

知识链模型研究^{*}

徐焕良¹ 李绪蓉²

(南京农业大学信息科学与技术学院 南京210095)¹

(南京航空航天大学信息科学与技术学院 南京210016)²

摘 要 知识链定义为有限个知识资源组成的有序集合。从“过程”的观点来看,伴随业务过程而存在的一个流动的“知识过程”,也就是知识链。知识链映射业务过程,支持并主导业务过程,更提升、激励业务过程。知识链模型包含了知识资源生产过程和知识资源应用过程。前者定义为知识生产链,其操作包括知识资源的挖掘、表示和存储。后者定义为知识应用链,其操作包括知识资源的识别和选用、执行、评估。一个完备、健壮的知识链模型可以通过定义知识链单元的特性参数来评估,以保证知识链操作过程的高效率、低成本和负荷平衡。

关键词 知识链,知识表示,知识挖掘,知识存储,知识集成,知识评估,领域本体,XML

Research on Knowledge Chain Model

XU Huan-Liang¹ LI Xu-Rong²

(School of Information Science and Technology, Nanjing University of Agriculture, Nanjing 210039)¹

(School of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016)²

Abstract The Knowledge Chain (KC) is defined as a set of knowledge resource units. Knowledge representation, knowledge discovery and knowledge storage are defined as knowledge production chain (KPC), while knowledge identify, knowledge integration and knowledge evaluation as knowledge application chain (KAC). A Knowledge Chain Model is put forward first, which is constructed by KPC and KAC, and then the construction of KPC and KAC is discussed. Further more, knowledge resource unit parameters are given to evaluate the KC qualities such as the high efficiency, low cost and balance load.

Keywords Knowledge chain, Knowledge representation, Knowledge discovery, Knowledge storage, Knowledge integration, Knowledge evaluation, Domain ontology, XML

1 知识链模型

所谓知识,是由一个包含语义信息的特征集以及与之相关的约束和规则集所组成的。在现代企业中,知识作为唯一独特的资源,像农业经济中的土地、工业经济中物质财富和资本一样,成为现代经济中的主要生产要素。知识的竞争,成为企业的核心竞争,也是保证企业可持续发展的动力源泉。以“流动态”的知识为中心的组织和应用成为企业组织生产的主要调控因素。这种无形的“流动态”的知识实际上就组成了企业运行的知识链(Knowledge Chain, KC)。那么,以供应→生产→销售为中心的传统企业的物流管理模式将发展成以知识链管理(Knowledge Chain Management, KCM)为中心的现代企业管理模式^[1]。

本文将知识链(KC)定义为有限个知识资源(Knowledge Unit, $KU_i, i=1, n$)组成的有序集合。

定义1 Knowledge Chain = {Knowledge Unit, $|KU_1, KU_2, \dots, KU_n$ }

令 Op 为定义在 KC 上操作的集合,则定义知识链模型为 {KC, Op }。

定义2 KC-Model = { KC, Op }

其中,一个知识链模型的操作集 $Op = \bigcup Op_i, (i = 1, 2, \dots, 5)$ 。

Op_1 , 定义为知识资源生产的操作集合;

Op_2 , 定义为面向过程的知识资源识别和选用操作集合;

Op_3 , 定义为知识资源修改操作集合;

Op_4 , 定义为知识资源执行操作集合;

Op_5 , 定义为知识链评估操作集合。

图1给出知识链模型的框图。

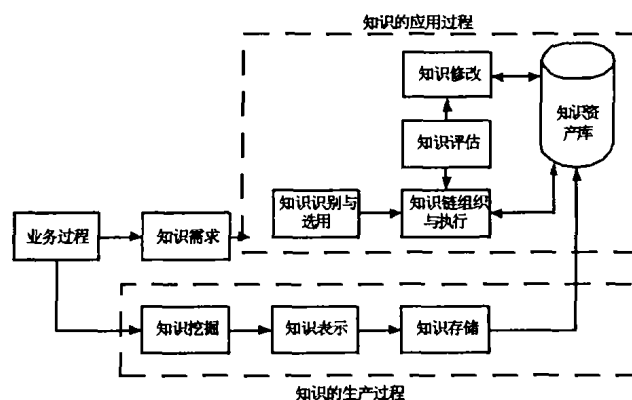


图1 知识链模型

在 KC 操作集合中,知识资源生产,亦即生产知识的过程,可定义为知识生产链,其操作包括知识资源的挖掘、表示和存储。

知识的应用过程,可定义为知识应用链,其操作包括 Op_2, Op_3, Op_4 和 Op_5 。知识资源的识别和选用操作指从知识资源库的选择合适的知识资源。知识资源修改操作指在应用实践中对知识资源的修正,如存储新的知识资源将以取代过时

