中医药知识发现可靠性研究

封毅 著

浙江工商大学出版社

中医药知识发现可靠性研究

封　毅 著

浙江工商大學出版社
ZHEJIANG GONGSHANG UNIVERSITY PRESS

图书在版编目(CIP)数据

中医药知识发现可靠性研究 / 封毅著. —杭州：浙江工商大学出版社，2016.12

ISBN 978-7-5178-1988-2

Ⅰ. ①中… Ⅱ. ①封… Ⅲ. ①中国医药学—研究 Ⅳ. ①R2

中国版本图书馆 CIP 数据核字(2016)第 304974 号

中医药知识发现可靠性研究

封 毅 著

责任编辑 谭娟娟 汪 浩
封面设计 叶泽雯
责任校对 尤建忠
责任印制 包建辉
出版发行 浙江工商大学出版社
(杭州市教工路 198 号 邮政编码 310012)
(E-mail:zjgsupress@163.com)
(网址:http://www.zjgsupress.com)
电话:0571-88904980,88831806(传真)
排 版 杭州朝曦图文设计有限公司
印 刷 浙江新华数码印务有限公司
开 本 710mm×1000mm 1/16
印 张 10
字 数 201 千
版 印 次 2016 年 12 月第 1 版 2016 年 12 月第 1 次印刷
书 号 ISBN 978-7-5178-1988-2
定 价 30.00 元

前　言

随着大数据时代的来临，知识发现可靠性已成为知识发现与数据挖掘领域中一个重要但容易忽视的主题。伴随着数据挖掘技术的广泛应用，有一个问题逐渐引起人们的关注，即在什么条件下知识发现是可靠的，或者说在什么条件下所发现的知识是可靠的。近年来在知识发现可靠性方面的研究，大多关注于某一具体数据挖掘模型下的可靠性问题。而对于不同模型间存在的可靠性共同主题，比如数据质量、评估方法等等，迄今为止仍没有一项系统性研究。针对知识发现可靠性的共同主题，进行分阶段、系统化的总结和梳理，已成为知识发现可靠性研究的一大迫切需要。

在知识发现技术所应用的各个领域，有一个领域特别需要知识发现可靠性的研究，即中医药领域。作为中华民族重要文化财富和学术成就的中医药，近年来面临着生存和发展的挑战。如何把这一挑战化为中医药发展的契机，利用知识发现技术促进中医药的发展，已成为中医药研究人员的一项重要课题。近年来的中医药信息化工作已为知识发现创造了有利条件。然而，由于中医药数据自然语言性强，数据表达涵义丰富，表达方式多样化，而且在数据质量上还面临较大问题，在具备这些特征的数据上所进行的知识发现，相比其他领域来讲，就更加需要关注和研究知识发现可靠性问题。

本书即是针对中医药知识发现可靠性的学术研究专著。全书共分八章，前两章为绪论部分，主要介绍知识发现可靠性、中医药知识发现的研究背景与现状。中间四章为主体部分，从知识发现整个生命周期的各个阶段对可靠性因素进行探讨，提出了知识发现可靠性框架 PBRF-KD，并针对中医药知识发现中比较突出的可靠性问题，重点探讨中医药知识发现中的结构性因素、表达性因素和信任性因素三大问题及优化方法。最

后两章为应用和总结部分，介绍了中医药知识发现原型系统及其对知识发现可靠性的关注和体现，在此基础上总结全书并展望未来工作。本书的研究工作与贡献主要包括如下几个方面：

(1)提出了基于过程的知识发现可靠性框架。

针对现有知识发现可靠性研究模型相关的特点，提出了一个与模型/应用无关的知识发现可靠性框架 PBRF-KD，该框架采用基于过程的思路对知识发现整个流程中的各个阶段和可靠性因素进行了梳理，归纳出了 7 种可靠性相关因素。该框架为知识发现项目设立了整套与可靠性相关的蓝本。

(2)提出了结构相关的可靠性因素的优化方法。

分析了中医药知识发现中与结构相关的可靠性因素，主要指数据完整性。针对文本型字段的完整性问题，提出了基于顺序半相关度量的中医药文本缺失字段填补方法。针对中医药文献类别标签缺失的问题，提出了基于 M-Similarity 的多标签文本分类方法。

