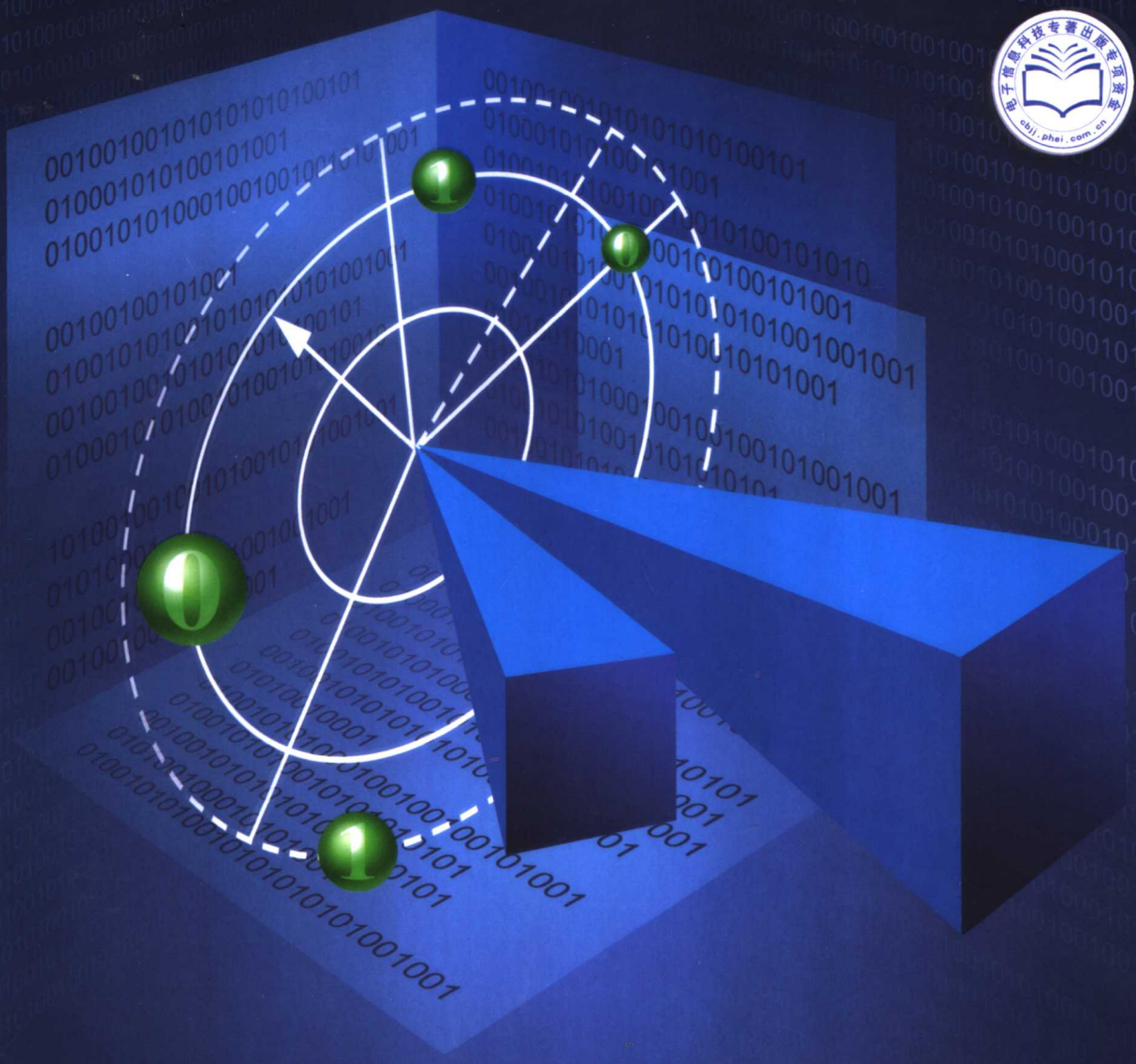


基于内在机理的知识发现理论及其应用

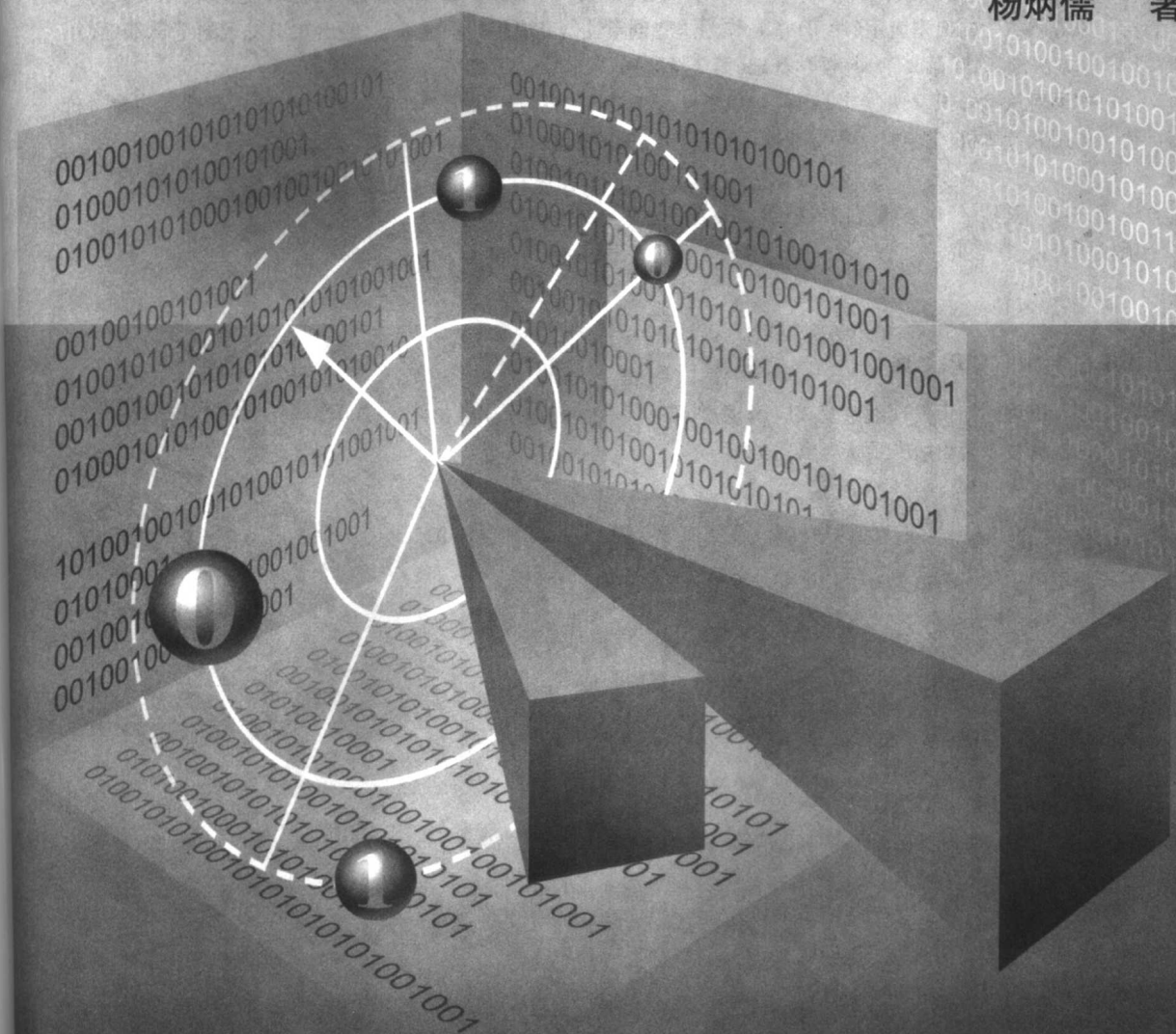
杨炳儒 著

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

电子信息科技专著出版专项资金资助出版

基于内在机理的知识发现理论及其应用

杨炳儒 著



電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
北京·BEIJING

内 容 简 介

本书定位在认知科学、认知心理学和认知生物行为的全新理念上，提出并实现了以三个核心定理为贯穿的三个原理或机制——内在机理研究（第4章），由此诱导出新结构模型（第5章），派生出新技术方法（第6章），引发出新型实用智能系统（第7章），并提出讨论复杂类型知识发现（第8章）；至此，系统地构建具有五个层次的基于内在机理的知识发现理论 KDTIM；在 KDTIM 的指导下，设计实现了具有自主知识产权的集成化组合构件式知识发现软件系统 ICCKDSS（第9章）；基于 KDTIM 与 ICCKDSS，给出在农业、现代远程教育网与气象等领域中的典型应用（第10章）；作为这些创新性研究成果的理论基础，给出知识发现的逻辑基础（第1章），方法论基础（第2章）与哲学基础（第3章）。

本书对从事知识发现、知识管理、知识工程、人工智能、计算机科学等研究的科技人员具有重要的参考价值。可用做计算机、信息技术等专业博士生、硕士生的高级教材。

图书在版编目（CIP）数据

基于内在机理的知识发现理论及其应用/杨炳儒著. —北京：电子工业出版社，2004.4
电子信息科技专著出版专项资金资助出版
ISBN 7-5053-9822-9

Ⅰ. 基… Ⅱ. 杨… Ⅲ. ① 认知科学 ② 知识学 Ⅳ. ①B842.1 ②G302

中国版本图书馆 CIP 数据核字（2004）第030432号

责任编辑：李 岩
印 刷：北京大中印刷厂
出版发行：电子工业出版社
北京市海淀区万寿路173信箱 邮编100036
经 销：各地新华书店
开 本：787×1092 1/16 印张：19.75 字数：492千字
版 次：2004年4月第1次印刷
印 数：3000册 定价：33.00元