^{*} 国家重点基础研究发展规划973项目(G1999032701)。徐焕良 博士,主要研究领域为知识工程。李绪蓉 博士,主要研究领域为知识工程/系统集成。

的知识。执行操作指运用知识资源对具体问题的解决。评估操作指对执行的效果、时间、价值等方面进行评价。

从“过程”的观点来看,“知识”存在于任何一项业务过程中,被包含,同时在运行,即存在一个流动的“知识过程”。“知识过程”映射于“业务过程”,支持并主导“业务过程”,更提升、激励“业务过程”^[3]。

综上所述,知识链模型的研究,包括两个方面的内容:

- 1) 知识生产链。包括知识表示、知识挖掘和知识存储。
- 2) 知识应用链。包括知识识别、知识执行、知识修改和知识评估。

2 知识生产链

2.1 知识表示

知识表示(knowledge representation)是研究基于某种约定的知识资源的描述,该描述应该易于为计算机所接受和处理。传统的知识表示方法很多,有产生式规则、框架、谓词和面向对象的方法等。本文提出一种新的知识资源表示方法--UKR(Unified Knowledge Representation)统一知识资源表示方法,即基于 XML 技术表示领域本体的共享语法方式^[9]。

2.1.1 XML 技术 XML,可扩展标记语言(eXtensible Markup Language, XML)是 XML 语义标准通用标记语言 SGML(Standard Generalized Markup Language)的一个子集,它定义一种 Internet 上数据表示和数据交换的新标准。

XML 提供了将大量信息组织成具有确定意义的整体结构(文档)的功能。XML 主要描述数据本身,因此能够很好地表达知识的语义。XML 的 DTD 可以用于定义一个知识表示的语法系统,通过定制的 XML 来解释实例化的知识表示文档。基于 XML 的知识表示显然就具有①语法独立。基于 XML 的统一语法,XML 文档通常包含一个文档类型声明,具有易懂性,也便于计算机处理,并以所需格式显示。XML 表示数据的方式真正做到了独立于应用系统,并且这些数据能够重用。②可扩充。通过对 DTD 的扩展增加新的知识表示能力。③标准化。XML 是国际通用的标准规范。④异构性。正是因为 XML 具有自描述性和可扩展性,所以 XML 可以作为一种通用格式,在异构系统之间进行数据传递。

2.1.2 领域本体 本体论(Ontology)是一个哲学概念,源于哲学上的一个学科。所谓本体(Ontologies)表述了哲学理论的一个术语,是“关于存在及其规律和本质的学说^[2]”,是物质存在的一个系统的解释。Standford 大学知识系统实验室认为,本体是关于主题的形式化和说明性表示,包括它的论域、论域中诸对象的名称、定义及相互关系。在计算机学科,“本体是概念化(conceptualization)的一个显式的规格说明”。

也可以这样说,本体是某一领域的概念化描述,往往表现为一组经过严格定义的词汇和一组刻画词汇意义和词汇关系的理论集合,也即对某一领域的共同理解。如果说面向对象方法适合于表示模块化、有纵向继承关系的知识,那么本体论则适合表示具有横向联系的知识。

本文将形式化描述为:

$\langle \text{Domain Ontology} \rangle ::= \langle D | \{W | \text{Concept, Instance, S | Relationship, Function, Axiom}\} \rangle$

其中, D 是某一领域,为非空集合。一个本体就是某个领域(或一个领域的某个方面)的实体的存在本质。 W 表示领域 D 内词汇集合,包括概念、实例等。 S 表示某一结构,用以明确 W 中词汇之间的关系,如关联、函数、公理等。