(3)提出了表达相关的可靠性因素的优化方法。

分析了中医药知识发现中与表达相关的可靠性因素，包括表达粒度和表达一致性。针对表达粒度，提出了基于规则的表达粒度细分方法。针对表达一致性，提出了基于本体的表达一致化方法。该套方法有助于提高中医药与表达相关的可靠性。

(4)提出了信任相关的可靠性因素的优化方法。

分析了中医药知识发现中与信任相关的可靠性因素，主要指数据可信度。针对中医药特有的数据可信度问题，提出了基于历史文献认可度的数据可信度衡量方法，和基于互联网知名度的数据可信度衡量方法。此外，基于这两种可信度衡量方法，提出了基于数据可信度的加权频繁模式挖掘算法，并在消渴方和脾胃方数据集上获得了有意义的结果。该套方法有助于提高中医药与信任相关的可靠性。

本书适于从事数据挖掘、中医药信息学的科技工作者阅读使用，也可作为高等院校数据挖掘、中医药信息学、管理科学与工程等信息类、医药类、管理类相关专业研究生和本科生的教学参考书。本书在写作过程中，得到了浙江工商大学、浙江大学、中国中医科学院等单位专家和学者的大力支持和帮助，在此表示衷心感谢。本书同时得到了浙江省“十二五”省级实验教学示范中心重点建设项目“现代商贸信息技术与工程实验教学中心”的支持，在此表示感谢。同时感谢业内专家对本书内容的指导、推荐和帮助。由于作者水平有限，书中难免有疏漏和不妥之处，恳请读者批评指正。

目 录

CONTENTS

第1章 绪 论

1.1 知识发现研究背景

20世纪80年代末以来，在“数据丰富，信息贫乏”的海量数据困境驱动下，在数据库、机器学习、统计学等学科的交叉影响下，知识发现和数据挖掘技术开始兴起，并获得了持续和高速的发展。目前，知识发现和数据挖掘已经成为子领域众多、内涵非常丰富的学科领域[1]。

KDD(Knowledge Discovery in Database，数据库中知识发现，可简称为知识发现)一词，最早于1989年8月在美国底特律市召开的第一届KDD国际学术研讨会上正式形成(当时还是IJCAI 1989的一个workshop)，用于强调数据所驱动的发现其终端产品是“知识”[1]。1992年，William J. Frawley, Gregory Piatetsky-Shapiro和Christopher J. Matheus[2]把KDD定义为：“从数据中抽取出隐含的、以前未知的和可能有用的信息的非平凡过程”。1996年，Usama Fayyad, Gregory Piatetsky－Shapiro和Padhraic Smyth[3]对KDD下了精确的定义：“从数据中获取有效、新颖、有潜在应用价值和最终可理解的模式的非平凡过程”。这是关于KDD的经典定义，被研究者和产业界广为接受。

数据挖掘(Data Mining，DM)的概念则是在1995年美国计算机年会(ACM)上提出的。Usama et al[3]认为，数据挖掘是KDD过程当中的一步，即通过使用各种数据分析和发现算法，在可以接受的时间内产生模式。但许多人在使用过程中把数据挖掘和KDD看作是同义词，不加以严格区分。所以，在大多数场合下，人们认为广义的数据挖掘等同于

KDD，即从存放在数据库、数据仓库或其他信息库中的大量数据中挖掘出有趣知识的过程[4]。

从数据挖掘技术类型来看，常见的主题包括频繁模式和关联分析、分类和预测、聚类分析等等[4]。而从数据挖掘的数据类型来看，早期以普通的关系数据库为主，近年已扩展到了基于 Data Stream、Graph、文本、多媒体、Web 数据等复杂类型数据的挖掘[4]。

近 20 年来，在学界和工业界，有大量研究着力于研发高效的数据挖掘算法，以提高其准确性、执行性能、可扩展性等。相对来说，知识发现的其他方面，所受到的关注就少得多。实际上，除数据挖掘算法本身外，知识发现领域中还有不少值得关注的主题。其中一个重要但容易忽视的主题，就是知识发现的可靠性(reliability)。

