科研合作网络的链路预测研究

张 斌/著

科学出版社

本书研究获
国家自然科学基金重点国际（地区）合作研究项目（71420107026）
国家自然科学基金青年项目（71704138）
中国博士后科学基金特别资助项目（2017T100585）
中央高校基本科研业务费专项资金
资助

科研合作网络的链路预测研究

张　斌/著

科学出版社
北京

内 容 简 介

科研合作网络是图书情报学领域的一个重要研究对象。本书介绍了科研合作网络的链路预测的基本理论和方法；借助中文社会科学引文索引（CSSCI）数据库，从中选取七门学科并构建合作网络，分析其网络结构，概括出我国人文社会科学领域的合作特点；针对所构建的合作网络，分别研究了静态链路预测和动态链路预测的预测效果；结合基础理论和实证结果，讨论了网络结构和链路预测的关系、网络演化与链路预测的关系等问题。

本书适合于从事信息管理、图书情报及相关领域的理论和实践工作者阅读参考，也可以作为高等学校相关专业硕士研究生和博士研究生的教学参考书。

图书在版编目（CIP）数据

科研合作网络的链路预测研究 / 张斌著 . — 北京：科学出版社，2018.6

ISBN 978-7-03-057852-5

Ⅰ . ①科… Ⅱ . ①张… Ⅲ . ①图书情报工作—研究 Ⅳ . ① G250

中国版本图书馆 CIP 数据核字 (2018) 第 129651 号

责任编辑：林 剑 / 责任校对：彭 涛

责任印制：张 伟 / 封面设计：无极书装

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

北京九州迅驰传媒文化有限公司 印刷

科学出版社发行 各地新华书店经销

*

2018 年 6 月第 一 版 开本：B5（720 × 1000）

2018 年 6 月第一次印刷 印张：11

字数：250 000

定价：108.00 元

（如有印装质量问题，我社负责调换）

序

科研合作是以科学研究为目的而形成的合作关系。20 世纪以来，科研合作在各个领域中日益频繁，甚至表现出了跨地域、跨机构、跨学科合作程度逐渐增强的趋势，亦成为解决重大科学问题的主要途径。对于科研合作规律和机制的研究，是科学学、科研管理和图书情报学领域长期以来持续关注的一个重要问题。

在情报学理论研究中，科研合作网络兼具社会网络和知识网络的双重属性，其网络演化的内在驱动因素，包括了人际交往、知识交流等机制，是非常重要的研究对象。科研合作网络是较早得到关注的一类复杂网络。很多学者都对其进行了深入研究，通过捕捉网络的静态结构和动态演化的特征，发现了复杂网络的很多新性质。随着近十几年网络科学的发展，各种聚类技术，如模块度、派系过滤算法、连边社团检测算法等被用于识别科研合作网络的社团结构，推动着科研合作分析走向更为精细的层次，并拓展到交叉学科合作趋势、研发群体科学等新的问题上。

与此同时，在一个网络中，如何通过已经观察到的节点之间的连接，来重现因为数据缺失而尚未观察到的连接，或者预测未来将要出现的连接，即复杂网络的链路预测问题，受到了不同学科领域越来越多的关注。链路预测的理论和方法在逐步建立和完善中，它有助于重新认识复杂网络的结构和功能，并在网络重构、网络演化模型评价、推荐系统等方面发挥着重要作用。链路预测作为一种新兴的兼具理论研究和实践应用的技术手段，理应受到图书情报学界的关注，并应用到知识网络，特别是科研合作网络的研究中。

知识网络的模型、结构与动力学机制等，是我以及武汉大学信息资源研究中心长期关注的课题。我们在 2012 年承担了国家自然科学基金面上项目“知识网络的形成机制及演化规律研究”，以此为依托，在 2015 年以“大数据环境下的知识组织与服务创新研究”为题申报国家自然科学基金重点国际（地区）合作研究项目并获批准。通过近六年的深入研究，这些项目已取得了很好的成果。张斌博士在读硕士和博士期间参与了这些项目的研究工作，取得了一些有益的成果，发表了多篇学术论文，最后以“科研合作网络的链路预测研究”为题完成了博士论文。