在领域工程中,领域本体反映了一个特定领域中可重用的、本质的概念。它提供特定领域的概念定义、概念之间的关

系(领域活动、领域理论和原理等),从而描述领域知识。

2.1.3 统一知识资源表示方法(UKR) 在 XML 技术中,XML 文档将文档的内容、结构和表现分开定义。XML 通过文档类型定义 DTD(或 XMLSchema 方法)的元素定义和属性定义建立 XML 的文档结构。元素定义规定了元素的内容;属性定义规定了每个元素中能够出现的属性。

而领域本体是研究特定领域里的对象、对象属性以及对象间的关系,从而为描述领域具体知识提供术语。

XML 技术与领域本体的结合,即领域本体中的对象类型与 XML 结构中的元素类型的对应表达,运用 XML 的属性表示本体对象的属性,通过 XML 结构的嵌套等来表示本体中的层次和引用等关系,这种领域本体在 XML 结构中的映射,即构成实现互操作和资源共享的统一知识资源表示。

2.2 知识挖掘

依据软件工程的观点,企业的知识可以分为可编码的知识、不可编码的知识两类。所谓可编码知识,是指可以表示为计算机接受和处理的知识。所谓不可编码知识,是指领域专家的思想、经验等。

可编码的企业知识是从大量异构数据源中发现隐藏的策略信息。这个过程就是所谓的知识挖掘。简言之,是从数据中提取知识模式^[7]。企业的知识模式又可以分为两类:数据模式,从结构化数据中挖掘的知识模式(KDD, Knowledge Discovery in Data);文本模式,从半结构化甚至非结构化文本信息中挖掘的知识模式(KDT, Knowledge Discovery in Text)。

为此,本文将知识挖掘分为数据模式的知识挖掘和文本模式的知识挖掘两种。

无论是数据模式的知识还是文本模式的知识,本质上都是从海量的数据资料中,提取隐含的、有用的、潜在的知识。而这些海量的数据可能分布在各种数据库或数据仓库中。如果定义分布的海量数据集 G 为输入,知识挖掘的结果是获取隐含的模式 m ,则可定义知识挖掘的过程为 G 到 m 的一个映射:

$$\text{即 } \xi(G, q) \rightarrow m$$

其中, G 为海量数据集; q 为知识挖掘方法; m 为知识挖掘的结果。

无论数据模式挖掘还是文本模式挖掘,都有数据源的预处理过程。而基于数据仓库技术的数据挖掘,是一个有效的、实现数据预处理的企业知识挖掘模型。

所谓数据仓库(Data Warehouse, DW),是指这样的一种存储方式:将来自于异地、异构的数据源和各种数据库的数据加工后,在一种称之为数据仓库的存储地存储、提取和维护。图2给出基于数据仓库的知识挖掘管理框图。

2.3 知识存储

知识资产库,是存储知识的机构,简称为知识库,如图2所示。

基于 XML 表示的领域本体的知识资产,自然可选择基于 XML 的知识库作为其存储机构。同时,根据挖掘的对象性质,企业的知识资源模式分为数据模式和文本模式。XML 文档也一般可以分为两大类,以数据为中心和以文档为中心。这种对应关系也极有助于利用 XML 对企业知识资源进行表示和存储。

基于 XML 的知识库(简称 XML 知识库),其中 XML 文档为知识资产存储区,XML 的 DTD 或者 XML Schema 为知识资产库的存储模式,SQL 为知识库的查询系统, SAX 或 DOM 为知识库的处理工具等。

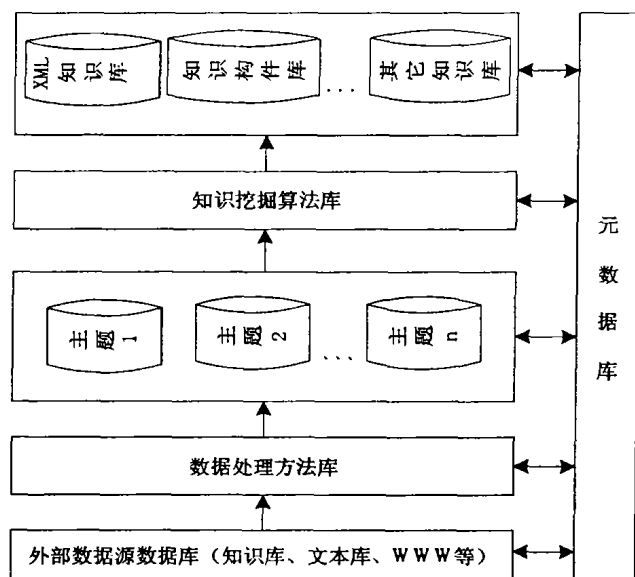


图2 基于数据仓库的知识挖掘管理

XML 资产库适于数据量一般、用户较少、性能要求不高的环境。但对日益开放的系统、多用户并行使用,还必须依赖数据完备、性能良好的传统意义上的数据库形式知识资产库,如知识构件库和其它形式的知识库。

