



数字信息资源开发利用与管理研究丛书

丛书主编 马费成

知识网络的形成与演化

Formation and Evolution of Knowledge Network

刘向 著



WUHAN UNIVERSITY PRESS

武汉大学出版社

本书出版获得以下资助：

- 国家自然科学基金重点项目“基于生命周期的数字信息资源深度开发与管理机制研究”
(项目编号：70833005)
- 国家自然科学基金项目“知识网络的形成机制及演化规律研究”(项目编号：71173249)



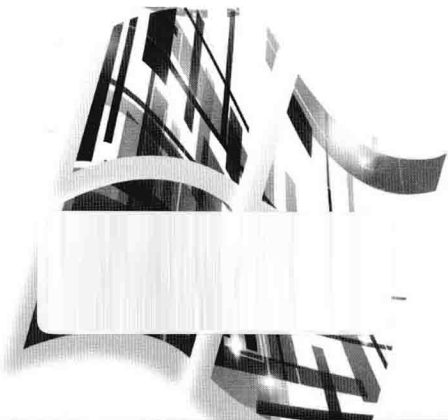
数字信息资源开发利用与管理研究丛书

丛书主编 马费成

知识网络的形成与演化

Formation and Evolution of Knowledge Network

刘向 著



WUHAN UNIVERSITY PRESS

武汉大学出版社

图书在版编目(CIP)数据

知识网络的形成与演化/刘向著. —武汉: 武汉大学出版社, 2014. 6
数字信息资源开发利用与管理研究丛书

ISBN 978-7-307-13240-5

I. 知… II. 刘… III. 互联网络—应用—知识学 IV. G302

中国版本图书馆 CIP 数据核字(2014)第 085727 号



责任编辑:詹 蜜

责任校对:汪欣怡

版式设计:马 佳

出版发行: 武汉大学出版社 (430072 武昌 珞珈山)

(电子邮件: cbs22@whu.edu.cn 网址: www.wdp.com.cn)

印刷: 武汉中远印务有限公司

开本: 720 × 1000 1/16 印张: 15.25 字数: 216 千字 插页: 1

版次: 2014 年 6 月第 1 版 2014 年 6 月第 1 次印刷

ISBN 978-7-307-13240-5 定价: 32.00 元

版权所有, 不得翻印; 凡购我社的图书, 如有质量问题, 请与当地图书销售部门联系调换。

总 序

自 20 世纪 50 年代数字信息资源产生之后，数字信息资源的开发利用和管理就一直受到各国政府、企业机构以及学术界的重视，仅相关法规就超过 2 000 部。尤其是近年来，数字信息资源的“井喷式”增长，给理论研究和实践管理都带来极大挑战。主要表现在四个方面：

第一，总量急剧膨胀，但组织无序，存取效率低；第二，数字信息资源生命周期极大缩短，老化速度加快，以致数字信息资源所依托的技术环境、业务环境、管理体制和政策难以适应，技术方案深陷“数据危机”；第三，数字信息资源的拥有和利用能力形成的数字鸿沟严重影响了社会公平，影响其可持续开发；第四，数字信息资源的集中保存机制弱化以致存储机构更加扁平化，数字信息资源内容的权威性面临极大挑战，使得有价值的资源得不到有效利用。而我们当前的数字信息资源管理方法、理念和机制还仍然停留在传统信息资源的生命周期环境，凸显出很多不适宜、不协调的因素。此外，数字信息资源本身不同于以往任何信息资源的特殊性，也给予管理和利用带来极大不便。

数字信息资源本身具有十分重要的作用。作为信息存储和传递的重要载体，是科学决策的重要依据，也是科学研究和创新的重要支撑，是提升公众社会福利的重要领域，还是一国或一个地区精神文化的重要组成。数字信息资源也是一个国家的数字资产，是学术研究信息的数字存档，一个国家的科技创新能力以及与此相关的国际竞争力都依赖于其快速、有效地开发与利用数字信息资源的能力。因而，必须结合我国的国情，深入研究，制定科学、可行的管理体系。具体到我国，数字信息资源总量还不够（仅占全球信息总



