国家自然科学基金项目·管理科学与工程系列丛书

基于异构社会网络的知识社区挖掘及学者相似度研究

刘萍 著

科学出版社

国家自然科学基金项目·管理科学与工程系列丛书

基于异构社会网络的知识社区挖掘及学者相似度研究

刘　萍　著

国家自然科学基金青年项目（71203164）

科学出版社
北京

内 容 简 介

本书围绕着异构社会网络中知识社区挖掘和学者相似度计算，在学科交叉的背景下，综合利用社会网络理论、社会资本理论、本体理论，提出：①基于社会资本理论的异构社会网络模型和知识社区发现方法；②基于关联网络链接分析的学者相似度计算方法；③基于本体的学者关联分析方法。全书以学者为研究对象，详细研究学者间多种学术关系的关联和融合，将单一节点、单一关系的学术网络扩展到多类型节点、多关系的异构社会网络，丰富和发展原有社会网络研究，对建立起更为广阔的语义社会网络研究范式具有一定的理论价值，同时也为科研组织创新团队管理提供决策支持。

本书可供管理类（如信息管理与信息系统、情报学、管理科学与工程等）、计算机类（如社会网络、语义网）专业或研究方向高校师生，以及各级科研管理和决策人员阅读参考。

图书在版编目（CIP）数据

基于异构社会网络的知识社区挖掘及学者相似度研究/刘萍著.—北京：科学出版社，2016

ISBN 978-7-03-048990-6

Ⅰ. ①基… Ⅱ. ①刘… Ⅲ. ①科学工作者-研究 Ⅳ. ①G316

中国版本图书馆 CIP 数据核字（2016）第 143517 号

责任编辑：徐 倩/责任校对：王 瑞

责任印制：徐晓晨/封面设计：蓝正设计

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京京华虎彩印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2016 年 6 月第 一 版 开本：720×1000 B5

2016 年 6 月第一次印刷 印张：11

字数：212 000

定价：68.00 元

（如有印装质量问题，我社负责调换）

前　言

在21世纪知识经济时代，知识创新是促进知识经济与社会可持续发展的基础和源泉，是推动科技进步和经济增长的革命性力量，也是提高国家综合国力和国际竞争力的强大保障。中国制定的《国家中长期科学和技术发展规划纲要（2006—2020年）》和《中华人民共和国国民经济和社会发展第十一个五年规划纲要》明确提出加快科学技术创新和跨越，大力推进自主创新，加快建设国家创新体系。

在从“中国制造”到“中国创造”的战略转型过程中，科研组织扮演着极为重要的角色。科研组织中具有共同兴趣、经验、目的和研究背景的科研人员聚集在一起进行沟通交流形成的知识社区，能满足科研人员在其科研活动中进行学习、开放、交流、分享、团队合作及创新研究的需求。知识社区有助于集聚智慧、启迪思路，提升科研人员知识共享、协同与创新的效率和效果。对科研组织内的知识社区识别和挖掘已成为知识管理研究的新热点。

传统的知识管理强调信息技术（如搜索引擎、数据仓库、文献管理等），目标在于快速准确地将有用的知识传递给需要的人。然而第1代知识管理也暴露出明显的缺陷，那就是知识共享的不足。第2代知识管理更多地考虑人力资源和过程的主动性，注重营造知识共享环境，强调知识社区对知识共享的催化作用，反映了知识管理与组织社区学习相融合的趋势。识别组织中拥有共同需求和兴趣的人员，挖掘潜在的知识社区对于组织知识管理的理论和实践具有重要的意义。

知识社区的本质是社会网络，因此社会网络的理论和分析方法也适用于知识社区及相应的知识管理研究。近年来，社会网络分析与知识管理相结合的研究受到广泛的关注和重视，第17届知识工程与知识管理国际会议（International Conference on Knowledge Engineering and Knowledge Management，EKAW）、第10届知识与系统科学国际研讨会（International Symposium on Knowledge and Systems Sciences，KSS）和第8届知识管理国际会议（International Conference on Knowledge Management，ICKM）都把社会网络分析与知识管理的交叉研究作为焦点议题。社会资本理论是社会网络三大理论之一，从认知、关系、结构三个层面反映组织成员间多种复杂关系。社会资本蕴涵于社会网络之中，个人只有成为