随着知识发现和数据挖掘技术的广泛应用，人们发现，在有些情况下，知识发现过程的鲁棒性无法完全满足，或者所发现的知识并不可靠。这就给知识发现界提出了一个重要的问题，即在什么条件下知识发现是可靠的，或者说在什么条件下所发现的知识是可靠的。毋庸置疑的是，这一问题对于知识发现应用成功与否影响重大。因此，为推进知识发现在各领域的成功应用，充分发挥计算机技术在数据分析挖掘任务中的优势，实现真正有意义、有价值的知识发现，很有必要对这一主题加以仔细研究。

近几年来，已有部分研究开始对知识发现可靠性这一主题进行探讨。其中，H. Dai 等[5]对 Graph Discovery 的可靠性进行了研究，探索了样本空间大小、图复杂性、链接数目等因素对可靠性的影响。E. Smirnov 等[6]提出 Reliable Classification 的概念，并提出基于 version space 的相关算法。Berka[7]在分类模型中引入一个 verification 步骤以考虑分类结果的可靠性。Wang et al.[8]对复合项关联规则的可靠性进行了分析，提出了相关度量和方法。2006 年起，数据挖掘主流国际会议 IEEE ICDM 已举办了三届专门讨论知识发现可靠性问题的 workshop(RIKD)。

已有的研究在很大程度上推进了对知识发现可靠性这一主题的理解和分析。不过，归纳起来看，已有的这些研究大多关注于某一具体数据挖掘模型下的可靠性问题(比如[5]中的图挖掘模型，[6][7]中的分类模型)。而在不同模型之间，实际上还存在着不少共同的知识发现可靠性主题，比如数据质量、评估方法等等。到目前为止，还没有一项研究从知识发现整个生命周期的各个阶段对可靠性问题进行探讨，而可靠性问题在实际应

用中又贯穿着知识发现项目的始终。因此，针对知识发现可靠性的共同主题，进行分阶段、系统化的总结和梳理，已成为知识发现可靠性研究的一大迫切需要。

1.2 中医药知识发现研究背景

本研究的应用背景是中医药领域的知识发现。近年来中医药的发展呼唤知识发现技术，而中医药本身的领域特点又决定了进行中医药知识发现迫切需要对知识发现可靠性加以探索。

1.2.1 中医药的跨越式发展需要信息技术

中医药学是中华民族五千年优秀文化和科学历史发展的积累，为人民的健康和生存质量的提高做出了极大贡献。然而，作为中华民族重要文化财富和学术成就的中医药，近年来面临着生存和发展的挑战。如何把这一挑战化为中医药发展的契机，实现中医药的跨越式发展，是中医药界需要解决的一个关键问题。在这中间，知识发现技术可以发挥重要的作用。

从中医来讲，中医整体的、系统的理念与西医以还原论为主的思维方式有着本质的区别。近年来随着生命科学的发展，人类逐渐认识到还原论的局限性，系统思维得到的重视与日俱增。系统生物学的兴起和发展[9]，即是一个明证。“人类基因组计划”的出现，更催生了生命科学的第三次“浪潮”——大科学研究[10]。以生命科学为核心的交叉大学科研究将成为21世纪科学研究的主体内容之一[11]，而计算机和信息技术是大科学研究不可或缺的组成部分。在这一浪潮中，以整体思维为方法论的中国传统医学，应抓住这个机遇，完善和发展自己的理论，加强与计算机和信息技术的交叉，推动大生命科学的发展。KDD作为知识发现的信息技术，必能在这一浪潮中发挥其独特的作用。对于中医理论的确证、完善和发展，对于中医人才培养的快速化，KDD技术都具有重要意义。

从中药来讲，近年来我国的中药发展面临着巨大的市场压力。目前我国的中成药产品缺乏国际市场竞争力，使我国中药产品在国际中成药市场的占有份额仅为3%—4%左右。与此对应的是，大量的原材料出口到日本、韩国等国家，被制成中成药产品。这中间的关键问题是我国对于

中药方剂有效成分的分析和提取，缺乏有效、快速的方法。此外，由于人类生活条件、生存环境的改变，使人类身心疾病增加，疾病谱发生了很大的改变，免疫功能障碍性疾病、环境污染疾病、肿瘤、有源性疾病、外伤及营养过剩或营养不良性疾病、老年性疾病明显增加，疾病从单纯治疗型向预防、保健、治疗、康复相结合的模式转变，现有的化学药品已不能完全适应社会的需要[12]，人类健康呼唤天然药物的大规模开发和应用。面对回归自然的发展趋势与现代市场化的要求，加快我国中药的发展已成为一个迫切的任务。因此，利用知识发现技术，加深对方剂配伍规律和药性理论的理解[13]，加快从中医药方剂当中提取有效成分的过程[14]，对于中医药产品的研究、开发、生产，对于推动中药理论和教学的发展，具有不可替代的作用。