科研合作网络的链路预测研究是一个交叉学科研究方向，涉及计算机、社会学、统计物理、复杂性科学、图书情报学等多个学科领域。该书作者综合运用信

息计量学、统计学、社会网络分析、链路预测、数据挖掘与可视化等理论与方法，就科研合作网络的结构链路预测效果、网络结构与链路预测的关系、网络演化与链路预测的关系等四个方面进行了探索性研究。综合来看，该书在三个方面取得了进展并具有一定特色：

一是构建了图书情报学视角下科研合作网络的链路预测的基本理论框架。该书通过不同链路预测指标和算法对不同学科合作网络进行静态和动态链路预测实验，并对预测效果进行比较研究，得出了对实际预测有一定指导作用的理论结论。在构建网络演化模型前，可以通过结构相似性指标来观察在合作网络中的链路预测效果，从而分析关于影响连接建立的主要特征；在构建网络演化模型后，可以利用链路预测理论设计量化评价方法，来评价所构建的模型及其演化机制的优劣。

二是通过对科研合作网络的网络结构和链路预测的研究，得出了一些基本规律，对选择合适的预测方法有一定参考价值。本书总结了不同类型网络的结构差异，将网络结构的改变程度与链路预测效果的变化情况进行关联分析，揭示预测算法性能和网络结构特征之间的关系。研究验证了当聚集系数很大时，最近邻的局部预测指标的预测效果会很好；随着聚集系数增大，稀疏网络中有叠加效应的局部随机游走指标的预测效果会比较好；稠密网络中 Adamic-Adar（AA）和资源分配指标的预测效果会比较好；网络结构的扰动对于 Katz 指标的预测结果影响最小。

三是将链路预测应用在我国人文社会科学合作网络上，有助于探索我国人文社会科学合作的一般规律和机制。研究发现，不管是对整体学科合作网络的静态挖掘，还是对核心作者合作网络的动态分析，合作关系的建立并不倾向于度择优机制，这意味着在构建演化模型时，应充分考虑科研合作网络兼具社会网络和知识网络的双重特点，来设计择优机制。

科研合作网络的链路预测毕竟是一个复杂的研究课题，而且科研合作网络的类型多样、结构复杂、动态变化，这都加大了研究难度，该书中的一些成果尚留有较大改进的空间。但我们也应该以宽容的心态看待一个年轻学者在学术道路上的成长过程，多给一些鼓励和支持，给他更多的时间来改进和提升。

张斌的本科、硕士和博士都就读于武汉大学，我更作为他硕士和博士期间的指导教师，和他交流有年。他勤奋好学、思想敏锐，具有探索精神，能够利用新知识、新方法和新工具来解决实际问题。该书是在他的博士论文基础上完成的。在本书付梓之际，他嘱我作序，写下上述感想，谨表祝贺。

2018 年 4 月于珞珈山

前　言

科研合作网络是由科研合作关系所形成的网络，是一种典型的社会网络，其最基本的节点就是参与科研合作的人。同时，科研合作网络还是一种典型的知识网络，是知识交流的产物，是一种建立在社会关系上的交流、转移、共享知识资源，特别是隐性知识后形成的网络结构。科学家在一篇论文上进行共同署名，可以看作完成了一次合作，这是显性表征。虽然合著论文并不代表科研合作的全貌，但却是行之有效的测量手段。本书讨论的科研合作网络具体指代作者合著网络。科研合作网络是较早得到关注的一类复杂网络，很多复杂网络新性质的发现都应归功于对科研合作网络的研究。在大数据时代，面对海量的科技文献信息，借助网络科学理论与方法，图书情报学界对于这一问题又重新有了巨大兴趣。