网络成员或建立网络联系，才能接触和利用社会网络资源，获取收益。对于知识社区的成员而言，这种收益就是共享知识及创新思想，在知识交流和共享的过程中不断形成及拥有创新思想的成员会逐渐成长为创新人才。

本书基于第 2 代知识管理以人为本的思想，营造服务于知识创新的知识共享环境，从异构社会网络（heterogeneous social networks）的视角进行“知识社区的挖掘和学者相似度研究”，本书的研究内容主要由三部分组成。

第一部分是知识社区的发现（第 2 ~ 5 章）。首先是对社区发现相关理论的概述，介绍社会网络理论、行动者网络理论、社会资本理论等相关理论。接着总结现有社区挖掘基础方法和最新研究进展。在此基础上提出基于社会资本理论的异构社会网络模型，并融合网络拓扑结构和节点属性进行知识社区发现及优化。

第二部分是学者领域知识结构挖掘（第 6 章）。第一部分社区挖掘是从网络中观角度进行的，然而即使是同一个社区的学者，个人研究专长也会有所不同。第二部分从网络的微观角度揭示每个学者的领域知识结构。本书提出学者领域知识网络模型，详细分析基于 LDA（latent Dirichlet allocation，即潜在狄利克雷分配）模型的学者研究主题挖掘方法，并以主题为节点，以主题间关系为边，构建知识网络，最终将关系紧密的主题划分为社团来揭示学者领域知识结构。

第三部分是学者相似度计算（第 7 ~ 9 章）。无论是潜在知识社区的发现或是个体领域知识结构的挖掘都是为了促进知识的交流和科研合作。探寻学者之间的关联度不仅能显化学者之间的关系，也能辅助科研合作推荐。本书重点阐述基于关联网络的学者相似度计算方法，以学者关键词共现网络为基础，以 SimRank 思想（即被相似实体指向的两个实体是相似的）为指导，充分挖掘网络中节点链接关系来计算学者间的相似度，实验结果证实基于 SimRank 的学者相似度计算能较好分析学者研究内容，有效提高学者间研究内容相似性的深度和准确性。本书第 9 章将本体理论与社会网络相结合，通过建立学术网络的本体概念模型，对异质学术网络的节点和关系进行抽象。通过定义本体推理规则，发现隐含的多维学者关系，构建加权学者网络。提出每一维度不同关系强度的计算方法，并按各关系的相对重要程度进行权重赋值，最终得到基于多种关系的学者相似度。

结合国内外研究，笔者认为本书的研究内容主要有以下三点创新之处：第一，针对单一的网络关系结构不符合现实世界中社会关系的复杂性、多样性特征，本书提出以社会资本理论为基础，从结构、认知、关系三方面考虑组织成员之间的关系，构建基于社会资本理论的多关系融合异质网络，具有首创性。第二，在社区发现和优化方面，充分利用节点属性信息中所包含的网络背景信息，提出基于词汇语义相似度的节点属性关联的计算方法，在基于模块度（modularity）优化的社区划分基础上，设计基于语义贴近度的信息熵方法来优化社区发现的结果。第三，将本体理论与社会网络理论相结合，补充异构社会网络中节点与关系的语

义信息，创新性地提出基于本体推理的学者关联分析方法。

本书的研究得到国家自然科学基金项目（71203164）的资助，在此表示衷心的感谢。

异构社会网络的知识社区挖掘和学者相似度研究是一项艰巨而又复杂的工程。本书从社会资本的角度来挖掘和测度学者间多种关联，在揭示隐性知识社区和推荐相似学者方面进行一些探索，还有很多工作有待深入研究，恳请同行批评指正。