1.2.2 中医药积累的海量数据需要知识发现

几千年来，中医药领域的无数临床实践与理论研究积累了海量的科学知识，这些知识包含在中医药古籍、文献以及当前的临床研究文献中。据统计，目前国内收藏的辛亥革命以前的中医药学古籍文献 1.3 万多种，其中在社会上流通较广的古籍近 1000 种。与此同时，现当代出版的大量中医药图书和期刊中也包含着有价值的大量信息。仅中国中医科学院图书馆就收录了 1911 年以后出版的中医药图书达 12000 多种，中医期刊 230 多种。根据中国中医药期刊文献数据库的数据显示，仅 1987—2003 年发表的中医药文献就高达 530700 篇。面对如此海量的中医药数据，如何有效地利用这些宝贵资源就成了发展中医药必须面对的一个问题。而知识发现所擅长的正是从海量的数据当中寻找有意义的模式、知识，完成普通人不能够完成的任务，是分析中医药的海量数据所需要的技术。

1.2.3 中医药信息化成果为知识发现创造了条件

应用知识发现技术的前提和基础是海量数据得以数字化。浙江大学计算机学院 CCNT 实验室和中国中医科学院（原中国中医研究院）于 1998 年就开始合作搭建中医药科技数据库群，并成功建立了集成全国 17 个分中心的分布式多库融合平台[15]。2002 年开始，双方逐步将原有多库融合平台转变为语义网格平台 DartGrid，提供动态的语义注册、语义查询等功能。通过全国 30 余家中医药学院、大学和科研院所近 300 名科技工作者的数据录入工作，该平台目前已集成了 50 余个数据库，100 余 G 数

据，其中包括中国中医药期刊文献数据库（收录了中医药文献约 80 万篇）、中国中药数据库（收录中药 8000 余种）、疾病诊疗数据（收录了各科疾病约 3700 种）、中国方剂数据库（收录古今中药方剂约 85000 首）、方剂现代应用数据库（9600 余种方剂的应用信息）、中国中药化学成分数据库（收录中药化学成分 3000 余种）等数据库。

同时，为建立中医药一体化语言系统并解决系统集成中出现的语义问题，浙江大学与中国中医科学院合作，开始基于 Semantic Web 技术搭建大规模的中医药领域本体——中医药学语言系统（Unified Traditional Chinese Medical Language System，简称 UTCMLS[16]）。该本体包含了中医药领域的基本概念，并定义了概念之间的关系。目前，中医药本体已经拥有包括中草药、中医化学、中医疾病等 8,000 多个概念和 50,000 多个实例，基本上涵盖了中医药领域的大部分领域知识。

以上的这些中医药信息化工作，实现了海量中医药数据的整理、存储和共享，为利用知识发现技术，从这些海量数据中发现有用的知识，实现数据的有效利用，创造了有利的条件。

1.2.4 中医药特点迫切需要知识发现可靠性研究

虽然中医药知识发现在各个方面已经取得了一定的进展，但由于下面的一些原因，目前的中医药知识发现还存在一些问题，迫切需要中医药知识发现可靠性方面的研究。

第一，中医药领域的语言特点，要求知识发现必须关注可靠性问题。中医药学是极具领域特色的一门传统医学学科，在理论基础和临床实践中都有其独特性，中医学的理论具有高度的哲学性、文化性和语言性特点，临床实践具有个体性、主观性和交互性特点。中医药学数据的一个显著特点是自然语言性，其数据表达往往含义丰富，表达方式多样化、个性化。中医药数据的这些语言特性，使得在其上进行的分析和挖掘，需要更多的关注可靠性问题。

第二，当前中医药数据的数据质量，使得知识发现尤其需要考虑可靠性问题。从目前的情况来看，当前的中医药数据基础，虽然在量上已经有了一定的积累。但在数据质量上，还面临着较大的问题。由于中医药领域的文化传统、历史变迁和语言特点，数据质量问题已成为中医药研究中不得不经常面对的问题。在质量有所欠缺的数据上进行知识发现，尤其需要进行知识发现可靠性方面的研究。