前　言

Acton 曾说过："科学家最必需的东西莫过于科学的历史和发现的逻辑……"。我们力求沿着知识发现领域研究进化的历史和发展的轨迹去完备、系统地概括与综合其基本内容；利用结构主义的合理内核与方法去展现每章内容的基本结构；合乎逻辑地集成我们自身历经十余年时间所做出的创新性研究成果（全书 90%以上的内容皆来自于我们独立发表的论著与研究报告）；注重在理论与应用的结合上阐明规律与技法。

本书在写作过程中，曾得到美国加州大学伯克利分校 L. A. Zadeh 教授与佐治亚州立大学 Hamid R. Arabnia 教授的热情鼓励与大力支持，在此深表谢意！此外，我指导的博士生宋威、梁开健、游福成、苏占东、李欣、曲文龙、樊广佺、李卫东与硕士生营志刚等在书稿的整理方面做了大量的工作；对于书中正文与参考文献中涉及到的专家学者、支持本书出版的出版社及有关人士，在此一并致以真诚的谢意！

由于本书论题具有知识发现领域前沿的动态跟踪的特征，加之作者水平有限，故决定了本书的探索性与尝试性；对于书中的错误与疏漏之处，诚请赐教，以期改进与完善。

杨炳儒

2003 年 10 月于北京

目　录

绪　论

20 世纪 30 至 40 年代，在罗森勃吕斯(Rosenn Boris)领导的科学方法讨论班的过程中，罗森勃吕斯和维纳(Winner)越来越深刻地感到："在科学发展的过程中，在那些已经建立起来的学科或部门之间，还存在着一些被人忽视的无人区，而正是在这些领域都可能得到最大的收获。"维纳把这些领域比喻为未开垦的科学处女地。一份最近的 Gartner 技术报告，揭示了在人工智能、认知科学、数据库技术等多学科交叉的"未开垦的科学处女地"上开放出的绚丽花朵——知识发现确具旺盛的生命力，并预示了它广阔的发展前景。

Karl R. Popper 曾指出："虽然我同意，科学知识只是日常知识或常识知识的发展，但是我坚决认为那些把自己限制在分析日常或常识知识，或者这种知识在日常语言中的表述的人们，必定完全看不见认识论的最重要、最令人激动的问题。"知识发现自 1989 年产生以来，经历了从结构化数据挖掘到复杂类型数据挖掘(Web 信息挖掘、多媒体信息挖掘等)，从简单应用到复杂应用的发展过程。目前，国际上知识发现的研究主要以知识发现的任务描述、知识评价与知识表示为主线，以有效的知识发现算法为中心。这是在相当长的一段时间内保持的主流与基调，完全表现出学科发展初期的自然与思维特征。在此背景下，不可避免地出现了一些现有结构模型与技术方法难以解决的问题。如：固有知识库的实时维护，领域专家局限性的束缚，知识库与数据库的同步进化，先验知识如何耦合到知识发现的过程中，动态挖掘进程中所发现规则的演化、评价和可理解性，数据量增大后引起的算法失效等。我们设想：能否跳出主流发展的运行轨道去俯视其现实发展的图景，从而另辟蹊径——将知识发现系统看做特定认知系统、生物有机体演化系统、认知物理(生物)系统，揭示其潜在的本质、规律与复杂性，然后再反作用于主流发展，提高其挖掘效率、扩展其应用的深度与广度。在 11 个科研项目的资助下，在经历了十余年的探索后，我们终于得到了肯定的回答。我们整个研究过程可划分为如下四个阶段：

第一阶段(1989—1997 年)：顺应主流发展，奠定理论基础，加强技术储备(方法论基础、逻辑基础和哲学基础)；

第二阶段(1997—2000 年)：针对结构化数据挖掘(KDD，Knowledge Discovery in Database)与基于知识库的知识发现(KDK，Knowledge Discovery in Knowledge base)，提出以三个机制为重要内核的内在机理研究，构建了五层次的 KDTIM 理论系统，设计与实现了 KDD* 软件系统，并成功地应用到农业生产领域中；

第三阶段(2000—2002 年)：针对复杂类型数据挖掘(Web 信息挖掘等)，完善了五层次的 KDTIM 理论系统，设计与实现了 ICCKDSS(V1.0)软件系统，并成功地应用到现代远程教育网领域中；