量的7.9%)，分布和结构都不合理，作为大国崛起的文化战略也需加强数字信息资源开发利用。因此，国民经济和社会发展“十一五”规划纲要、国家中长期科学和技术发展规划纲要(2006—2020)、2004年《关于加强信息资源开发利用工作的若干意见》以及《2006—2020年国家信息化发展战略》都将网络(数字)媒体信息资源的开发利用计划列为我国的战略发展计划，还是今后很长一段时间内的重要任务。

著名的数字图书馆专家 Neil Beagrie 和 Maggie Jones 指出，“数字信息的飞速增长，已经严重影响我们的文化、知识库和经济。”尤其是数字信息资源总量呈高指数增长，已经影响到信息资源的有效开发利用，并危及人类的知识再创造活动，因而围绕数字信息资源而展开的研究在全球范围内兴起。美国加州大学伯克利分校的博格曼教授坦言，数字信息的研究是歧义最多，解释视角最为广泛的模糊研究领域；DPC 报告也指出，数字信息资源研究“作为新生学科，最大不足是缺乏科学的术语定义和术语分类体系”。目前，国内专家学者关于数字信息资源的狭义定义大体上能达成一致，即电子形式的信息资源，指一切以数字形式生成和发行的信息资源。但是对数字信息资源管理的内涵，却存在三类不同的研究范式：系统(技术)学派专注于其技术实现和信息管理，而对于其资产、内容及文化的价值属性没有有效考虑；经济管理学派专注市场、产业、制度和社会行为，却面临公平与效率、创新与保护的两难抉择，而且对信息商品生产过程的复杂性关注也不够；过程学派能够比较综合考察技术、管理和经济行为，却往往局限于图书情报的应用领域，未能建立面向政府和企业的综合理论模型。因而，亟须建立一种数字信息资源管理的“全息视野”，数字信息资源的生命周期模型正是分析数字信息资源管理的一个理想的思路 and 工具。

数字信息资源生命周期系指数字信息资源从生成到其价值失去的整个时间区间，尽管不同主体、不同类型的数字信息资源生命周期各不相同，但都可以映射到由生产—分发—采集—加工—存储—服务—再生产等环节构成的循环上。这一循环实际上对应于数字信息资源的业务流程，为观察和研究数字信息资源生命周期提供了一



条清晰的主线，同时也为管理和控制数字信息资源的生命周期提供了可能，使得我们可以采用先进适用的技术方法、合理的政策法律工具通过生命周期循环来管理数字信息资源，以提高其有效开发和利用的效率，而且可以实现数字信息资源的深度开发。

鉴于以上原因，我们在2008年以“基于生命周期理论的数字信息资源深度开发与管理机制研究”为题申请国家自然科学基金重点项目获得批准。我们希望以数字信息资源的生命周期为主线，研究数字信息资源开发、利用和管理的主体、客体和过程的相互影响关系，构建数字信息资源深度开发和管理机制，确立数字信息资源的政策内容架构，有效协调各主体矛盾，并建立过程环节的有效关联，实现数字信息资源的有序管理、深度开发、有效利用和可持续发展，发挥数字信息资源在社会经济、科技、文化发展中的最佳效能。具体实现以下目标：

(1)研究数字环境下信息资源的新特征、新趋势，构建数字信息资源的生命周期模型，并将这一模型应用于数字信息资源开发、利用和管理的不同主体、客体和过程，揭示其相互联系与影响的内外因素，建立数字信息资源深度开发模式和管理机制。

(2)开展政府、企业、学术等不同领域数字信息资源生命周期的实证研究，并对不同生命周期的数字信息资源管理成本、风险、交换和流通效率进行评估，分析其影响因素，建立更为通用的方法论体系。

(3)以数字信息资源生命周期为工具和主线，解决数字信息资源开发过程中的各环节关联性差、技术标准不统一、多元主体冲突、产业高度集中而创新分散、政策法规不完备等给数字信息资源开发带来的问题，促进数字信息资源的深度开发，建立我国数字信息资源产业生态模式并揭示其运作机制。

(4)建立基于数字信息资源生命周期的、结合架构体系和信息资源目录体系的数字信息资源管理架构，并开发或引入相应的辅助工具软件，为政府、企业的数字信息资源管理制度提供参考工具。