研究科研合作网络的结构和演化，可以用来分析学者之间的相互影响、合作的演化机理、知识传播和发展脉络等问题，这为我们观察到的科研合作和知识交流现象提供了解释，同时也为科研合作的演化动态和发展趋势奠定了预测的基础。然而，刻画网络结构特征的统计量非常多，不同的演化模型和机制也有各自的适用性，很难用一个统一尺度来判断哪个更加优秀。近些年，复杂网络中的链路预测方法受到越来越多的关注，链路预测在网络重构、网络演化模型评价、推荐系统等方面有着重要应用。链路预测作为一种新兴的兼具理论研究和实践应用的技术手段，有必要得到图书情报学界的关注，并应用到知识网络，特别是科研合作网络的研究中。

本书的选题正是在这样的学术前沿和实践需求的背景下提出来的。本书选择我国人文社会科学合作网络作为研究对象，在分析网络结构的基础上思考其中的链路预测问题，合作网络不涉及异质网络情形，主要实现以下目标：①研究合作网络的静态结构特征和动态演化特征，在此基础上，研究基于结构相似性的链路预测方法在合作网络中的预测效果；②一方面研究不同网络结构对链路预测效果造成的影响，另一方面识别和分析合作网络中的异常链路，研究异常链路对网络连通性造成的影响，揭示网络结构与链路预测之间的关系；③研究科研合作网络演化的内在驱动因素，建立利用链路预测来构建和评价网络演化模型的理论框架。

对应上述研究目标，本书的主要内容是：从图书情报学的角度梳理和总结链路预测相关理论和方法，借助中文社会科学引文索引（CSSCI）数据库，尝试从链路预测的角度对真实的科研合作网络进行分析，从科研合作网络的结构、科研

合作网络的链路预测效果、网络结构与链路预测的关系、网络演化与链路预测的关系这四个方面展开研究，挖掘我国人文社会科学领域的合作特点，在此基础上为科研管理人员提供较为合理的政策建议，提高科研合作效率和生产力，具有重要的理论意义和实践意义。

在理论方面，对科研合作网络进行链路预测研究，可以产生对实际预测工作有指导作用的理论结论，并有助于建立图书情报学角度的链路预测的理论框架。目前有关科研合作网络（以至于知识网络）的演化机制的提法较多，如富者愈富、好者变富、马太效应、累积优势、同质性等，但大部分研究都是通过演化模型生成模拟网络，将模拟网络的某些统计特征量与真实网络进行对比，来说明所提出的演化模型能够描述真实网络的演化过程，缺乏评价这些演化模型的研究。借助链路预测的理论框架和评价方法可以从另一个角度来观察真实网络及其演化过程，这是对以往通过理论分析和数学建模形成的演化模型研究的有效补充。

在实践方面，科研合作网络的未知链路和未来链路可能代表着被忽略的或潜在的合作关系，而异常链路可能是影响到整个科研合作网络的连通性的重要合作关系。识别和分析这些链路，并结合相关的背景知识，可以为科研管理工作提供重要依据，从而制定出较为合理的科研发展战略，进而提高科研合作效率和生产力。这些工作属于科研管理工作的一部分，同时也是图书情报学领域长久以来所进行的一项实践工作，具有重要的应用价值。

科研合作网络的链路预测研究是一个十分必要但又很有挑战性的课题。有别于以往研究科研合作网络的描述模型和过程模型，本书研究采用链路预测思想来挖掘真实科研合作网络中被忽略的或潜在的合作关系，并利用识别出来的链路来分析合作网络中重要的合作关系和组合，从另一个视角来补充和完善科研合作网络的结构与演化的研究。研究从一开始的对象选择、数据获取、切入点的确定，到后面实证研究和理论构想，无时无刻不面临着巨大挑战。将链路预测应用在我国人文社会科学合作网络上，结合具体场景进行深入分析和探讨，这本身就具有很强的实用价值，有助于揭示链路预测方法本身存在的优势和局限性，相关结论对实践工作也有一定参考价值。然而，由于时间、知识结构以及本书篇幅的限制，还存在着许多有争议或有价值的问题值得进一步深入思考和研究。