第三，中医药去芜存精的发展趋势，对知识发现的可靠性提出了要求。作为一门具有几千年历史的传统医学，中医药与其他学科不同的一点，在于其内容精华与糟粕并存。利用知识发现技术研究中医药，其一大目的就是为了能发现一些有价值的规律和模式，并能有助于专家去除一些不科学的内容。这样去芜存精的过程，就需要知识发现本身是比较可靠的。

总之，对于中医药知识发现来讲，中医药的上述领域特点，促使我们必须对中医药知识发现可靠性这一主题加以充分关注和仔细研究。

1.3 本文的研究内容与主要贡献

本文围绕中医药知识发现可靠性这一主题展开研究，从知识发现整个生命周期的各个阶段对可靠性因素进行了总结和梳理，并据此建立了知识发现可靠性框架。针对中医药知识发现中比较突出的可靠性问题，重点探讨中医药知识发现中的结构性因素、表达性因素和信任性因素三大问题，其主要研究内容及其贡献如下：

1.3.1 提出了基于过程的知识发现可靠性框架

针对现有知识发现可靠性研究模型相关的特点，提出了一个与模型/应用无关的知识发现可靠性框架 PBRF-KD，该框架采用基于过程的思路对知识发现整个流程中的各个阶段和可靠性因素进行了梳理。该框架为知识发现项目设立了整套与可靠性相关的蓝本。

1.3.2 提出了结构相关的可靠性因素的优化方法

分析了中医药知识发现中与结构相关的可靠性因素，主要指数据完整性。针对文本型字段的完整性问题，提出了基于顺序半相关度量的中医药文本缺失字段填补方法。针对中医药文献类别标签缺失的问题，提出了基于 M-Similarity 的多标签文本分类方法。

1.3.3 提出了表达相关的可靠性因素的优化方法

分析了中医药知识发现中与表达相关的可靠性因素，包括表达粒度和表达一致性。针对表达粒度，提出了基于规则的表达粒度细分方法。

针对表达一致性，提出了基于本体的表达一致化方法。该套方法有助于提高中医药与表达相关的可靠性。

1.3.4 提出了信任相关的可靠性因素的优化方法

分析了中医药知识发现中与信任相关的可靠性因素，主要指数据可信度。针对中医药特有的数据可信度问题，提出了基于历史文献认可度的数据可信度衡量方法，和基于互联网知名度的数据可信度衡量方法。此外，基于这两种可信度衡量方法，提出了基于数据可信度的加权频繁模式挖掘算法。该套方法有助于提高中医药与信任相关的可靠性。

1.4 本文的组织结构

本文共分8章，其结构如下。

第1章 绪论，介绍知识发现可靠性的研究背景，中医药知识发现的研究背景，本文的主要研究内容和贡献。

第2章 中医药知识发现研究现状，分析中医药知识发现的数据基础，并从中医方剂知识发现、中药知识发现和中医证候知识发现三个方面，论述中医药知识发现的研究现状。

第3章 知识发现可靠性框架PBRF-KD，提出知识发现可靠性的过程视角；提出一般知识发现过程中的可靠性框架，并扩展到基于CRISP-DM的可靠性框架PBRF-KD；归纳并分析PBRF-KD框架中的7种可靠性相关因素。

第4章 结构性因素的分析与优化，分析中医药知识发现中与结构相关的可靠性因素，主要指数据完整性。针对文本性字段的完整性问题，提出基于顺序半相关度量的中医药文本缺失字段填补方法。针对中医药文献类别标签缺失的问题，提出基于M-Similarity的多标签文本分类方法。

第5章 表达性因素的分析与优化，分析中医药知识发现中与表达相关的可靠性因素，包括表达粒度和表达一致性。针对表达粒度，提出基于规则的表达粒度细分方法。针对表达一致性，提出基于本体的表达一致化方法。

第6章 信任性因素的分析与优化，分析中医药知识发现中与信任

相关的可靠性因素，主要指数据可信度。针对中医药特有的数据可信度问题，提出基于历史文献认可度的数据可信度衡量方法，和基于互联网知名度的数据可信度衡量方法。此外，基于这两种可信度衡量方法，提出基于数据可信度的加权频繁模式挖掘算法。