(5)通过对数字环境下的政策需求、政策目标的研究，确立数字信息资源政策的内容架构，为制订结构合理的数字信息资源政策



提供依据。

课题组通过 4 年多的探索和研究,基本实现了上述目标。我们在国内外重要杂志上发表学术论文近百篇,提供咨询研究报告 3 份,并按课题的预定目标完成了多部专著,其中《网络信息序化原理》和《网络信息生命周期模型及规律研究》两部著作已经在科学出版社出版,《数字信息资源优化配置》被列为武汉大学学术丛书即将出版。本套丛书作为上述成果的补充,我们计划出版 6 本,分别讨论数字信息资源规划、信息搜索、知识网络演化、数字信息资源管理政策等方面的问题,能够比较全面地展示生命周期视角下数字信息资源的开发、规划、管理和利用方面的研究成果。

感谢参加丛书撰写的课题组成员查先进、唐晓波、陆伟、裴雷、丁韧、刘向、罗琳、张晓娟、王晓光、刘萍、陆泉、望俊成等同志,他们不仅承担了课题研究的任务,而且为本套丛书的写作付出了艰苦的努力,特别感谢武汉大学出版社詹蜜女士,她为本书的出版付出了许多辛劳。

数字信息资源开发管理的过程和环境涉及多主体、多维度、多环节、多类型的数字信息资源,相当复杂。依靠一个项目来解决其有效开发利用中的各种问题是不现实的,需要我们继续沿着这一方向深入探索和研究。

马费成

2012 年 10 月于珞珈山

前 言

知识网络将人类知识构建成一个时间上和结构上无限延展的关联系统，为分析知识的结构、探索创新形成的过程以及指导科技实践提供了一条有效的途径。本书针对知识网络的结构及其演化过程、知识网络中的增长老化、知识网络中热点与趋势的涌现以及知识空间中创造者的迁移分布等问题展开了讨论。

对于知识网络的结构及其生成过程方面，针对 Price 经典模型在描述知识产生上与实际大规模统计数据的差异性，探讨了知识演化的马太效应过程中潜隐的时间因素的作用；针对研究者的专业领域限制问题，分析了视域受限条件下的局域演化模型；针对研究者研究问题的聚集性问题，构造了基于邻域演化的过程模型。研究结果表明：度择优所体现的马太效应的作用是全局性的，时间择优所体现的后发优势的影响则是局部的，时间效应一定程度上平抑了度择优所导致的马太效应的负面影响；跨领域交叉机制既保证形成一定集聚拓扑结构又满足学科知识交叉引用的要求；而基于邻域演化的过程模型则生成了兼具小世界和无标度特征的动态增长网络。

对于知识网络中的增长老化问题，采用了知识节点泛增长函数的方式，考察了节点在非平稳增长条件下知识网络的拓扑结构及节点的历时老化；讨论了知识产生的时点与知识增长老化之间的关系，提出了对知识老化曲线形成的一种客观性的新解释；研究了连接边非平稳增长条件下，知识网络的增长老化问题。研究发现：当知识增长率为线性等收敛函数时，知识网络的度分布与增长模式无关，且知识节点的历时被引是一直衰减的；当增长率函数为发散性的指数函数时，知识节点的度分布指数较前者要小，节点的历时被引数则是单调上升的；指数增长模式具有较平坦的度分布，而线性



增长等在知识利用上较为均匀,知识利用率的层次性没有体现出来;在所属学科的扩展期产生的知识节点的历时被连接数先上升后下降,而在衰退期产生的节点的历时被连接数一直是衰减的;知识的利用效率随其所属知识领域的扩张而增加,随衰退而减少;出度线性增长形成更平坦的度分布,而出度对数增长对度分布影响不大,出度对数增长更加符合现实情况。

对于知识网络中热点与趋势的形成过程及其探测问题,构建了知识网络中的群体动力学模型。模型认为知识类群的热点与趋势受到关联类群的影响,通过考察知识类群在相互作用的环境中主题词向量的时序变化,进而得到知识类群的演化方向。与传统分析方法的比较研究证实了这一方法的有效性:群体动力学方法对趋势的预测比传统方法具有优越性,对于少数据、不确定性较大的小知识领域的预测效果明显优于传统方法。