刘　萍

2016年2月于武汉大学

目　录

第1章 绪 论

1.1 研究背景与目的

知识创新是通过科学研究获得新的基础科学和技术科学知识的过程，目的在于追求新发现、探索新规律、创立新学说、创造新方法和积累新知识。科研组织尤其是科研院所是国家创新体系建设的重要组成部分和创新主体之一，担负着基础性、战略性和前瞻性的研究工作。科研组织作为一种知识密集型组织，其生命力和竞争力在于不断创新并产生新知识。科研组织中拥有相同研究兴趣的科研人员自发地聚集在一起形成社区，促进了组织成员学习、开放、交流、分享、团队合作及创新研究，为头脑风暴等集体创新行为提供了方便的场所。

随着知识管理研究的不断深入，知识社区的概念逐渐吸引了国内外越来越多的学者关注，成为知识管理的一种重要的方法。知识社区是组织成员之间由于合作、交流、共享等形成的相对稳定的团体，知识社区可以是组织中的正式团体，如项目团队；也可以是非正式的团体，如有着共同兴趣爱好且经常相互交流的组织成员组成的团体。知识社区在知识的创造和传播过程中发挥着重要作用。知识社区的成员围绕相同研究兴趣进行交流，各种观点相互碰撞可以激发创造性，产生更多的新观点以及问题的解决方案，并在对别人的设想与方案的思考和质疑过程中发现现实可行的方法。知识社区有利于融汇集体智慧，促进创新思想的形成以及发现和培养创新人才。对知识社区的识别和挖掘也成为知识管理研究的新热点。

知识社区是社会网络的一种社区结构，因此社会网络分析方法也适用于知识社区挖掘研究。近十多年来不同领域的研究人员提出了很多社区挖掘方法，分别采用了来自物理学、数学、计算机科学等领域的理论和技术。尽管社区挖掘已取得了许多令人鼓舞的成果，但该问题还未被很好地解决。主要原因如下：①当前关于社区挖掘的绝大多数方法都假定社会网络中只存在一种关系，也就是说绝大多数被研究的社会网络都是同质网络。而现实是社会网络的节点间存在多种复杂的关系，如学术网络中学者间存在的合著关系、共事关系、引用关系等。采用传统方法挖掘的结果并不完全符合用户的真实需求。揭示多关系社会网络中的社区

结构仍是一个挑战。②绝大多数社区挖掘技术关注的是网络拓扑结构，而社区的识别基于链接或结构的相似度，在聚类过程中没有考虑节点的属性信息，忽略了节点间的内在联系。

针对目前社区发现研究的不足，本书将社会网络理论和分析方法引入知识管理的研究，从社会资本的角度识别和挖掘知识社区。研究针对科研组织成员的多关系社会网络模型，探索结合网络结构和内容的社区发现方法，识别科研人员知识领域，分析计算科研人员相似性，推动科研人员的知识交流与共享。

1.2　国内外研究现状

社会网络理论的理论基础被认为是 Stanley Milgram 提出的“六度分割理论”（six degrees of separation），Stanley Milgram 通过一个连锁通信实验证明了“六度分割理论”，理论指出，最多通过六个人就能够认识任何一个陌生人，该理论也被称为“小世界理论”。1998 年，Watts 和 Strogatz（1998）提出基于人类社会网络的网络模型，通过调节一个参数它就可以从规则网络向随机网络过渡，该模型被称为 WS 小世界模型；1999 年，Barabasi 和 Albert（1999）提出了复杂网络的无标度性并建立了无标度网络模型。至此，社会网络形成了初步的理论研究基础，应用范围越来越广泛，吸引了大量不同学科、不同背景的研究者和实践者。本书从社会网络中社会资本的角度挖掘潜在的知识社区，涉及三个方面的研究：一是社会网络模型；二是知识社区；三是社区发现。以下简要分析有关方面的研究进展。