第 7 章　中医药知识发现系统 DartSpora，在回顾知识发现系统发展历史的基础上，介绍中医药知识发现原型系统 DartSpora，说明 DartSpora 的系统架构、虚拟组织模型、系统功能，并论述 DartSpora 中对于中医药知识发现可靠性的关注和体现。

第 8 章　总结与展望，总结全文并提出将来进一步工作。

第 2 章
中医药知识发现研究现状

本章首先分析中医药知识发现的数据基础，并从中医方剂知识发现、中药知识发现和中医证候知识发现三个方面，论述中医药知识发现的研究现状，最后概括了中医药知识发现的发展趋势，指出中医药知识发现迫切需要知识发现可靠性研究。本章介绍的这些工作成果构成本文进一步研究的基础。

2.1　中医药知识发现数据基础

由于知识发现是从海量数据中挖掘有意义模式的技术，因此，要进行中医药知识发现，其前提是具有一定的中医药数据基础。历年来中医药领域进行了多个国家科技项目研究基本数据库资源的建设问题。“九五”攀登预选项目“中药现代化关键问题的基础研究”中开展了“中药复方人工智能信息系统的研究”；国家 973 项目“方剂关键科学问题的基础研究”中也开展了有关方剂基础科学数据的信息处理研究。此外，全国各个中医药大学和科研机构也从不同的主题出发陆续建立了一些中医药数据库，如北京中医药大学建设了中医药古方剂数据库，福建中医药大学建设了台湾省中医药文献数据库和台湾药用植物资源数据库等。

当前规模和影响较大的中医药数据资源包括 CTCMPD(China TCM Patent Database，中国中医药专利数据库)[17]，TradiMed 数据库[18]，中医药化学数据库(TCM chemical database)[19]，以及 TCM-Online Database System[20]。

CTCMPD 是国家知识产权局“十五”信息技术重点应用性研究项目，

由知识产权出版社专利数据研发中心(PDC)和化学审查一部中药室合作完成,收录了1985年至今公开的全部中国中药专利。目前,该数据库收录的专利文献记录量已达19000余件,收录中药方剂近4万个。

TradiMed数据库是由韩国首尔国立大学Natural Product Research Institute建立的中医药数据库。基于中国和韩国的医学古籍,TradiMed对现代医学知识与传统医学进行了整合。到目前为止,TradiMed收录的信息包括3199种中药,11810首方剂,20012种中药的化学成分,以及4080种疾病。

中医药化学数据库由中国科学院过程工程研究所生化工程国家重点实验室研发。该数据库目前收录了从将近4000种中药中分离出来的9000种化学成分的详细信息,同时对于其中的很多化合物提供了详尽的生物活性数据。

TCM-Online Database System是迄今为止世界上规模最大的中医药数据库群。TCM-Online Database System的原型始于20世纪90年代末。1999起,国家科技部连续多年通过基础条件平台项目支持中医药科学数据的共建、共享和应用工作。在此背景下,中国中医科学院(原中国中医研究院)和浙江大学计算机学院CCNT实验室开始合作搭建中医药科技数据库群,并成功建立了集成全国17个分中心的分布式多库融合平台[15]。2002年开始,双方逐步将原有多库融合平台转变为语义网格平台DartGrid,提供动态的语义注册、语义查询等功能。通过全国30余家中医药学院、大学和科研院所近300名科技工作者的数据录入工作,该平台目前已集成了50余个数据库,100余G数据,其中包括中国中医药期刊文献数据库、中国中药数据库、疾病诊疗数据、中国方剂数据库、方剂现代应用数据库、中国中药化学成分数据库等数据库。目前,TCM-Online Database System可以通过网站[20]和CD-ROM两种方式进行访问。

本文知识发现研究的中医药数据基础即为中国中医科学院的TCM-Online Database System,其中主要用到的数据库为中国方剂数据库、中国中药数据库、中医药期刊文献数据库、中医药学语言系统。下面分别予以介绍。

2.1.1 中国方剂数据库

中国方剂数据库收录了来自710余种古籍及现代文献中的古今中药方剂84464首,分别介绍每一方剂的不同名称、处方来源、药物组成、功