对于知识空间中知识创造者的创造活动在时间和空间上的迁移分布问题,统计分析了高产者、低产者和创造者全体在连续知识创造活动中的时间间隔和空间迁移。研究发现高产者的创造活动的时间间隔满足重尾分布,具有阵发特征,长时静默和密集阵发同时存在;低产者的时间间隔分布近似指数函数,随机性和偶然性稍强一些;知识创造者连续创造活动的空间迁移也满足重尾特征,密集的局限于某一领域进行知识创造,然而常常也会进行一些长距的知识领域探索。

本研究受到国家自然科学基金重点项目“基于生命周期理论的数字信息资源管理及深度开发机制研究”(No. 70833005)、国家自然科学基金项目“知识网络的形成机制与演化规律研究”(No. 71173249)、国家自然科学基金青年项目“专利网络中群体行为视角的科技创新趋势涌现研究”(No. 71303090)的资助。

目 录

第1章 绪论	1
1.1 研究背景	1
1.2 国内外研究现状	4
1.3 内容与结构	19
1.4 方法与技术路线	21
1.5 创新点	25
1.6 研究的难点	26
第2章 基础理论与方法	28
2.1 知识网络	28
2.2 拓扑结构	32
2.3 网络演化	47
2.4 群体动力与网络耦合	49
2.5 本章小结	54
2.6 附录：仿真实验	54
第3章 知识网络的结构生成	59
3.1 引言	59
3.2 时间的作用机制	61
3.3 局域演化网络	72
3.4 结构生成模型	82
3.5 本章小结	93
3.6 附录：仿真实验	95



第4章 知识网络中的增长老化	101
4.1 引言	101
4.2 知识节点的增长与老化	103
4.3 增长老化与知识节点产生时点的关系	117
4.4 连接边的增长机制	127
4.5 本章小结	139
第5章 知识网络中热点与趋势的涌现	142
5.1 引言	142
5.2 研究背景	144
5.3 研究设计	151
5.4 群体动力模型	154
5.5 操作与实施	158
5.6 基于直接引证网络的实验	161
5.7 基于文献耦合网络的实验	168
5.8 结果讨论	174
5.9 本章小结	176
5.10 附录：不同类型知识网络探测新兴趋势的效能比较	177
第6章 知识空间中知识创造者的迁移分布	181
6.1 引言	181
6.2 研究背景	182
6.3 研究设计	185
6.4 数据与方法	187
6.5 统计分析	188
6.6 结果讨论	193
6.7 本章小结	195
第7章 专利网络中科技创新趋势的探测	196
7.1 引言	196

7.2	当前主要的研究概况	198
7.3	专家判断视角发现科技创新趋势	200
7.4	信息分析视角探测科技创新趋势	202
7.5	专利可视化视角观测科技创新趋势	205
7.6	网络演化视角分析科技创新趋势	207
7.7	科技创新趋势分析系统和工具	209
7.8	简单评述	210
7.9	本章小结	212
第 8 章 结语		214
8.1	内容总结	214
8.2	存在的局限与研究展望	216
参考文献		218

第1章 绪 论

► 1.1 研究背景

知识网络的研究立足于理论和实践两个方面的需求，而对这两个方面问题的探讨又是紧密关联的：理论方面的需求是人类日益庞杂、繁复的信息积累，使得人们试图理清知识发展脉络变得尤为困难。随着现在信息和通信技术的发展，信息快速增长，学科领域纷繁复杂，知识门类在高度分化的同时不断趋向综合，已逐步演化为一个复杂的网络体系。面对如此庞大而复杂的知识系统，人们要理清其中知识发展的脉络，对知识的创新领域和发展趋势做出准确的判断也变得越来越困难。这一情势已经危及人类的知识生产、利用和再创造活动，学术界形象地称其为情报危机^①。实践方面的需求则与当前经济社会的进步和科技创新密切相关。科技创新是推动经济社会发展的源动力，自近年爆发全球范围的金融和能源危机以来，世界各国纷纷启动了一轮声势浩大的科技创新和产业复兴计划——美国、欧洲、日本等发达经济体投入巨资支持本国在新能源、节能环保、信息技术等方面的科技研发和产业提升^②；我国政

① 马费成，等．信息管理学基础[M]．武汉：武汉大学出版社，2002：29.

② 东方证券研究所策略团队．推动中国经济增长的下一个支柱产业——经济发展转轨期的战略性新兴产业[R]．战略性新兴产业专题系列投资策略报告之一，2010-02.