1.2.1　社会网络模型

社会网络模型通常用来描述网络社会结构和社会联系。在社会网络模型研究中，通常将社会网络表示为一个图——$G=(V, E)$，其中，V 代表社会行动者，E 代表行动者之间的关系。社会网络模型构建的重点在于行动者之间关系的抽取。社会行动者之间可能存在多种关系，如血缘关系、朋友关系，或是人与机构的隶属关系、机构与机构的隶属关系等，如何在复杂多样的关系中抽取提炼出有价值的关系信息是一项至关重要的工作。针对这个问题，研究人员从体现行动者之间关联的数据库中抽取关系作为社会网络的边，构建网络。这些关系包括邮件关系、合作关系、交流关系等。代表性的研究包括：彭玲（2010）通过搜集分析邮件数据集，获取邮件的收发关系，以及相应的邮件主题、内容、收发时间等信息，在此基础上计算两

个人之间的联系频率，以人为节点，以联系频率为边的权重，构建了邮件网络，并通过改进的社区挖掘算法得到邮件网络的社区结构；王福生等（2009）通过提取合著论文作者的合作关系，建立科研合作网络模型，验证了网络节点的度分布；肖连杰（2010）依据合著论文确定节点间边的关系，并结合合著论文的篇数与学术价值，即学术影响因子、期刊级别两个方面共同计算边的权值；杨洪勇和张嗣瀛（2008）通过统计期刊的学术论文，建立科研合作网络的演化模型，通过网络节点度的分析，研究作者间合作方式及团队发展模式；王辉等（2011）从典型的大规模 Web 社会网络 DBLP（Computer Science Bibliography，即计算机科学文献库）中抽取数据源，构建大型科研合作网络；袁毅和杨成明（2011）跟踪微博用户在时间周期内关于特定话题的交流数据，依据用户在信息交流过程中形成关注、评论、转发和引用四种行为的频率，分别探究微博用户信息交流过程中所形成的社会网络；孟徽和邓心安（2009）通过采集图书情报学（library and information science，LIS）领域核心期刊的论文数据，建立关于情报学的社会网络，并通过对节点度、介数和接近度的综合比较，分析了作者的影响力。

上述的关系抽取都是针对网络中的单一关系进行的，根据单一关系构建的网络是一种单关系同质网络，这种网络中，节点之间只存在一种单一关系。但随着社会网络研究的不断发展，研究者发现单一网络关系结构已经不足以解决现实中的复杂问题，因此，异构社会网络应运而生。在社会网络研究中引入异质关系的概念，反映了多种关系的共同存在，使分析更具有个性化，满足不同角度观察者的需求。研究者通过选取能够表征实体间某种特种的多种关系，构建存在多种关系的异构社会网络。代表性的研究有：张福增等（2007）通过提取科研人员之间的合作关系、引文关系及讨论等情景建立了一个不同权重的有向科研影响关系网络，根据该网络的关系矩阵给出了衡量科研人员影响力的指标和计算方法；Mucha 等（2010）通过扩展模块度研究异质关系网络的多尺度社区结构，并将之用于分析美国某大学在校学生因四种社会关系构成的异质关系网络；Szell 等（2010）以一个大型多用户在线网络游戏的用户社会关系网络为对象，研究了异质关系网络中多种关系并存对网络结构（包括社区结构）的影响；张伟哲等（2012）构建了以论坛中主题及其回复为关系的异构网络模型，并提出了一种基于异质网络的意见领袖社区发现算法来挖掘论坛中的意见领袖社区。在上述的研究中，研究人员虽然考虑了通过抽取多种关系来构建异构社会网络，但是并没有考虑到这些关系之间重要程度的差异性。对于研究对象的某个需要表征的特征，多种关系体现出的重要程度不同。如何根据研究对象的特征和实验目的来确定每一种关系的重要性程度，是研究工作面临的一个挑战。本书将尝试通过社会资本理论来抽取三种能够表征研究人员的研究兴趣相似度的关系，并通过机器学习的方法来确定每一种关系所对应的权重系数，以此来构建科研组织中基于研究兴趣的由多种关系融

