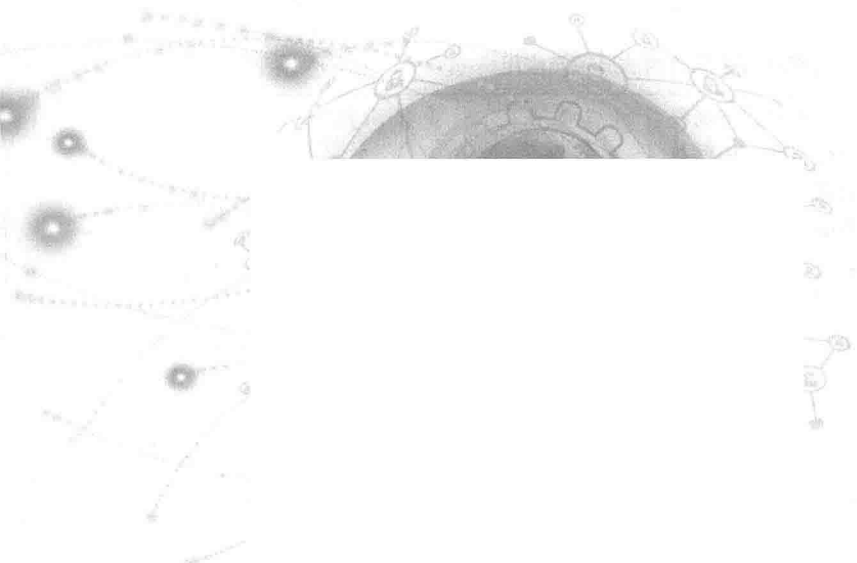
人工智能与数据挖掘的原理及应用

黄尚科◎编著

延边大学出版社

人工智能与数据挖掘的原理及应用

黄尚科◎编著



延边大学出版社

图书在版编目(CIP)数据

人工智能与数据挖掘的原理及应用 / 黄尚科编著
. -- 延吉 : 延边大学出版社, 2019.7
ISBN 978-7-5688-7289-8

Ⅰ. ①人… Ⅱ. ①黄… Ⅲ. ①人工智能－研究②数据采集－研究 Ⅳ. ①TP18②TP274

中国版本图书馆CIP数据核字(2019)第138222号

人工智能与数据挖掘的原理及应用

编　　著：黄尚科
责任编辑：田莲花
封面设计：孟　微
出版发行：延边大学出版社
社　　址：吉林省延吉市公园路977号　　邮　　编：133002
网　　址：http://www.ydcbs.com　　E-mail：ydcbs@ydcbs.com
电　　话：0433-2732435　　传　　真：0433-2732434
制　　作：山东延大兴业文化传媒有限责任公司
印　　刷：北京建宏印刷有限公司
开　　本：880×1230　1/32
印　　张：7.25
字　　数：166千字
版　　次：2019年7月第1版
印　　次：2019年7月第1次印刷
书　　号：ISBN 978-7-5688-7289-8

定价：30.00元

前言

Preface

本书系统地阐述了人工智能与数据挖掘的基本原理、方法和应用技术，比较全面地反映了国内外人工智能和数据挖掘领域的最新进展和发展方向。

人工智能是计算机科学的一个分支，是一门研究机器智能的学科，即用人工的方法和技术，研制智能机器或智能系统来模仿、延伸和扩展人的智能，实现智能行为。人工智能的机器学习、数据挖掘、计算机视觉、专家系统、自然语言处理、模式识别、规划和机器人等相关的应用带来了良好的经济效益和社会效益。人工智能的长期目标是建立人类水平的人工智能。

数据挖掘是指在大量的数据中挖掘出信息，通过认真分析来揭示数据之间有意义的联系、趋势和模式。而数据挖掘技术就是指为了完成数据挖掘任务所需要的全部技术。数据挖掘是近年来伴随数据库系统的大量建立和万维网的广泛应用而发展起来的一门技术。数据挖掘是交叉性学科，它是数据库技术、机器学习、统计学、人工智能、可视化分析、模式识别