本书是国家自然科学基金重点国际（地区）合作研究项目“大数据环境下的知识组织与服务创新研究”（项目编号：71420107026）、国家自然科学基金青年项目“心智空间视角下科学知识生成与演化机理研究”（项目编号：71704138）和中国博士后科学基金特别资助项目“科研合作网络的演化模型与动力学研究”（项目编号：2017T100585）的研究成果之一；同时，得到“中央高校基本科研业务费专项资金”资助。

感谢我的硕士生和博士生导师马费成教授，给我充足的时间去学习、思考和转变，耐心为我答疑解惑，一步步引导我走上科研道路，并且在我工作后还给予巨大的物质和精神支持。感谢黄长著教授、朱庆华教授、胡昌平教授、李纲教授、陆伟教授等在本书初稿时，给予的大量修改建议。我的师妹李亚婷、贾茜、戴怡清，参加了本课题的实证研究，在数据收集处理、算法分析、模型构建等方面做了大量工作，尤其是牺牲了宝贵的节假日时间，在此表示感谢。

在课题研究过程，尤其是本书成稿过程中，参考了许多学者的论著，他们的成果为本书提供了丰富的素材和理论支撑，书中都以参考文献的形式进行了标注，如果有不慎遗漏的，亦表示特别的歉意。

张　斌

2018 年 4 月于珞珈山

目　录

第1章　知识网络的链路预测

长久以来，人们一直在研究文献单元、信息单元、知识单元之间的内在联系，从而可以更好地组织、导航、检索和利用文献信息资源。这一研究路线主要是对文献信息等显性知识进行组织、管理和优化，形成了以文献计量学、科学计量学和信息计量学为核心的研究领域。近些年，科学知识网络（简称知识网络）逐渐成为图书情报学、计算机科学、管理学、教育学、心理学等诸多领域共同关注的热点问题。在具体的研究中，以 de Solla Price（1965）为代表的统计物理方法和以 Brookes（1981）为代表的认知地图构想，反映和奠定了当前的两种主流研究思路（刘向等，2011），并且在知识网络的结构和演化问题上取得了很多研究成果（张金柱等，2012）。

研究知识网络的结构和演化是厘清知识发展脉络，对知识的发展趋势和演化动态进行预测的基础。然而，刻画网络结构特征的统计量非常多，不同的演化机制和模型也很难判断孰优孰劣。最近几年，复杂网络的链路预测方法受到越来越多的关注，链路预测在网络重构、网络演化模型评价、推荐系统等方面有着重要应用。科研合作网络是知识网络的一种，也同样是复杂网络，因此本书先从图书情报学的角度梳理和总结有关知识网络链路预测方面的研究成果，以分析当前国内外研究现状。

1.1　链路预测的研究思路与方法

给定某个时点 t 下的网络 $G(V, E)$，其中 V 为节点集合，E 为连边集合。链路预测的目的在于预测未来一个时点 t'（$t' > t$）下节点之间新的连边（未来链路，future links），或者挖掘当前网络中缺失的或者未被观察到的连边（未知链路，missing links）。链路预测可以被看作是一项统计分类任务，它的分类对象是网络中的节点对。以检索式 TS=("link prediction") 在 Web of Science（WoS）的 SCIE 和 SSCI 中进行检索，时间跨度为 2000 ～ 2017 年，检索时间为 2018 年 1 月 14 日，共获得 478 篇文献记录，它们主要集中在计算机、工程、物理、通信等科学领域，近年来研究呈现快速上升趋势，如图 1-1 所示，并有成为图书情报学领域新兴热点的趋势。

图1-1　链路预测研究论文时间分布

使用 CitNetExplorer（van Eck and Waltman，2014）对这 478 篇文献间的直接引证关系进行聚类并可视化，如图 1-2 所示。其中，被引得分前三的文献作者及内容是：① Liben-Nowell 和 Kleinberg（2007）提出基于网络拓扑结构的相似性定义方法，将这些指标分为基于节点和基于路径的两大类，并分析了若干指标在物理学合作网络链路预测中的效果；② Lü 和 Zhou（2011）关于复杂网络中链路预测问题的综述；③ Zhou 等（2009）分析 10 种基于节点局部信息的相似性指标在六个真实网络中进行链路预测的表现。这 478 篇文献存在两个大的类群，分别包含了 198 篇和 140 篇文献，代表了链路预测研究的两大分支，即未知链路预测和未来链路预测，类群展开如图 1-3 和图 1-4 所示。下面基于链路预测相关的高被引文献来回顾基本的研究思路与方法。