第四阶段(2002—2003 年)：针对复杂类型数据挖掘(多媒体信息挖掘等)，进一步拓展了五层次的 KDTIM 理论系统(成为基于复杂系统信息挖掘的一般性的知识发现理论)，完善了 ICCKDSS(V1.0)技术，并成功地应用到气象领域中。这一阶段的工作重点是加速推广应用与

产品化进程。

本书集成了以上各阶段的研究成果，经过融会贯通与逻辑加工后，于这一年轻学科的发展期，在国际上率先尝试性地提出一套较为完备的理论系统（第 1～8 章）、软件系统（第 9 章）与应用系统（第 10 章）；以抛砖引玉，为更富创新性的知识发现的知识发现理论、方法与技术体系的诞生做点奠基工作。

作为本书贯穿性的主线是基于内在机理的知识发现理论（KDTIM），其总体架构如图 0-1 所示。

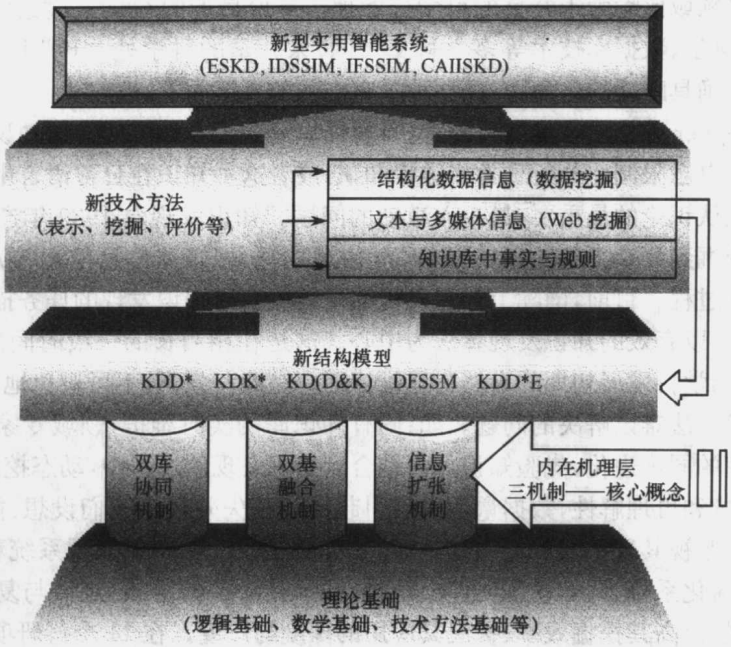


图 0-1 KDTIM 总体架构

（1）理论基础层（第 1～3 章）

提出广义归纳逻辑因果模型，提出因果关系能行可判定方法、因果关系定性推理模型与方法，提出专家知识的归纳获取机制，提出语言场与语言值结构的知识表示方法，提出基于事实与规则的 KDK 归纳发现算法等。通过理论基础层研究，我们着重解决知识发现现实发展中的重要问题之一是基于知识库的知识发现。目前，国际上仅见到极少数文献涉及到这方面的研究，但都没有较为系统地形成结构模型和算法。我们针对海量知识库的特征，综合运用归纳发现算法和归纳逻辑评价方法（基于卡尔纳普归纳逻辑与基于柯恩归纳逻辑的），解决了基于知识库的知识发现这一崭新的研究专题的奠基与探索。其他理论基础研究成果所解决的知识发现现实发展中的问题，在此不一一赘述。

（2）内在机理层（第 4 章）

不拘泥于算法改进等跟踪性研究，而另辟蹊径——提出三个机制（原理），即双库协同机制（结构对应定理、可达关系概率估计定理、启发与维护协调算法等）、双基融合机制（过程等价定理，R，S，T 三类协调算法等）和信息扩张机制（参数演化定理、“不动点原理”与“突变性原理”

等),揭示了知识发现系统的潜在本质、规律与复杂性;形成了 KDTIM 的理论支柱。通过内在机理层研究,我们着重解决知识发现现实发展中的重要问题之一是在动态挖掘进程中,知识的评价、可理解性和知识的优化问题。目前,对于这个问题的研究基本上都是停留在一个挖掘进程中的剖面上现实时态的给予解决(尚未见动态性的研究)。到底在数据库和知识库不断变化的情况下,如何给出所发现规则的价值度量,预见其取舍?如何理解规则?如何表征算法在形成规则中的有效性,从而启发对算法复杂性的改进和对算法的优化?信息扩张机制对于这些问题都给予了明确的回答。其他内在机理研究成果所解决的知识发现现实发展中的问题,在此不一一赘述。

(3) 新结构模型层(第 5 章)