所谓知识构件库存储的是那些可以构件化的知识资源。构件的可复用性、进化性正实现了知识资源的复用和进化的要求(有关问题参见论文“基于最大熵的构件库动态设计”)。

3 知识应用链

3.1 知识集成

实现分布、异构的应用系统的领域知识共享,就必须建立共享的领域知识表达(描述知识的语法,如用 XML 表示)和知识语义协定(基于领域本体抽取领域词汇集)^[6]。

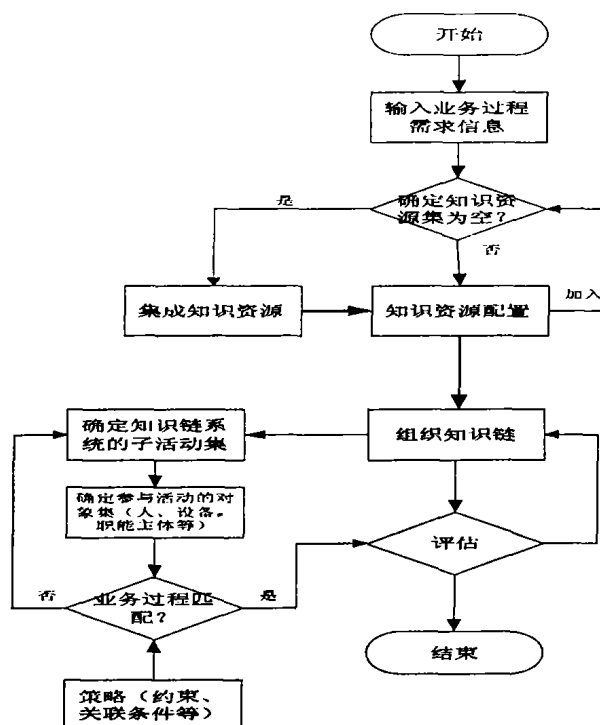


图3 知识资源集成的一般过程

知识链的形成,也就是将有限个 KU_i (即知识资源单元)集成起来以解决某一问题域的问题,而 n 个这样的知识链形成的知识链系统,即可解决更大空间问题域的问题。换句话

说,知识应用链,本质上就是知识资源的集成问题。本文将知识资源集成过程表示为:

知识资源集成 (KC) = $\phi\{KC\text{-Model}, BP\text{-Requirements}, KC\text{-evaluation}\}$

其中, ϕ , 表示集成过程; KC-Model, 知识链模型; BP-Requirements, 业务过程需求; KC-evaluation, 知识链系统的评估。

为说明问题,图3给出知识资源集成的一般过程。

由图可以看出,执行某一业务过程,就会产生知识资源的需求,理论上至少就会有一个有限的知识资源集与之对应,并通过识别、评估、选择出一个最优的知识链,以支持业务过程。如果知识资源集为空,则通过集成、配置知识资源以组成一个知识链。所组成的知识链与业务过程匹配与否,将决定是否需要重新配置知识资源。配置好的知识链,还需进行评估,判断是否满足设计要求,满足与否,则决定是否需要重新组织新的知识链,直至满足要求。

以上是一条知识链的形成过程, n 个这样的知识链就形成知识链系统,即知识链集。知识链系统实际上就完成了整个业务系统。

3.2 知识链评估

知识链评估实际上是对知识链节点上的知识单元的评估,本文给出以下参数作为评估指标:

(1) 知识资源的能力参数— p_1

知识资源的能力参数反映知识资源本身对某一问题域的求解能力,甚至包括执行的效率、求解的成本等。能力越强,复用频度越高,其综合性能越好,则定义为能力参数值越大,相应竞标者优先级越高。

(2) 知识资源的状态参数— p_2

知识资源一旦获取任务,其状态空间就不为空,状态参数相应取值。定义知识资源的状态参数是保证知识资源的合理分配,系统间的负载平衡。分配任务的执行优先级高,剪要性强,工作量大,工作时间跨度长,则状态参数就越大。