早期关于链路预测的研究思路和方法主要基于马尔可夫链和机器学习（Sarukkai，2000；Zhu et al.，2002），这在计算机科学领域已有较深的研究。而在 Lü 和 Zhou（2011）给出的关于复杂网络链路预测的综述中，基于网络结构的链路预测主要有三种研究思路与方法，如图 1-5 所示。

第一种研究思路是基于相似性（similarity）进行链路预测。相似性表达的是一种接近程度（proximity），其研究的前提假设是两个节点之间的相似性越大，它们之间存在链路的可能性就越大。刻画相似性的方法有很多，其中最简单的是利用节点的属性。在社会网络中，如果两个人具有相同的年龄、性别、职业、兴趣、爱好等，就说明他们很相似；而在知识网络中，如果两篇论文具有相似的标题、关键词、主题词、摘要等，也可以说明它们很相似。基于节点自身属性来判断相似性的确可以得到很好的预测效果，如论文相似性检测，但在很多情况下获取这些信息并不容易。目前讨论更多的是基于网络结构信息的相似性链路预测方

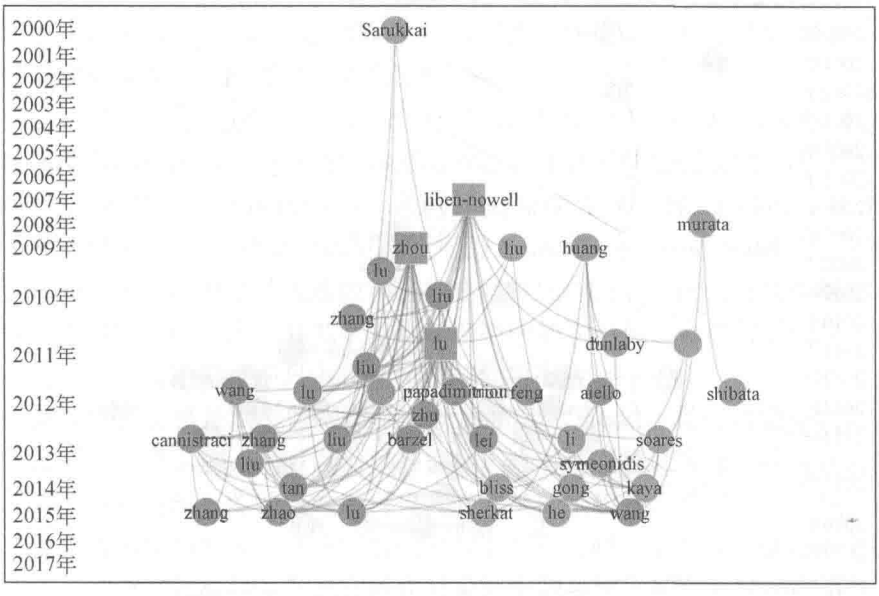

图1-2　链路预测相关文献间的引证关系可视化

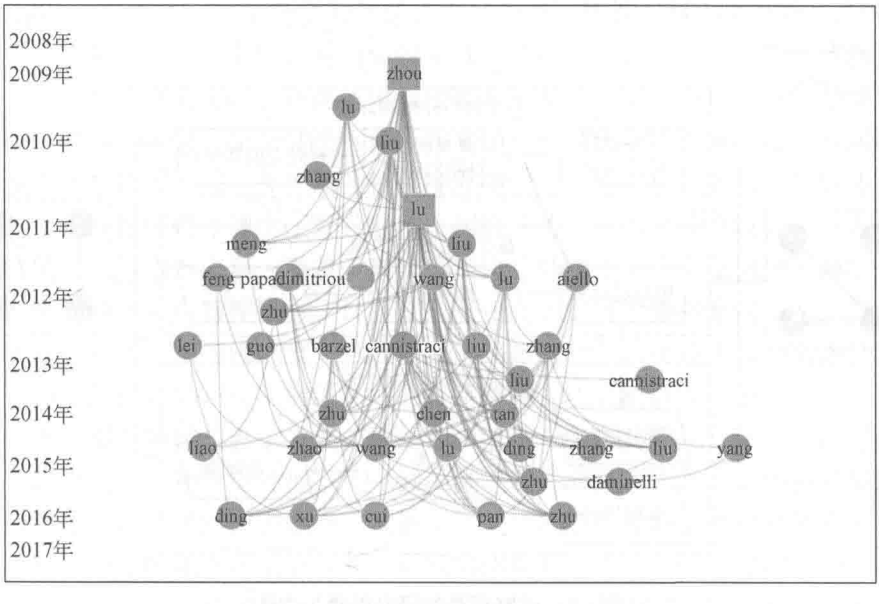

图1-3　链路预测相关文献中的最大类群展开

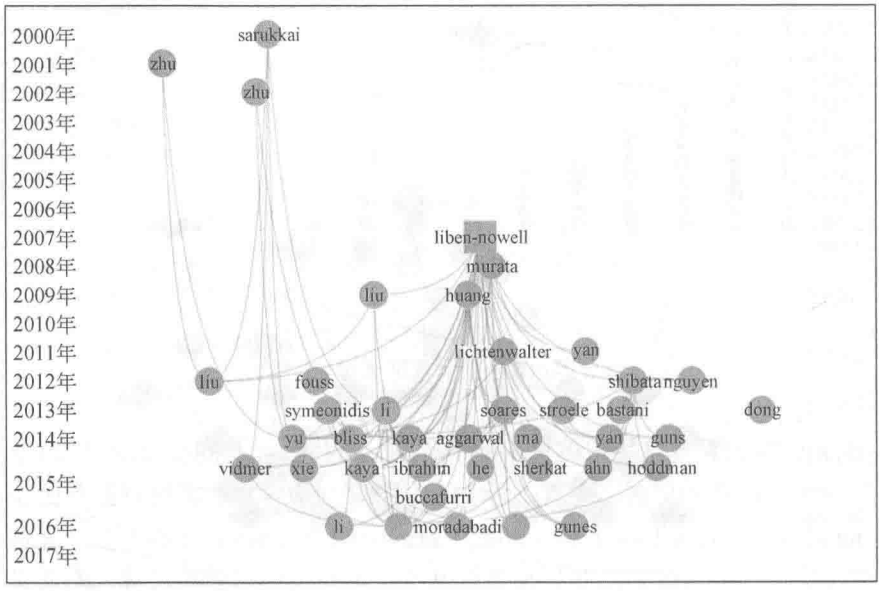

图1-4　链路预测相关文献中的次大类群展开

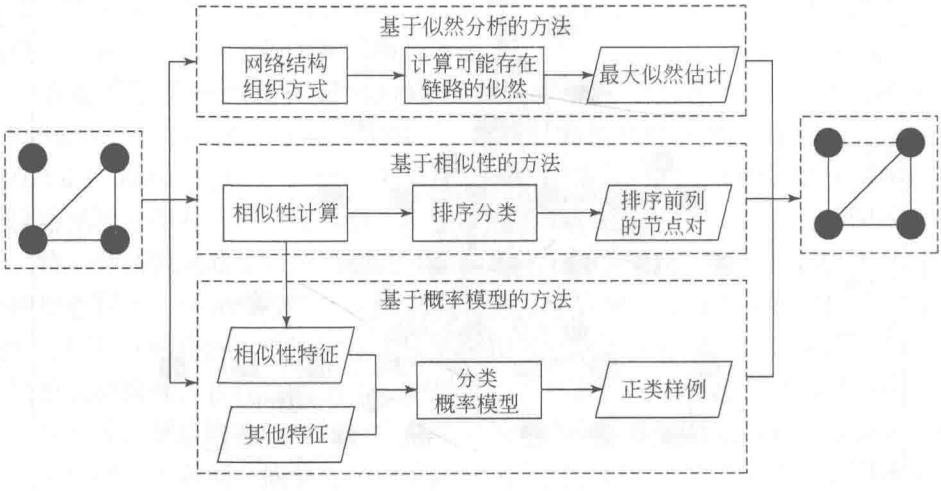

图1-5　链路预测的研究思路与方法