等多门学科的融合。

本书采用深入浅出的语言，将枯燥的理论知识用易于理解的形式表达出来。本书力求科学性、实用性、可读性。内容由浅入深、循序渐进、条理清晰，让读者在有限的时间内，易于理解所学内容，掌握人工智与数据挖掘能的基本原理与应用技术。

博观而约取，厚积而薄发。本书是通过吸取国内外多种人工智能与数据挖掘的资料，参考最新的同类教材和有关文献的研究成果编写而成，在此谨向这些教材和文献的作者表示感谢。由于人工智能与数据挖掘发展很快，又是一门统计、计算机、数据库等多个专业领域的交叉学科，并且由于作者水平和成书时间所限，难免存在挂一漏万之嫌，书中不妥和错误之处在所难免，恳请各位专家和广大读者不吝指教和帮助。

目录

Contents

原理篇

第1章 人工智能的基本原理

第1节 知识表达与语言

知识是人类智慧的源泉，知识是对客观事物及其联系的认识和知道如何去改造客观环境的能力。知识产生于人类的实践和思维活动。

知识表示是智能系统的重要基础，是人工智能中最活跃的研究部分之一。知识表示就是要研究用机器表示知识的可行的、有效的、通用的原则和方法，即把人类知识形式转化为机器能处理的数据结构，是一组对知识的描述和约定。我们主要介绍几种常用的知识表示方法：谓词逻辑、产生式系统、框架表示、语义网络、脚本、概念图、面向对象表示等。

随着互联网的高速发展，特别是语义Web服务的需要，本体(Ontology)成为愈来愈重要的知识表示方法。

一、知识的定义

人类社会正从工业社会转入信息社会。信息社会就是大

量生产知识，将物化的知识力量运用于生产，运用于社会，为社会带来生产力新的飞跃，相应地带来社会生活新的变化。20世纪40年代香农通过研究通信和控制系统中信息传送的共同规律，以及提高信息传输系统的有效性和可靠性问题，创立了信息论。

在计算机科学中，信息是根据表示数据所用的约定，赋予数据的意义。数据是事物、概念或指令的一种形式化的表示形式，以适合于人工或自然方式进行通信、解释或处理。信息是数据所表达的客观事实，数据是信息的载体，与具体的介质和编码方法有关。

知识是人通过实践，认识到的客观世界的规律性的东西。知识是经过加工的信息，它包括事实、信念和启发式规则。知识一般可分为陈述性知识、过程性知识和控制性知识。陈述性知识提供概念和事实。例如，在一个智能检索系统中，陈述性知识包括说明具体事实的数据库内容。规则中表示问题的知识称作过程性知识。智能信息检索系统中利用过程性知识处理陈述性知识。用控制策略表示问题的知识常称为控制性知识。控制性知识包含有关各种处理过程、策略和结构的知识，常用来协调整个问题求解的过程。从计算机程序组织来看，一般智能系统可以看成是三级结构，即数据级、知识库级、控制级。数据级是关于求解的特殊问题及其当前状态的陈述性知识。知识库级是具体领域问题求解的知识，它常常是一种过程，说明怎样操纵数据去达到问题求解，反映动作的过程。控制级是过程性知识的控制策略，相应于控制性知识或元知识。

对于很多大型而复杂的基于知识的应用系统，常常包含多种不同的问题求解活动，不同的活动往往需要采用不同方式表示的知识，是以统一的方式表示所有的知识，还是以不同的方式表示不同的知识，这是建造基于知识的系统时所面临的一个选择。统一的知识表示方法在知识获取和知识库维护上具有简易性，但是处理效率较低。而不同的知识表示方法处理效率较高，但是知识难以获取，知识库难以维护。那么在实际当中如何来选择和建立合适的知识表示方法呢？这可以从下面几个方面考虑：①表示能力，要求能够正确、有效地将问题求解所需要的各类知识都表示出来。②可理解性，所表示的知识应易懂、易读。③便于知识的获取，使得智能系统能够渐进地增加知识，逐步进化。同时在吸收新知识的同时应便于消除可能引起新老知识之间的矛盾，便于维护知识的一致性。④便于搜索，表示知识的符号结构和推理机制应支持对知识库的高效搜索，使得智能系统能够迅速地感知事物之间的关系和变化；同时很快地从知识库中找到有关的知识。⑤便于推理，要能够从已有的知识中推出需要的答案和结论。

