the 2005 IEEE International Conference on e-Business Engineering, (ICEBE'05). 2005.

[6] 陈思雨. 论信息的本质定义[DB]. http://www.entropy.com.cn, 2002.

[7] 汪小龙. 信息获取科学的若干问题研究[D]. 中国科技大学, 2003.

[8] 毕家祥. 全息空间的拓扑结构与人工智能模型[J]. 科学探索, 1985, 36: 73-85.

[9] 杜智华. 信息空间的数学模型[J]. 新疆师范大学学报(自然科学版), 2000, 19(2): 12-14.

[10] Schorlemmer. Ontology Mapping: The State of the Art[J]. The Knowledge Engineering Review, 2003, 18(1): 1-31.