合而成的异构社会网络。

1.2.2 知识社区

知识社区的概念是在知识管理的基础上提出来的。随着知识管理研究的逐渐深入，知识社区作为知识管理的一种重要的方法，已经被越来越多的研究者关注。知识社区是组织中以知识交流和知识共享为目的，组织成员自发或半自发由于合作、交流、共享等形成的相对稳定的团体。陈永隆和庄宜昌（2003）认为知识社区是“透过网络社群的互动与分众特色，辅以实务社群的搭配运作，建立以专业技术与知识领域为主的讨论区、专栏区、留言版、聊天室、读书会、研讨会等，让企业内部的知识工作者能够经由选择特定的专业领域，与其他具有相同专业领域或对该专业领域有兴趣的跨部门员工，进行互动并创造知识、分享知识的平台”。知识社区分为两种形式，即实体知识社区和虚拟知识社区。

实体知识社区是指组织中一群具有特殊专长或工作的群体成员，为了促进彼此间的知识交流与共享，使工作更加高效率，在广泛的交流学习和互相帮助过程中，形成了有着共同的兴趣或目标，以及分享或研究与工作相关的知识和经验的共同愿望，由此所形成的一种特殊的建立在工作与实践基础之上的组织或团体（任曼，2011）。知识社区可以是组织中的正式团体，如项目团队；也可以是非正式的团体，如有着共同兴趣爱好且经常相互交流的组织成员组成的团体。在知识社区中，组织成员通过交流和学习、分享的活动，可以不断地促进显性知识和隐性知识的转化与创新，进而实现组织的知识共享与知识创新。

随着信息技术的不断发展，互联网开始渗透到人们生活的每个环节中，人们迫切地需要一种更加便捷化、高效率的知识管理方式，于是，知识交流开始逐步转入互联网的平台，互联网成为知识交流的全新趋势，因此，基于互联网的虚拟知识社区应运而生。秦鸿（2007）认为，虚拟的知识社区是现实社区的网络缩影，它是通过现代信息技术的支持，以知识创造和传播为目标的、现实与虚拟载体相结合的一种有着空前灵活性和创造力的新型社区。

目前的研究中，对知识社区的研究主要集中在这些方面：①数字图书馆的知识社区关系及构建的研究。2004 年，日本筑波大学知识社区研究中心召开了以“网络化信息社会中数字图书馆与知识社区”为题的研讨会，对数字图书馆和知识社区的关系进行了重点分析，会议认为，数字图书馆是网络基础设施的重要组成部分，知识社区是更高层次的信息组织和交流模式。秦鸿（2007）分析数字图书馆和知识社区的关系，并探讨了知识社区的分类及知识组织工具。王利萍等（2007）通过对图书馆 2.0 服务模式的分析，研究了图书馆 2.0 开放式