我们从逻辑抽象的角度出发，可把知识分为以下几类：①对象知识（object knowledge）：关于客观事物及其联系的知识。②进程知识（processes knowledge）：关于事态发展或从事活动的知识。③技巧（know-how）：通过经验体会而获得的知识。④常识（commonsense）：泛指普遍存在而被普遍认识了的客观事实这一类知识。⑤元知识（meta-knowledge）：关于知识的知识。

二、知识表达问题

如果将AI问题到产生式系统(即GDB、RB、CS)的映射统称为问题的表达,那么,问题域知识到全局数据库GDB和规则库RB等结构上的映射,便是知识的表达问题。

知识的表达从构造知识库系统角度看,需要注意以下几个方面:

语法数据结构——知识存储和访问的形式。

语义解释过程——它给知识结构赋予含义。带语义的数据结构也就是知识用于问题求解程序中将产生有知识的行为。

用自然语言表达知识,对人类来说早已是很普通的事情。但是自然语言存在两个主要问题,即多义性(ambiguity)和模糊性(fuzzy)。

(一)多义性

①在灯谜会上,猜谜语,猜中者可得到一份包括一本笔记本和一支铅笔或一支圆珠笔的奖品。

②They are flying planes.

在句子①句法结构中,"和"和"或"这两个连接词先后出现在同一个句中,造成了这一句子语义上的二义性。既可理解猜中的人可得奖品是一本笔记本加上一支铅笔或圆珠笔(二者择一),也可理解为奖品是一本笔记本和一支铅笔(同时获得两样奖品)或者一支圆珠笔(或者选择一支圆珠笔作为奖品)。句子②的意思是"他们正在飞飞机";若把"flying plane"理解为一个专业术语,指的是"牛头刨床",则句子②的意思是"它们是些牛头刨床"。

通过上面的简单例子可以看出，自然语言存在着语法和语义的多义性。

即使是人，面对这种多义性，也往往会产生误解。对于计算机，更是自然语言理解或机器翻译所面临的一个难题。

(二)模糊性

人类无论是在日常生活中，还是在业务活动中，都运用着大量的不精确的、不完全的、不确定的、不肯定的概念。例如，很大、比较小、相当好、可能有、这个比那个大、贵一点、慢一点走等。这些都属于所谓的模糊概念(fuzzy concepts)。

再如，用自然语言来表达这样一个常识：

“如果要卖的电视机是旧的电视机而且很便宜，那么这台电视机的质量很可能不好。”分析这个句子就会发现，含有很多模糊概念。例如，像“旧的”“很便宜”“不好”等，都是一些模糊谓词；这台电视机的质量可能不好就是模糊事件。

除了上述多义性和模糊性外，自然语言句子的结构不规整性，句子成分彼此相关，模块性很差等，也都给直接用自然语言表达知识带来处理和理解上的困难。何况，任何一种自然语言的语法和语义都有例外情况，尚未被人们充分认识。自然语言本身也在不断发展变化。因此，为了建立知识库，实现机器智能，必须寻求不同于自然语言的形式化的知识表达式。

目前研究者们已经提出了多种知识表达方式，它们具有不同的数据结构和解释过程。究竟什么样的知识表达方式较好，目前还没有一个确定和通用的评价标准。因为一种知识表达方式的性能往往直接与求解问题类型及性质相关。从广

义的意义上讲，所有知识的表达方式是等效的，它们最终都要被归入某种计算机语言。然而，对一个具体的问题域，却只有部分表达方式能简洁、恰当地突出它的结构特征，从而便于计算机高效处理。