法，其中最简单相似性指标是共同邻居，即两个节点如果有更多的共同邻居就更可能产生连边。在社会网络中，如果两个人拥有更多相同的朋友，那么这两个人很可能也是认识的（Kossinets，2006）；而在知识网络中，有更多共同合作者的两个科学家在未来合作的可能性较高（Newman，2001a）。

第二种研究思路是基于似然分析进行链路预测。Clauset 等（2008）提出了一种利用网络的层次结构进行链路预测的方法，在具有明显层次结构的网络中，如恐怖袭击网络、草原食物链等，预测的精确度表现较好。Guimerà 和 Sales-Pardo（2009）提出了一种基于随机分块模型（stochastic block model）（Holland et al.，1983）来预测网络缺失连边和识别错误链路的方法。随机分块模型是一种基于最大似然估计的方法，其基本思想是网络中的节点可以分为若干模块，而两个节点之间存在连边的概率只由它们所在的模块决定。

第三种研究思路是基于概率模型（probabilistic model）进行链路预测。其基本思路是建立一个含有一组可调参数的模型，之后使用优化策略寻找最优的参数值，使得所得到的模型能够更好地再现真实网络的结构和关系特征，网络中两个节点产生连边的概率就等于在该组最优参数下它们之间产生连边的条件概率。

上述三种研究思路的立足点和研究对象的不同导致了研究方法和应用范围存在明显差异。目前，关于第一种研究思路探讨得最多，应用范围也最广，其相似性指标也有不同的划分标准，如：①基于邻居节点和基于路径的（Liben-Nowell and Kleinberg，2007），②基于局部信息、基于路径和基于随机游走的（吕琳媛，2010），③基于局部、全局和类局部的（Lü and Zhou，2011）。对于具体预测指标和方法可参照相关文献，在此不做详细展开。