网络知识社区的构建原则及其功能的实现。陈红勤和曹小莉（2011）将图书馆网络社区的知识传播机制划分为知识建构、知识转移和共享、知识创新、知识服务机制四种，进而探讨了知识建构系统框架、知识转移过程、知识创新的通用模型和途径，以及知识服务的构成要素和过程。陈廉芳（2012）分析了图书馆知识社区联盟构建的必要性和可行性，并探讨了图书馆知识社区联盟的数字资源整合、联合参考咨询和用户教育等功能模块的实现。曹志辉（2008）分析了数字图书馆与知识社区的关系并论述了数字图书馆知识社区的构建方式。②基于工作与实践社区研究。Lave 和 Wenger（1991）提出实践社区（communities of practice，CoP）的概念，将其定义为“关注某一个主题，并对这一主题都怀有热情的一群人，他们通过持续地互相沟通和交流增加自己在此领域的知识和技能”。实践社区是指成员间的那种非正式的工作联系性群体，从知识管理的角度出发，实践社区常被称为知识社区。目前，实践社区的研究还主要集中于国外的研究，主要讨论了实践社区的定义、类型、构建原则、社区成功及失败的影响因素等。Bielaczyc 和 Collins（1999）认为实践社区是知识管理的重要工具，可以提供讨论的场合，使知识可以通过讨论扩散。Wenger 等（2002）提出了设计实践社区的七条基本原则。石文典等（2008）针对人性特征研究实践社区知识传播的影响作用。③基于知识社区的 E-learning 模式的研究。王知津和谢瑶（2008a，2008b）论述了 E-learning 及知识社区的概念与特征，提出了基于知识社区的 E-learning 模式的构建，并以美国应用材料公司（Applied Materials Inc.，AM）成功实施基于知识社区的 E-learning 为案例，从战略、技术、流程和人力资源四个管理层面来分析其基于知识社区的 E-learning 的构建过程，并探讨了企业实施基于知识社区的 E-learning 的关键成功因素。④知识社区在企业中的应用研究。李毅心和任南（2007）针对我国中心型企业如何建立实体知识社区和虚拟知识社区进行研究，他们认为知识社区的建立直接支持知识获取和知识共享，有助于员工间隐性知识的共享，为知识创新提供了基础。谈涟亮（2003）认为企业知识社区的构建是多层次、全方位的，从上至下、从里向外涉及了企业的各个方面，包括员工的意识和工作的方式。⑤知识社区对教育推新及对信息社会影响的研究。李文娟和王宇辉（2008）对远程教育中知识社区的建设问题进行了分析，并提出远程教育中知识社区建设的重点在于规范显性知识，挖掘隐性知识，促进交流共享，同时也提出了要合理应用技术、开展制度建设、构建网络平台等建设意见。黄禧凤（2012）结合高校英语文化知识社区构建的研究背景，阐述了英语文化学习中引入知识协同理论的优势，研究了基于知识协同的英语文化知识社区的功能模块和服务流程。

1.2.3 社区发现

社区发现是社会网络研究的重要内容之一，并直接关系到网络系统中的中观度量与对应的共性规律，是一个基础问题，在过去十多年内吸引了很多学者的关注。目前国内外关于社区发现的研究比较多，取得了一些重要的进展，本书将从社区的定义及社区发现的方法两方面来总结国内外的社区发现研究现状。当前主要的社区发现算法有图划分方法、子图聚合算法和基于优化的方法。第一，图划分的方法一般是自顶向下把图分成不相连的子图，如 Flake 等（2002）提出的最大流/最小割方法；Girvan 与 Newman（2002）提出的基于边介数的社区发现的 GN 算法；Tyler 等（2003）将统计方法引入基本的 GN 算法中提出的近似 GN 算法；Radicchi 等（2004）提出了连接聚类系数（link clustering coefficient），取代 GN 算法的边介数；Kim（2007）提出基于边的谱分解算法。第二，基于子图聚合的方法，是基于一定的子图距离度量方法，自底向上合并子图，生成社区结构。例如，Newman（2004a）提出的模块度，以及基于模块度的聚合算法就是属于一种平均连接聚合方法。第三，基于优化的方法是指基于局部搜索的优化策略，主要有 Kernighan-Lin 算法（KL 算法）（Newman and Girvan，2004）、快速 Newman 算法[①]（Newman，2004a）和 Guimera-Amaral 算法（GA 算法）（Guimera and Amaral，2005）。另外，现有的社团结构分析算法大多数将网络划分为若干相互分离的社团，无法对彼此重叠、互相关联的社团结构进行分析，针对这一问题，赵鹏等（2008）根据交联网络的结构特点，提出了交联网络中的可重叠社团结构分析算法（algorithm to analyze overlapping community structure of intersection networks）。朱大勇等（2009）以相异性指数作为网络节点的距离度量，结合模块度提出用遗传聚类来分析和发现网络社团结构。这些算法为社会网络中社区结构的发现提供了技术支持。