内在机理诱导出五个新的结构模型,基本上是传统的结构模型融入相应的机制而成,即 KDD* = KDD + 双库协同机制,KDK* = KDK + 双基融合机制,KD(D&K) = KDD* + KDK*,KDD*E = KDD* + 信息扩张机制。另外,针对复杂类型数据挖掘构造了"发现特征子空间模型 DFSSM",Web 文本挖掘结构模型与 Web 访问信息挖掘结构模型等。新的结构模型优化了知识发现进程与运行机制。通过结构模型层研究,我们着重解决知识发现现实发展中的重要问题之一是在同时存在海量数据库和海量知识库的时候,如何利用 KDD* 的方法在数据库中发现知识,同时利用 KDK* 方法在知识库中发现知识,即实现在"综合基"(基于数据库同时基于知识库)上的知识发现。这时知识发现的结构模型是何种形态?我们提出的 KD(D&K)结构模型较好地解决了这一问题。其他结构模型研究成果所解决的知识发现现实发展中的问题,在此不一一赘述。

(4) 新技术方法层(第 6 章)

由内在机理与结构模型派生出九种经大量实例验证优于同类方法的新技术方法;如优于 Apriori 算法及其改进型的挖掘关联规则的 Maradbcm 算法,基于遗传算法与梯度下降法的聚类规则挖掘算法,基于 DFSSM(优于 VSM)的 Web 文本分类与聚类挖掘算法,源于气象信息的相似模式挖掘算法,基于小波神经网络的混沌模式挖掘算法,基于广义归纳逻辑因果模型的因果关联规则挖掘算法,基于事实与规则的 KDK* 归纳发现算法,关联规则的系统自动化评价方法等。目前,在 KDD 现实发展中,如何产生高效的、可扩展性的挖掘算法是其主流发展中的核心问题。通过技术方法层研究,我们着重解决知识发现现实发展中的重要问题之一是提出优于经典的 Apriori 算法的 Maradbcm 算法。目前,挖掘关联规则的较权威的 Apriori 算法尽管做了若干改进,但有一个根本性的问题始终得不到解决,即知识库中的领域知识与背景知识没有通过一定的、具体的计算机程序直接参与挖掘进程,从而局限在一个封闭的挖掘系统中。本层中提到的 Maradbcm 算法解决了这一问题,经过大量的对比实验证实其在缩小搜索空间、不失有意义规则、增强自主性、快速运行、提高挖掘效率等方面的优越性。其他技术方法所解决的知识发现现实发展中的问题,在此不一一赘述。

(5) 新型实用智能系统层(第 7 章)

