数据与知识工程导论

胡运发 编著

清华大学出版社

数据与知识工程导论

胡运发　编著

清华大学出版社
北京

内 容 简 介

本书全面介绍了数据、信息和知识共享的理论、方法和技术。第1~3章介绍了数据工程，包括数据表示、数据模型、数据设计、数据分析和数据挖掘等理论和方法。第4~7章介绍了知识工程，包括知识表达、知识推理、知识管理、知识获取、知识利用等理论和方法。数据工程为知识工程提供重要的支持手段，为知识获取提供了无尽的源泉；反过来，知识工程也为数据工程提供了更加智能化的提取信息的手段。软件工程可以也应该从数据工程和知识工程独特的理论、方法和技术中获取有益的借鉴。

本书可作为计算机科学、信息科学、管理科学及人工智能等学科的大专院校师生、社会各界信息化管理人员、工程技术人员的教科书或参考书。

图书在版编目(CIP)数据

数据与知识工程导论/胡运发编著. -北京：清华大学出版社，2003

ISBN 7-302-06240-4

Ⅰ. 数… Ⅱ. 胡… Ⅲ. ①数据管理②知识工程 Ⅳ. ①TP311.13②TP182

中国版本图书馆 CIP 数据核字（2003）第 002260 号

出 版 者：清华大学出版社(北京清华大学学研大厦，邮编 100084)

http://www.tup.tsinghua.edu.cn

责任编辑：钟志芳

印 刷 者：北京昌平环球印刷厂

发 行 者：新华书店总店北京发行所

开　　本：787×1092 1/16　**印张**：26.75　**字数**：612 千字

版　　次：2003 年 4 月第 1 版　2003 年 4 月第 1 次印刷

书　　号：ISBN 7-302-06240-4/TP・3735

印　　数：0001～4000

定　　价：34.00 元

目　录

第0章 绪 言

数据工程是数据库工程的简称。数据库工程已经有许多著作和教材做过介绍。由于数据库工程和知识工程关系密切，是任何知识工程必不可少的一个步骤，因此为了讲解清楚知识工程，必须对数据工程有一个起码的说明。知识工程比数据工程起点要晚一些。知识工程之父费根鲍姆，在1977年的国际人工智能联合会议上，首先提出了知识工程的概念。该概念来源于他在1968年研究成功的一套智能系统(DENDRAL)，这是第一个结合启发式程序和大量专门知识的实用智能系统。足够的物理、化学知识使得它能够按人类的思维方式，根据分子式和质谱仪数据，高效、切合实际地推断出分子结构，其水平达到了人类专家水平。在此影响下，一大批专家系统从化学、数学、医学、生物工程、地质采矿、石油勘探、气象预报、地震分析、过程控制、系统设计、计算机配置、集成电路测试、电子线路分析、情报处理、法律咨询、航天航空、经济决策和军事决策等方面涌现出来。目前，知识工程在知识获取、知识表示以及知识利用等方面已经取得一系列成果。当今，国际软件市场上形成了一门旨在生产和加工知识的新产业——知识产业。国际社会已经从传统的工业经济走向知识经济的新时代，因此研究和学习知识工程相关的原理、方法和技术，就显得十分必要了。

本书是介绍数据工程或知识工程的书籍，存在以下特点：

(1) 数据共享、信息和知识共享在任何情况下都是计算机技术发展的原动力。本书把实现数据、信息和知识共享的理论、方法和技术放在首要的位置上，而不是各种理论、方法和技术的简单汇集。

(2) 数据工程和知识工程关系十分密切。数据工程是表达外延性知识的共享理论、方法和技术；知识工程是表达内涵性知识的共享理论、方法和技术。数据工程为知识工程提供重要的支持手段，为知识获取提供了无尽的源泉；反过来，知识工程也为数据工程提供了更加智能化的提取信息的手段。