基于结构相似性进行链路预测只涉及网络的结构信息，相似性指标计算起来会比较简单，但不同指标在不同网络中的预测能力却不一致，其预测的精确度取决于对网络结构特征刻画的好坏。基于似然分析进行链路预测由于要针对整个网络结构，计算复杂性较高，不太适合在规模较大的网络中应用。而基于概率模型进行链路预测，不仅使用了网络结构信息，还会涉及节点属性信息，计算复杂度更高，且需要数据库的支持，但其优势在于具有较高的预测精确度。三种研究思路各有所长，受益于知识网络的结构和演化相关研究成果，基于网络结构的链路预测方法有望得以运用。

图书情报学领域充满了各式各样的知识网络。从知识节点上看，论文、作者、概念是最常见的三种节点类型。从知识关联上看，共现关系和引证关系是最重要的两种关联关系，其他诸如耦合、共被引、互引网络等体现的依然是共现关系和引证关系，只是构建方式稍显特殊。根据网络中是否含有不同类型的连边，可以将知识网络划分为同质网络（homogeneous network）和异质网络（heterogeneous network）。表 1-1 列举了知识网络类型和典型代表，这里不再详细对单一网络展

开介绍。这些知识网络规模不一，有的规模还极其庞大。因此，在链路预测问题上主要采用的是基于相似性进行链路预测的思路以降低计算的复杂性，同时也会引入一定的语义和属性信息来保证预测的精度。

表1-1 知识网络类型和典型代表

网络类型		典型代表
同质网络	无权有向网络	引证网络
	无权无向网络	各种网络的二值特殊形态
	加权有向网络	作者互引网络、期刊互引网络等
	加权无向网络	合作网络、共词网络、耦合网络、共被引网络等
	二分网络	科学家－论文二分网络、疾病表现－致病基因二分网络
异质网络		作者多重交互网络