(3) 知识资源的成本参数— p_3

知识资源的成本参数是表示知识资源在挖掘、表示、开发和集成等方面所累积的成本值。成本参数越高,复用的代价越大。但,知识资源复用频度越高,知识资源的价值越大,动态的资源成本也会降低。所以说,Agent 的成本参数表示的不仅仅是静态的开发成本,其数值将随着知识资源的复用频度动态变化。

那么,在知识链构成的时候,其构造优先值表示为:

$$Value = f(\alpha p_1, \beta p_2, \gamma p_3, \lambda)$$

式中, α, β, γ , 为加权系数,由经验选取; λ 为知识资源信任度。 λ 的取值将根据知识资源评估值确定,并以作为该知识资源选用的优先级别。于是,建立优选规则:

若存在一知识链构成, $KU_i, \forall KU_i \in \{KU_i | KU_i, i=1, 2, \dots, n\}$; 存在有, $Value = \max\{Value | BidValue, i=1, 2, \dots, n\}$ 那么该知识链构成。

I) 如果 $\lambda_i > \lambda_j$, 则 KU_i 优于 (superior) KU_j ;

II) 如果 $\lambda_i = \lambda_j$, 则确定 $Bidvalue = g(\alpha p_1, \beta p_2, \gamma p_3, \lambda)$

定义 如果 $p_{1i} > p_{1j}$, 则 p_{1i} superior p_{1j} ; 如果 $p_{2i} < p_{2j}$, 则 p_{2i} superior p_{2j} ; 如果 $p_{3i} < p_{3j}$, 则 p_{3i} superior p_{3j} 。

① 有 $p_{1i} > p_{1j}, p_{2i} < p_{2j}, p_{3i} < p_{3j}$, 则 KU_i 优于 (superior) KU_j 。

② $\forall p_{2i} \geq \{p_{2j} | p_{2j}, j=1, 2, \dots, n\} = p_{20}$ 状态系数阈值, KU_i 淘汰。

(下转第192页)

Methodology for Modeling the Control of a Discrete Event Manufacturing System[J]. IEEE Trans. on Systems, Man, and Cybernetics, Part A: Systems and Humans, 1999, 29(6): 573~586

9 和延立, 杨海成, 何卫平, 等. 信息集成与知识集成[J]. 计算机工程与应用, 2003, 4: 38~41

10 Masters J. Structured Knowledge Source Integration and Its Applications to Information Fusion[C]. In: The 5th Intl. Conf. on Information Fusion, Annapolis, MD, July 2002. 1340~1346

11 Salerno J. Information Fusion: A High-Level Architecture Overview[C]. In: The 5th Intl. Conf. on Information Fusion, Annapolis, MD, July 2002. 680~686