(3) 数据工程和知识工程是软件工程的特殊情况，有着许多独特的理论、方法和技术；软件工程可以也应该从中获取有益的借鉴。

为了方便读者阅读本书，下面就三个问题做简单的介绍，作为本书入门的向导。

(1) 什么是数据工程？

(2) 什么是知识工程？

(3) 数据知识工程与软件工程有什么关系？

0.1 什么是数据工程

数据工程是设计和实现数据库系统以及数据库应用系统的理论、方法和技术，是研究结构化数据表示、数据管理和数据应用的一门学科。为了说清什么是数据，首先应说明什么是信息？信息似乎人人皆知，又似乎各人理解不一。辞海的解释是“信息是收信人事先不知道的报道”；控制论创始人维纳的定义是“信息就是信息，不是物质，也不是能量”；申农认为“信息是消除不定性的东西”；耗散论者普里高津则说：“信息是熵”。信息具有二重性：首先它是客观的，是客观事物状态及变化的一种表现形式；其次它是主观的，是主体的一种感受，能够引起主体认识发生变化的客体表现形式称为信息，如客体的表现不能引起主体认识的变化，则不能称为信息。信息在人们的生活中，在人类社会运行中，有着十分重要的作用。信息是中介体，人总是通过感知客观事物表露的信息，才能认识客观事物本来的面貌；信息是粘合剂，由于客观事物表露的信息具有同一性，社会中的人们才可能有共同的感受，形成共同的看法，组成统一的社会；信息是放大器，信息一旦产生出来，可以被无限制地学习、传播和复制，极大地节省其他人开发信息的资源耗费，从而极大地提高人类的生产力。由于计算机技术的飞速发展，数据可成为载荷信息的物理符号，信息生产空前扩大，已成为系统的大规模的生产活动。在计算机领域，信息和数据不可分离，又相互区别。并非任何数据均可表征信息，信息仅是消化了的数据。信息是更本质、更直接地反映现实，而数据仅是信息的一种表现。信息不依赖载荷它的物理设备的改变而改变，而数据则不然。对于计算机而言，信息处理就是数据处理。信息的收集、存储、加工和传播就是数据的收集、存储、加工和传播。数据处理的基本目的在于提取信息，提高人们的判断和决策能力。

数据工程处理的对象是大规模数据。为了处理大规模数据，数据库技术，即大规模数据管理技术，在 20 世纪 60 年代中期以后，逐步地发展起来。其基本特征是数据表示有了统一的模型，数据的使用有了统一的操作，数据管理有了系统化的方法；强调数据的共享，数据和应用数据的程序相互独立。数据库技术成为数据工程中心和基础的环节，数据工程本质上就是数据库工程。数据库工程相关的理论、技术和方法极大地推动了计算机处理大规模数据的能力及处理信息的能力。到了 20 世纪末，人类的信息存量极大地增长，人类的信息流量极快地加速，以数据库为基础的计算机应用，已占到全部计算机应用的 70%以上，许多社会学家一致公认信息时代已经到来。

数据工程设计包括哪些基本环节呢？数据工程设计分为三个基本环节：概念数据模型的分析与设计、逻辑数据模型的分析与设计、物理数据模型的分析与设计。概念数据模型是十分关键的，它是组织化了的不受时间限制的结构化数据模型，具有简明性、完全性、理论性和通用性。广泛使用的实体联系图(E-R 图)提供了一种简单可行的概念模式分析与设计方法。概念数据模型独立于逻辑模式，也独立于特定数据库管理系统(DBMS)，其主要特

点是：(1)能真实地反映现实世界；(2)易于为用户理解；(3)易于修改和扩充；(4)易于向关系、网状和层次等数据模式转换。逻辑数据模型分析与设计的任务就是把概念数据模型结构转换为与 DBMS 所支持的数据模型相符合的数据结构过程。在此过程中，充分运用关系数据库规范化理论成果，指导关系模式的设计并做极小化处理。物理数据模型设计的内容包括：根据 DBMS 提供的功能，确定数据的存储结构、存取路径、存放位置和存储分配。其设计过程需要对时间、空间效率、维护代价和各种用户进行权衡，经评价择优选择较优方案。

当然，组织数据入库是数据库实施的最主要的工作。数据库工程还包括数据库的实施和维护，相当于软件工程中的编码、调试。数据库维护工作又包括安全性、完整性控制，性能的监督、分析和数据库的重构等。