1.2 同质网络的链路预测

同质网络中只存在一种类型的连边，因此二分网络也是一种同质网络。大部分聚焦在同质网络的链路预测研究，主要是在引证网络、合作网络和共词网络等方面进行了探索，本节就它们中的研究成果进行梳理。

1.2.1 引证网络

科学引文索引创始人 Garfield（1955）认为，引证行为体现了知识的前后承接关系，利用引证网络可以实现领域知识内容结构上的聚合。引证网络是最早提出的一种知识网络形态，它的主要研究对象是学术论文。论文之间的引证关系一经发表一般不会出现再变动，绝大部分都会服从时间先后关系，且大部分论文都经由相关主题的其他文章来引用，很少出现相互引用，网络中没有环状结构。理论上讲，这种引证关系一定要满足时间先后关系，且不出现相互引用，在时序上具有单向无回路的特征。但在实际情况中，由于存在优先出版等导致出版周期错乱的因素，引证网络中存在极少的不符合理论情形的连边。这些单向关系无法使用无向网络中的相互关系或共现关系进行刻画，因此需要借助有向网络的形式。

引证网络是一种无标度网络。Barabási 和 Albert（1999）提出了通过引入增长和择优连接机制构建增长网络模型（简称 BA 模型）来解释真实网络无标度特性出现的内在机理。事实上，这一思想可以追溯到 de Solla Price（1965，1976）

和 Simon（1955）的研究，只不过当时使用的是累积优势和马太效应来描述这一过程。当然，还有其他的因素影响着引证网络的演化过程。例如，时间效应会在一定程度上平抑度择优所导致的马太效应的负面影响（刘向和马费成，2012）；知识学习和知识引证通常还是基于局部知识领域（马费成和刘向，2013；刘向等，2013）。因此，真实的引证网络往往是由多种机制混合作用而成。

学术论文中的引证与被引证体现出了同质性（homophily），使得大量文献得以分群。在一些文献中，聚类效应被当作一种特殊的同质性，即拥有共同邻居的节点被看成是拥有相似的网络环境。同质性体现了节点间的相似性及相容性能对它们之间是否形成连边产生重要影响（McPherson et al.，2001）。显然，论文间的引证关系倾向于产生在具有相似内容的文献之间（Cheng et al.，2009）。在文献计量学中，通常会使用文献耦合和共被引两种方法来揭示论文的主题相似性，以及相互之间的作用和联系。文献耦合和共被引都建立在相似的假设基础之上，即具有耦合关系和共被引关系的论文可以认为它们在学科内容上存在某种相关性。

引证网络的形成机制非常复杂，而同质性的存在使得可以利用链路预测来探究网络结构的形成机制。无向网络中三个节点的相互关系非常简单，在已经拥有共同邻居的两个节点之间产生连边，只能形成一种新结构，称为三角形结构（Rapoport，1953）。而在有向网络中，同样考虑三个节点，情况就复杂得多（Brzozowski and Romero，2011）。在不考虑互惠连接机制和具体节点间连边的情况下，引证网络的三角形结构可分为前馈回路和反馈回路两种。Milo 等（2002）引入模体（motif）来描述网络中反复出现的相互作用基本模式。相比三节点结构而言，四节点结构更加复杂。Milo 等（2002）发现在很多网络中，四节点的双风扇结构和双平行结构是非常显著的。实际上，可以将这些识别出来的模体看成网络生成和演化过程中在特定限制条件下形成的特殊局部结构。于是，在引证网络中三个节点和四个节点的情况下可得六种子图结构，如图 1-6 所示。

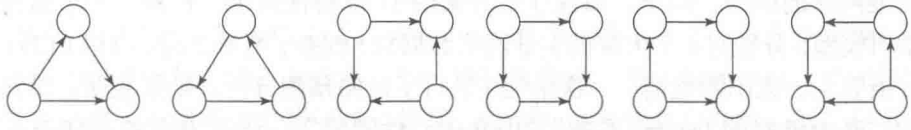

图1-6　六种含有回路的最小子图

Zhang 等（2013）提出的势理论（potential theory）为强化上述假设和描述引