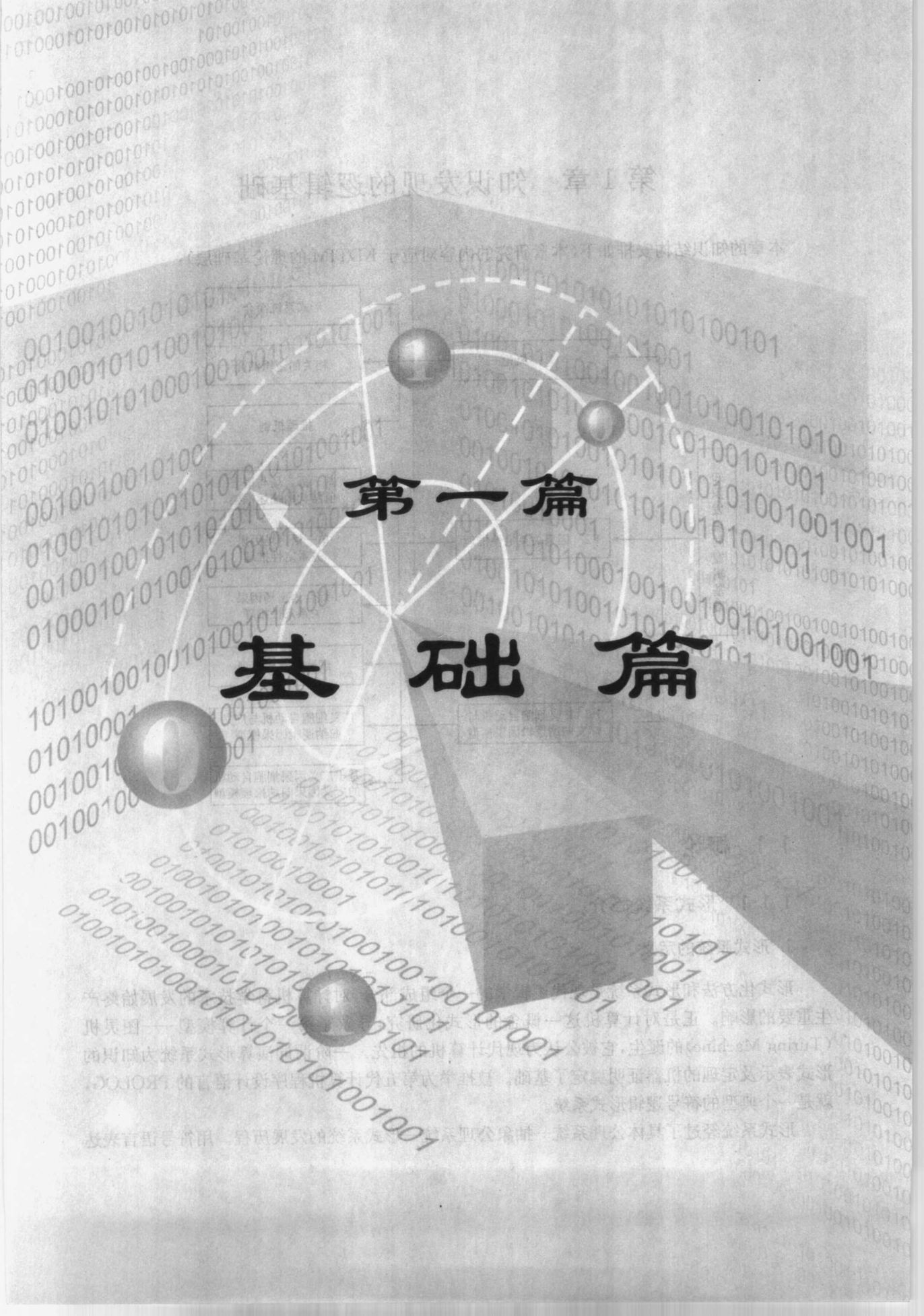
将上述知识发现的创新性的基础、机理、模型、技术与经典实用智能系统相融合,引发出四类典型的新型实用智能系统 ESKD,IDSSIM,IFSSIM,CAIISKD 等。新型实用智能系统在结构、技术、功能与智能化程度等方面均大大优于经典实用智能系统。通过技术方法层研究,我们着重解决知识发现现实发展中的重要问题之一是智能系统中"知识贫乏"这一瓶颈问题。以

经典的专家系统为例，无论如何进行改进，都只能从书本知识和领域专家那里通过知识获取构件，以及通过推理机制等得到知识库中的知识，再没有获得新知识的其他有效途径。而 ESKD 恰恰通过新型 KD(D&K)结构模型，从新的知识源即数据库与知识库中产生新的知识，从而在根本上改变了“知识贫乏”这一窘境。其他新型实用智能系统所解决的知识发现现实发展中的问题，在此不一一赘述。

注：第 8 章“基于复杂类型数据的知识发现(信息挖掘)”的内容，可对应地插入到内在机理层、新结构模型层、新技术方法层与新型实用智能系统层中。

在 KDTIM 的指导下，我们设计并实现了具有自主知识产权的大型集成化组合构件式知识发现软件系统 ICCKDSS(第 9 章)，其 1.0 版本涵盖了全新模型与算法的结构化数据挖掘(KDD*)与部分非(半)结构化数据挖掘(Web 挖掘、图像挖掘)。ICCKDSS 的主要特点是各模块可以按需选取和集成；且由于设计了完善的接口，也可以单独与其他系统集成；另外易于功能的扩展和构件的重用；可以面向相当广泛的一类信息挖掘问题。

经近期的四次查新证实：综合利用 KDD*，KDK*，KD(D&K)，DFSSM 和结构化数据挖掘技术、Web 挖掘技术以及 ICCKDSS 来解决农业、现代远程教育网以及气象云图的信息处理问题(第 10 章)均未见相关报道。相应的应用领域专家在经过大量的实际运行与对比实验后，对运用基于内在机理的新结构模型、新技术方法与新软件系统所得到的结果给予充分的肯定。这些新的、成功的应用也是对 KDTIM 与 ICCKDSS 的有效验证，并在不同程度上解决了在实际应用中，对于不确定和不完全信息的处理、非规范知识的处理和知识发现主流发展中某些关键技术的处理等问题，具有很强的学科带动性和对知识发现主流发展的驱动性。