0.2 什么是知识工程

知识工程是设计和实现知识库系统及知识库应用系统的理论、方法和技术，是研究知识获取、知识表示、知识管理和知识利用的一门学科。

知识工程处理的对象是知识。但什么是知识？这似乎是为许多人熟知的较为浅显的概念。实际上，当知识进入经济领域，知识概念就发生了变化。广义地说，知识是人们通过学习、发现或感悟到的对世界的认识的总和，是人类认识的结晶。狭义地说，知识是一种有组织的经验、价值观、相关信息和洞察力的组合。利用知识所构成的框架，人们可进而评价和吸收新的经验与信息。与知识相关联的两个概念是数据和信息。数据是未加工的信息，必须被处理以使其成为有意义的事实的集合。信息是通过将事实和约定的语境关联而导出的。知识是从不同语境中得到的信息之间的关联。当从完全不同的语境中导出一般性原理时，智能出现了。世界发生的知识革命，主要是信息革命引发的一场知识领域革命。信息技术全面渗入到知识活动的全过程，引发了知识的生产、流通和利用的深刻变革。知识的全面数字化，知识活动的计算机化和网络化，彻底改变了知识存在的形式和知识活动的时空关系，促进了知识活动与经济活动的互动联系。由于知识具有不可替代性、不可逆反性、无限制的重复作用性和无限增值的可能性，知识使用越多，价值越大，使用的频度越高，效率越明显。知识是资源、资本、财富，甚至是比资本、财富更重要的资源。因此，知识管理和利用是十分重要的。我们处理“数据”将近 40 年，抽取“信息”也已 20 年，现在所面临的挑战是如何以实用的方式抽取知识，并利用知识帮助人们更深刻、更科学、更迅速地抽取信息与知识。换句话说，我们面临着知识工程上的挑战。

知识工程是一个远比数据工程复杂得多的领域，也是一个比数据工程更富有挑战性的领域。

首先，知识有很多种类。第一类是关于事实和现象的知识(Know-what)；第二类是自然原理或领域规律性知识(Know-why)；第三类是关于技能和能力的知识(Know-how)；第四类是关于是谁的知识(Know-who)。最适合经济生产的或最接近市场商品的知识是 Know-what 和 Know-why 类知识；Know-how 和 Know-who 虽然是客观存在的，但往往是隐式，很难

加以表示。其次，知识表示方式很多，有产生式、函数式、逻辑式、对象式、语义网、框架结构、过程式。逻辑式中有单调逻辑式和非单调逻辑式、精确逻辑式和非精确逻辑式。非精确逻辑式又有概率的、模糊的、粗糙的、证据的、统计的。每种知识表示都对应不同的推理方法，不同的推理方法又有不同的管理方法。直至今天，我们仍不知道是否存在最好的表达方式和最好的知识管理机制，更加困难的是，我们还需要有一种机制能够帮助我们不断地发现知识，因为只有发现知识，才能谈到表示知识和利用知识。

面对这么多困难，我们怎么办？事情总得有一个开始。20 世纪 80 年代以来，关于知识工程的书籍已经出版许多本。这些书籍的一个共同特征是简单地汇集各种知识表示、知识利用和知识获取方法，缺乏与数据工程的紧密联系。我们从数据工程的发展历史中得出一个结论：知识工程必须有自己的坚实的理论基础，必须有一种主要的知识表示方式，其他任何知识表示都应与此知识表示互换。在此条件下，才可能有统一的知识库系统模型和知识管理。在上述基础上，我们才可能谈得上知识的共享和知识的价值。

虽然本书介绍了多种知识表示、多种推理方式，但介绍的目的是为了说明多种知识表示如何与主知识表示的关系，多种推理机制如何在主推理机制的基础上加以扩展后实现。

从知识表示角度看，本书突出逻辑表示方法，这是因为：(1)逻辑表式法有坚实的理论基础；(2)逻辑表示法是结构化表示方式，利于从控制的细节摆脱出来，利于利用与共享；(3)逻辑表示法和数据工程中的数据模型有共同的逻辑理论基础，数据模型表达是外延型知识，逻辑式表达是内涵型知识；(4)语义网是框架表示的特殊情况，而对象表示是框架表示的扩张，它们都可用逻辑方式表示，函数式是逻辑式的特殊情况；(5)知识表示方式与知识利用方式(推理机)关系十分密切，从分级观点看问题，推理机也是一种元级逻辑知识，不同的推理机可以用不同的元级知识表示；(6)多种推理都是逻辑推理的自然扩张。