由于各类社会网络的节点和边的意义不尽相同，而且目前对社区的定义还没有一种统一的认识，不同研究领域的研究者由于研究对象和目的的不同，对社区划分结果的期望值不同，因此，也没有一个统一的评价标准来对上述算法的好坏以及社区划分结果的优劣进行评价。目前，对社区划分结果的评价主要是基于网络的拓扑结构，通过计算网络的模块度、密度、凝聚力等结构指标来判断。这种评价的方法将任何类型的网络都看做网络图，从图论的角度出发来判断划分结果的优劣。但这种做法仅仅考虑了网络的结构信息，却忽略了网络自身所拥有的内容信息。因此，研究者认识到使用语义结构模型可以很好地表达社会网络中的语

① 即社区发现 FN 算法，fast algorithm for detecting community structure in networks。

义信息，他们将网络的内容信息与网络结构特征结合起来，针对具体的问题对社区划分的结果进行优化。利用已有的语义模型对社区进行建模或者将语义信息融入传统的社会网络模型中，使用语义网资源描述框架（resource description framework，RDF）和网络本体语言（Web ontology language，OWL）等工具可以对本体与网络资源进行语义描述。目前互联网上最流行的社会网络本体应用是FOAF（friend-of-a-friend）（FOAF，2015），其提供了RDF/XML（extensible markup language，即可扩展标记语言）字典来描述个人信息。在人际关系基础上形成信任网络，可以进行知识的发布和共享（Ding et al.，2003），从而形成Web社区（Lawrence and Schraefel，2006）；J. Tang等（2008）构建了一个ArnetMiner系统，用扩展的FOAF本体来标注社会网络；Jung和Euzenat（2007）构建了三层的语义社会网络分析框架，包括社会网络层、本体网络层和概念网络层，社会网络成员间关系的衡量采用基于概念层的概念相似度和本体层的本体相似度。

但语义模型大多都用于对社区进行发现和挖掘的过程，不太适用于社区优化尤其是进行社区分割、合并和用户动态迁移的场景。现实可行的办法是在网络中为节点加上具有语义信息的属性。节点属性信息包含了节点在网络中的背景信息，能够反映成员感兴趣的内容。这些丰富的非拓扑信息使社会网络分析不再局限于网络拓扑结构层面，而是从不同的语义层面展开。Cantador和Castells（2006）通过对概念及社会网络成员的聚类得到一种多层的语义社会网络。Cruz等（2011a，2011b）在社区发现的过程中将语义信息与社会网络的拓扑结构相结合，提出基于整体模块度最优的社区挖掘算法。曹源（2008）通过分析个人文献提取每个科研人员的关键词及权重形成用户兴趣向量，构建用户兴趣模型，计算出用户研究兴趣的相似度并以此为边的权重，构建基于用户兴趣的科研社会网络。Dang和Viennet（2012）在构建的社会网络模型中为每个节点添加了属性向量，并在此基础上提出了两种基于节点属性与网络拓扑结构分析的社区发现方法。Steinhaeuser和Chawla（2010）综合考虑节点属性与网络结构特征，提出了基于节点相似性的边的赋权方法（node attribute similarity，NAS），并以此为基础提出基于随机游动的社区发现方法。Zhou等（2009）提出了将网络结构与节点属性的相似性相结合计算距离的方法，平衡了图中结构与各点属性的关系，将属性转化为一种附加的结构，达到属性与结构的统一。

1.2.4 相关工作小结

综上所述，国内外学者从不同角度对社会网络进行了持续不断的探索，取得了丰硕的成果，但同时也存在一些不足。首先，国内外学者所研究的社会网络绝

大多数是单一关系的同质网络，不能很好地刻画现实网络中多种关系相互交织的现象，因而异构社会网络将成为未来研究的重要趋势。其次，目前社会网络中社区发现模型多数依赖节点关系，忽略了节点属性所附带的语义特征，因此有必要结合节点属性和网络结构对社区发现进行定量与实证研究。本书在借鉴和吸收国内外已有相关研究成果的基础上，提出利用社会资本理论选取科研组织中基于研究兴趣的多种关系并构建社会网络模型，增加节点属性来丰富社会网络的语义内容，结合结构和内容来优化社区划分。将社会资本、知识组织与社会网络有机结合进行研究，既具有理论创新性，也具有实践应用性。