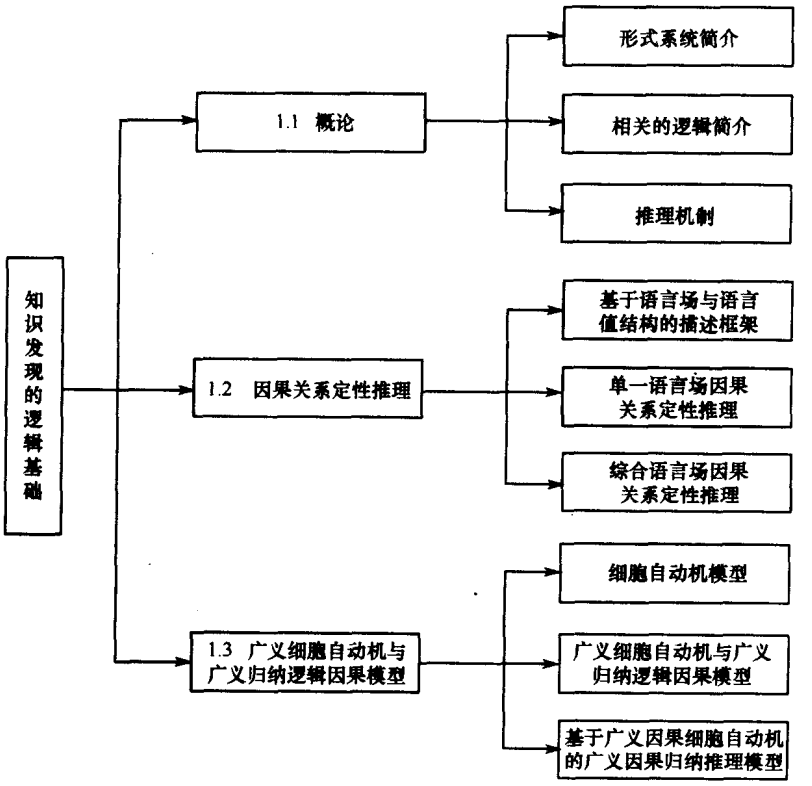


第一篇

基础篇

第 1 章　知识发现的逻辑基础

本章的知识结构安排如下(本章研究的内容对应于 KDTIM 的理论基础层)：



1.1　概论

1.1.1　形式系统简介

1. 形式系统的定义

形式化方法和形式系统是现代逻辑学的一个组成部分，对计算机科学技术的发展始终产生重要的影响。正是对计算机这一概念的形式化研究，导致了第一个计算模型——图灵机(Turing Machine)的诞生，它被公认为现代计算机的祖先。一阶谓词演算形式系统为知识的形式表示及定理的机器证明奠定了基础。被推举为第五代计算机程序设计语言的 PROLOG，就是一个典型的符号逻辑形式系统。

形式系统经过了具体公理系统—抽象公理系统—形式系统的发展历程。用符号语言表达

的形式系统是一个符号体系，它由两大部分组成。第一部分是符号体系的组成部分，包括形式语言所使用的符号及各类符号串的形成规则。这一部分也称为形式系统的语言部分。第二部分是符号体系的推演部分，称为理论部分，包括称为公理的符号串集合，称为推理规则的符号串重写规则集合，以及由它们重写生成的被称为定理(Theorems)的那些符号串集合。更严谨一些，我们可以如下定义形式系统。

定义 1.1：一个形式系统 FS 可用五元组〈Σ, TERM, FORMULA, AXIOM, RULE〉来表示。

(1) 非空集合 Σ 称为 FS 的符号表(Alphabet)，其元素称为符号。

(2) Σ 上全体字的集合 Σ^* 的一个子集 TERM，其元素称为 FS 的项(Terms)，TERM 有子集 VARIABLE，其元素称为变元(Variables)。

(3) Σ^* 的一个子集 FORMULA，其元素称为 FS 的公式(Formulas)；FORMULA 有子集 ATOM，其元素称为原子公式(Atomic Formulas)；公式集与项集不交，即 TERM$\cap$FORMULA$=\varnothing$，而项和公式常统称为表达式(Expressions)。

(4) FORMULA 的一个子集 AXIOM，其元素称为 FS 的公理。

(5) FORMULA 上的 n 元关系的集合 RULE，即

$$\text{RULE}=\{r \mid \exists n(n \text{ 是正整数 } \wedge n\geqslant 2 \wedge r\leqslant \text{FORMULA}^n)\}$$

其元素称为 FS 的推理规则。FORMULA^n 表示笛卡儿积：

$$\underbrace{\text{FORMULA}\times\text{FORMULA}\times\cdots\times\text{FORMULA}}_{n\text{个}}$$

其中，$\Sigma(\Sigma^*)$，TERM，FORMULA 是 FS 的组成部分，而 AXIOM，RULE 构成 FS 的演绎部分，由它们出发可造就整个系统的全部定理，即 FS 的理论，可进行给定前提下的演绎。在 FS 中给出了关于定理、证明、理论及演绎等概念的定义，从而实现逻辑学的形式化。