从上述观点看问题，我们可把任一知识库系统中的知识分为三级：

第一级是数据级(上下文)，这在数据工程部分已做了简述。

第二级是知识库级，我们称为规则级。规则级知识反映自然原理或领域内规律性知识，它分解或综合数据级数据，即操作数据级数据。

第三级是推理机级，知识可以控制知识库级知识，通过增加精度计算模式，精确的推理机可以变成非精确推理机。通过例外情况的不一致性的消除，单调的推理机可变成非单调推理机。

知识工程或知识库工程设计一般分为三步：

第一步，数据库工程或数据工程设计，我们在 0.1 节中已经做了说明。数据工程中的概念模型描述了问题领域解空间。

第二步，逻辑模式的设计。知识库工程中的逻辑模式，指的是在问题领域内概念空间中概念移动的规则，即在什么条件下，从概念空间中的一点移动到另一点。

例如

is-a(x,y)表示概念 y 是概念 x 的实例。

规则“如果 is-a(x,y)并且 is-a(y,z)则 is-a(x,z)”表示如果“概念 y 是概念 x 的实例，并且概念 z 是概念 y 的实例，那么概念 z 是概念 x 的实例”。

第三步，针对特定的知识库，确定规则级知识的逻辑结构和可信度表示。

在特殊情况下，还可用元级知识表达推理机推理方式和推理策略。

知识工程中最为困难的问题是知识获取。因为知识是稀有资源，并非任何人都可轻易获取。一般说来，获取知识的一类方法是知识工程师获取领域专家们的知识；另一类方法就是 KDD(Knowledge Discovery in Database，[Fayyad 96])，即从数据库数据中发现知识。数据库中的数据实际是组织了的客观事实，我们可以从中挖掘出所需的知识，这就是所谓数据挖掘(Data mining)。数据库知识是外延知识，知识库知识是内涵知识，同类外延知识的共同属性就是内涵知识。数据挖掘的确很有意义，它是连接数据工程和知识工程之间的桥梁。数据挖掘的核心是挖掘算法，常用的数据挖掘算法有关联规则挖掘算法、序列模式挖掘算法、分类挖掘算法、聚类挖掘算法、异常测算法等。

0.3 数据知识工程和软件工程的关系

数据知识工程是为了创建一种数据(知识)库系统以及数据(知识)库应用系统。数据(知识)库及其应用系统都是一种软件系统，所以数据和知识工程可以且应该遵循软件工程的一般原则。

软件工程的一般原则是什么呢？Frity(NAV69)在 NATO 会议上给出的软件工程定义至今仍是各种软件工程定义的基础。

软件工程定义 1：软件工程是为了经济地获得可靠的和能在实际机器上高效运作的软件而建立和使用的好的工程原则。

该定义给出软件工程原则的一般属性：(1)经济性，即低成本(低的人员、设备、时间的消耗)；(2)可靠性，即软件的创建和使用运行要可靠，不能导致错误和意外；(3)实际运行，软件是不同计算机上可运行的代码(不是文档)，也隐含了可移植性；(4)高效性，即软件具有高的时空性能。

符合上述性质的软件工程原则是好的。“好”也是一个相对的概念，好中求好，永无止境。上述定义是从结果看问题。结果好，一切皆好，这当然没有问题。问题是什么原则才能产生好的结果呢？定义中并没有给出好的原则是什么。IEEE(IEE-93)看出这一点，所以给出了软件工程另一定义。

软件工程定义 2：(1)将系统化的、规范的、可度量的方法用于软件的开发运行和维护过程，即将工程化方法应用于软件中。(2)对(1)中所述方法的研究。

此定义明确好的软件工程方法必须是系统化的、规范的和可度量的，好的软件结果必须用好的工程化方法作用于软件开发过程。

这个定义同样是开放的，它没有明确界定什么是系统化的、规范的、可度量的。

任何软件系统均需要 3 个阶段，即定义阶段、开发阶段和维护阶段。

定义阶段要解决的问题是“做什么”(what to do)，包括系统或信息工程、项目开发和需求分析。开发阶段要解决的问题是“怎么做”(how to do)，包括软件设计、代码生成和