1.3　研究的主要内容与方法

本书探讨基于异构社会网络的知识社区挖掘与学者相似度研究，主要包括以下五个方面的研究内容。

（1）基于社会资本的网络模型研究。组织的社会网络模型是对成员集合及其关系的一种抽象表示。现有的社会网络模型通常是基于图论的模型，将网络建模为一个图 $G=(V, E)$，V 代表网络的节点集合，E 表示网络中边的集合，该模型不能很好地反映科研组织中人员的多种关联，并且忽略了人员自身的属性特征。本书提出多关系社会网络模型，主要研究多关系的选择、关联强度的计算及节点属性的表示与相似度计算。其中对多关系的选择，主要是基于社会资本理论。作为社会网络的三大理论之一，社会资本是个人或群体的人际关系网络，由认知维度、关系维度和结构维度组成。在科研组织中，认知维度的关联体现为成员之间具有相同的研究兴趣，关系维度的关联体现为成员之间的合作经历，结构维度的关联体现在科研组织的层级结构上。

（2）知识社区发现及优化。社区结构是社会网络的重要特性，代表了社会网络中具有相同兴趣或偏好的团体。社区发现旨在挖掘网络中的一些关系比较紧密而又具有相似兴趣的子团体。目前针对社区发现的研究主要集中在内容分析及聚类和结构分析两个方向。本书将采用结构分析和内容分析相结合的社区发现方法。首先通过加权 WGN 算法（weighted GN algorithm）获得社区划分的初步结果，其次利用节点属性的相似度对社区划分结果进行优化。

（3）学者领域知识结构挖掘。从网络的微观角度揭示每个学者的领域知识结构挖掘。本书提出基于知识网络的学者领域知识结构表示，主要研究学者领域知识的构成、学者领域知识网络模型及构建方法，以及知识网络的划分，以此来揭示学者领域知识结构。

（4）学者相似度计算。学者相似度计算是学科知识结构探测、相关学者推荐、链接预测的基础研究问题。现有的学者相似度计算都是通过属性间的某种直接关联来计算学者间的相似度，忽略了属性间的间接关联。本书提出一种新的基于关联网络的学者相似度计算方法，以学者关键词关联网络为基础，采用SimRank算法充分挖掘网络中节点链接关系来计算学者间的相似度，实验结果证实基于SimRank的学者相似度计算能较好分析学者研究内容，有效提高学者间研究内容相似性的深度和准确性。

（5）语义社会网络建模及学者关联分析。目前社会网络分析主要是对行为者之间的社会关系进行定量化分析，侧重的是网络结构的分析，而忽略了节点与节点间关系的含义。本书将本体理论与社会网络相结合，主要研究针对学术网络的本体构建、基于规则的学者间关系推理与发现，以及学者间多维关系的测度。最后用一个计算语言学领域学术网络对语义社会网络及学者关联分析方法的有效性进行验证。

本书在研究和写作过程中主要用到文献调研法、多学科交叉法、建模分析法及实证研究法四种：

（1）文献调研法。广泛搜集国内外相关文献资料，把握国内外在社区探测、社区划分、个体知识结构挖掘、学者关联分析等领域的前沿动态，修正与完善本书的研究思路和内容，细化工作思路和研究内容。

（2）多学科交叉法。知识社区探测与挖掘是一个跨学科、跨领域的前沿课题，本书将综合使用社会网络分析方法、网络聚类方法、图划分、社区探测等方法，对实验进行分析与评价。

（3）建模分析法。通过数学模型和形式化模型的建立与分析，探求面向科研型组织的异构社会网络构建，在此基础上设计基于网络整体模块度最优的知识社区划分优化方法，对提出的方法进行验证，检验模型的可靠性。

（4）实证研究法。在相关理论研究的基础上，以典型科研型组织——武汉大学信息管理学院为例，对所提出的异构社会网络模型、社区探测及学者相似度计算方法进行实验、分析、评价和优化。