2. 元语言

对象语言是形式系统本身所使用的语言，它是一种形式语言。对形式系统的研究，在某种意义上就是对这一语言的研究；对象语言中的符号一方面是客观对象的抽象，另一方面它们作为符号客体又是被研究的对象。

元语言是对形式系统及其语言进行研究时，进行通信、交流所使用的语言。通常使用的元语言是一种借助形式符号的自然语言。

元语言中的成分包括：

(1) 对形式系统各组成成分的称谓，例如术语、项、公式、公理等；

(2) 对系统分析讨论时所使用的逻辑术语，例如“当且仅当”、“如果……那么”、“所有”、“存在”等；

(3) 描述形式系统有关性质的元语言术语，例如一致性、完备性、可判定性等。

元语言中使用的形式符号，大体分为以下三类。

第一类：元语言中继续使用对象语言中的符号，但它是在不同意义上使用这些符号。对象语言是使用这些符号的客观实体，而元语言使用这些符号时，只是把它们看做它们自身的名来使用。

第二类：在元语言中要使用语法变元来表示对象语言中的一类符号或一类表达式。这些语法变元的使用，使公理、规则的交代简单明了，一个模式便交代了可数无穷多的公理或规则。

第三类：元语言也采用一些自身所需的符号来代替一些需反复使用的术语。例如 $\vdash_{FSPC} A$ 表示 A 是 FSPC 的定理。

3. 元理论

元理论是研究形式系统时所取得的理论成果的总体。它包括研究形式系统使用的理论工具及研究形式系统的理论成果。元理论中关于形式系统的研究成果可分为三部分。一是关于形式系统语构(Syntax)的研究，在这类研究中，形式系统只是一个没有具体意义的符号体系，被研究的是符号串的推演(重写)规律；二是关于形式系统的语义(Semantics)的研究，在这类研究中，元理论赋予形式符号一定的意义，研究形式系统在被做出各种解释时的性质，特别是系统中公式的真值性质；三是关于形式系统语构与语义关系的研究，特别关注形式系统的整体特性，并对系统做出评价。

1.1.2 相关的逻辑简介

相关的逻辑系统结构如图 1-1 所示。

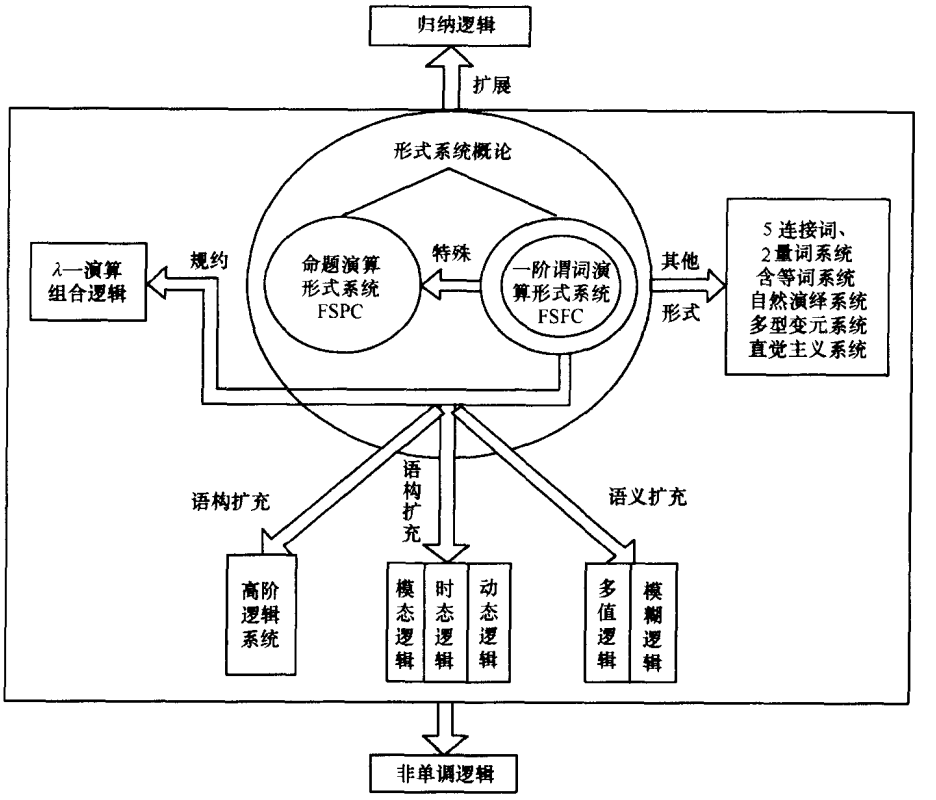


图 1-1 相关的逻辑系统结构图

一般把逻辑学发展划分为四个时期：古代逻辑、中世纪逻辑、近代逻辑和现代逻辑。古代