虽然难以得到评价知识表达方式的通用标准，但根据AI问题求解的共同特征，可归纳得到以下几条评价准则：

1.表达范围广泛和准确性好。表达方式既能正确反映客观领域知识，又可表达多种类型的知识。

2.模块性和可理解性好。模块性好的知识表达方式具有以下特点：①容易理解。②便于修改。③适于并行处理且并行度高。④访问效率高。知识库的合理组织依赖于知识的表达形式，知识库的组织形式将直接影响知识系统的效率。

智能来源于知识，知识表达是AI研究的核心课题之一。逻辑、语义网络、过程以及框架都是目前被普遍采用的知识表达方式，它们都有各自的侧重点，有突出的优点和弱点。

总之，知识表达的研究方向可概括为：①探讨新的表达方式，以便更有利于知识库和知识处理效率的提高。②标准化。主要指表达术语、原语(primitives)以及表达技术的标准化。③非精确性知识常识、时间变化以及关于知识库自身知识的表达和处理方法的研究。④知识表达的综合模式。在这种方式下，知识表示系统将具有自动任务划分和任务与表达模式匹配的能力。

三、知识在AI问题求解中的作用

(一)问题求解的前提

前提知识存储在产生式系统的全局数据库和规则库中，它

们是问题求解的依据。如果前提知识不充分，得到的解答就可能不完全。

(二)控制问题求解(搜索)的进程

人们的实践证明：

推理方法固然重要，但问题求解中的启发式知识更重要。要克服组合爆炸，提高搜索效率，必须利用知识制导，就是将特殊问题领域知识嵌入搜索策略中去。

知识是有关信息关联在一起形成的信息结构，具有相对正确性、不确定性、可表示性和可利用性等特点。对知识的表示可以分为符号表示法和连接机制表示法。

目前的知识表示一般都是从具体应用中提出的，后来虽然不断发展变化，但是仍然偏重于实际应用，缺乏严格的知识表示理论。而且由于这些知识表示方法都是面向领域知识的，对于常识性知识的表示仍没有取得大的进展，是一个亟待解决的问题。

知识表示对专家系统十分重要。知识可以用许多种方法来分类，如，先验知识和后验知识，过程的、说明的和缺省的知识。逻辑方法、产生式、语义网络、框架是专家系统中常用的知识表示方法。脚本和概念图是自然语言理解中常用的方法。面向对象的知识表示方法是一种综合的方法。每一种知识表示方法都有优缺点，在设计一个基于知识的系统前，应先决定选用哪种方法可以更好地解决问题。

与其用一个工具去解决所有的问题，不如对特定的问题选用最合适的工具。

四、自然语言理解

自然语言是人类特有的用于交流的手段，对它的理解是一件困难的事情，这不仅需要有语言学方面的知识，而且还需要有与所理解话题相关的背景知识，必须很好地结合这两方面的知识，才能建立有效的自然语言理解程序。人类对自然语言理解和处理开始于机器翻译，是人工智能领域中早期较活跃的研究领域之一，但是由于它难度很大，至今仍未能达到很高的水平。

自然语言是指人类语言集团的本族语，如汉语、英语等，它是相对于人造语言而言的，如C语言、JAVA语言等计算机语言。语言是思维的载体，是人际交流的工具，人类历史上以语言文字形式记载和流传的知识占到知识总量的80%以上。就计算机应用而言，有85%左右的应用都是用于语言文字的信息处理。在信息化社会中，语言信息处理的技术水平和每年所处理的信息总量已成为衡量一个国家现代化水平的重要标志之一。

自然语言理解作为语言信息处理技术的一个高层次的重要研究方向，一直是人工智能领域的核心课题，也是困难问题之一，由于自然语言的多义性、上下文有关性、模糊性、非系统性和环境密切相关性及涉及的知识面广等原因，使得很多系统不得不采取回避的方法；另外，由于理解并非一个绝对的概念，它与所应用的目标相关，例如是用于回答问题、执行命令，还是用于机器翻译。因此，关于自然语言理解，至今尚无一致的、各方可以接受的定义。从微观上讲，自然语言理解是指从自然语言到机器内部的一个映射；而从宏观上看，自然语言是