12 凌玲, 王学林, 胡于进, 等. 基于知识生命周期的知识集成模型[J]. 华中科技大学学报, 2002, 30(10): 23~28

13 刘炎禄, 俞勇. 面向语义 Web 的知识表示框架[J]. 上海交通大学学报, 2002, 36(9): 1309~1312

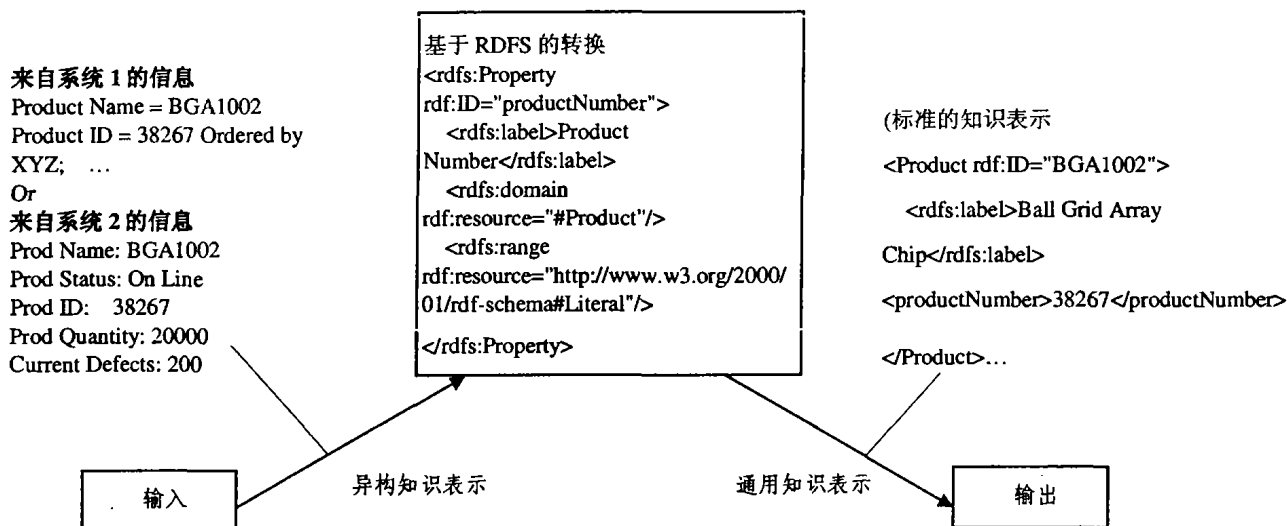


图6 基于RDFS的知识转换实例

(上接第187页)

③ 存在其他, 计算

$Value = \max\{Value | BidValue, i=1, 2, \dots, n\}$ 中标。

同时, 调整 KU 的参数, 赋值

$$\begin{cases} p_1 = p_1 + \delta_1 \\ p_2 = p_2 - \delta_2 \\ p_3 = p_3 - \delta_3 \\ \lambda = \lambda + \delta(\delta_1, \delta_2, \delta_3) \end{cases}$$

式中, $\delta_1, \delta_2, \delta_3$ 为调整值。可根据影响程度建立相应的系数取值算法或数值表。

通过对知识资源的评估, 就决定了知识库中知识资源的淘汰机制。

结束语 当今, 国际管理信息发展从信息管理走向知识管理, 从信息资源开发走向知识资源开发^[8]。知识链模型则是研究知识资源的开发和管理, 即由知识生产链和知识应用链组成的知识资源的开发和管理^[10], 同时, 传统的业务过程价值链也将转化、提升为知识价值链。可以想象, 一个完善的知识链管理系统, 才是真正使企业全面增值的核心动力。本文研究有以下阶段性的成果:

1) 提出了知识链模型, 它由知识生产链和知识应用链两个部分组成。

2) 在知识生产链中, 给出了一个统一的、全局的知识表示方法, 即 UKR; 给出了基于数据仓库的知识挖掘的过程模型; 论述了知识资源的存储方式。

3) 在知识应用链中, 给出了知识集成过程定义、方法及知识链评估体系。

由于知识资源研究涉及的领域和学科很多, 还有许多值

得进一步研究的工作如下: 1) 进一步完善知识生产链中所涉及的知识资源的表示、挖掘和存储; 如知识的表示语法、共享语义的规范(如本体, 领域本体的概念抽取、具体知识的表示、领域本体库的建立等); 知识资产库的构建; 知识挖掘算法的研究, 算法库的建立。

2) 进一步完善知识应用链中知识资源集成机制。

3) 将知识链模型转化为可实现的知识链系统。

参考文献

- 徐焕良. 企业知识资源计划及其关键技术研究: [南京航空航天大学博士论文]. 2003. 12
- 陆汝钊. 世纪之交的知识工程与知识科学. 清华大学出版社, 2001. 9
- 徐焕良, 丁秋林. 知识管理的平台-企业知识资源计划的研究. 软科学, 2003. 6
- 金芝. 基于本体的自动需求获取. 计算机学报, 2000(5)
- Tom Finneran -CIBER, Inc. A Component-Based Knowledge Management System. <http://www.tdan.com/i009hy04.htm>
- 顾新建, 祁国宁. 知识集成初探. 计算机集成制造系统, 2000, 6(1)
- Han Jiawei, Kamber M. 数据挖掘概念与技术. 机械工业出版社, 2002. 5
- Davenport T H, et al. Successful Knowledge Management Projects. Sloan Management Review, Winter 1998. 43~57
- 徐焕良, 丁秋林. 企业知识资源计划系统的集成研究. 计算机工程, 2004(2)
- Schreiber C. 等著, 史忠植, 等译. 知识工程和知识管理. 机械工业出版社, 2003. 4