府也迅速将创新提升到国家战略的高度^①，提出大力培育和支持战略性新兴产业^②。然而，哪些技术是未来发展的核心和前沿，哪些领域是当前亟须突破的重点，成为摆在产业界和科技界面前的首要问题。要对这个问题进行回答，就需要识别知识发展的主干，探讨科技发展的前沿。

知识网络将人类知识构建成一个时间上和结构上无限延展的关联系统，也为分析知识网络的结构、探索创新形成的过程提供了一条良好的途径。美国科技信息研究所创始人 Garfield 在 20 世纪 50 年代便意识到科学引证网络 (Citation Network) 可以反映科学知识之间传承、发展的关系，并且尝试利用引证网络研究科学知识发展的历史、脉络和结构^③；Bernal、Price、Leake 和 Shryock 等学者均表示对这一想法的认同，针对几个领域的引证网络分析也证实了其有效性^④。Price 借助物理统计的方法对引证网络的拓扑结构及人们的引证行为进行了分析，通过对文章参考文献的时序分布的研究，指出引证过程中经典理论和研究前沿 (Research Front) 的区分^⑤。1981 年，英国情报学家 Brookes 又提出了认知地图 (Cognitive Map) 的构想，他认为可以将紧密相连的若干学科领域的固有联系表示成概念联结网络，每一片段情报都成为网络经脉之上的一个要素，从而形成表示科学认知结构的知识发展脉络，他还利用加拿大学者 Farradan 的关系索引和数据库，针对液晶领域的引证数据绘制出了

① 中华人民共和国国务院. 关于发挥科技支撑作用促进经济平稳较快发展的意见[Z]. 中华人民共和国国务院工作文件, 2009-03.

② 温家宝. 让科技引领中国可持续发展[R]. 首都科技界大会工作报告, 2009-11.

③ Garfield E. Citation Indexes for Science [J]. *Science*, 1955, 122: 108-111.

④ 转引自 Garfield E. *Citation Indexing—Its Theory and Application in Science, Technology, and Humanities* [M]. Philadelphia: ISI Press, 1983: 69-123.

⑤ Price D. J. Networks of Scientific Papers [J]. *Science*, 1965, 149: 510-515.



认知地图的雏形^①。对于 Brookes 的认知地图构想, 马费成对其进行了高度评价, 认为认知地图如果成功构建, 可以有效遏制人类的情报危机, 会使情报学研究取得突破性进展^②; 赵蓉英、邱均平等学者也认为这一构想指明了未来知识组织的最高目标^③。

基于类似的想法, 人们探索了分析知识网络结构和知识发展脉络的多种方法和工具, 例如科学知识图谱 (Mapping Knowledge Domains)^④、概念地图 (Concept Map)^⑤、合作网络 (Collaboration Network)^⑥、知识超网络 (Knowledge Super-network)^⑦等。然而, 这些研究更多的关注于知识网络的可视化形态或相互关系的描述上, 多是定性的概念模型和实证性研究思路。知识网络是一个复杂的时变动态系统, 当前的方法在揭示知识发展的动态演变过程及洞察知识的创新领域与发展趋势上遇到困难。

本书基于引文与主题词形成的混合知识网络, 通过复杂网络和群体动力学的方法, 构建知识网络的演化过程与动力学模型, 重点研究知识网络的生成机制和演化规律; 分析知识网络演化中的动力学, 探索研究热点的形成过程和创新趋势的涌现机理; 寻找分析知

① Brookes B. C. The Foundations of Information Science (Part IV) [J]. *Journal of Information Science*, 1981(3): 3-12.

② 马费成. 论布鲁克斯情报学基本理论[J]. 情报学报, 1983, 2(4): 318-321.

③ 赵蓉英, 邱均平. 知识网络研究(I)——知识网络概念演进之探究[J]. 情报学报, 2007, 26(2): 198-209.

④ Börner K., Mane K. K. Mapping Topics and Topic Bursts in PNAS[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, 101(1): 5287-5290.

⑤ 马费成, 郝金星. 概念地图在知识表示和知识评价中的应用(I)——概念地图的基本内涵[J]. 中国图书馆学报, 2006, 32(3): 5-9.

⑥ Barabási A. L., Jeong H., Néda Z., Ravasz E., Schubert A., Vicsek T. Evolution of the Social Network of Scientific Collaborations [J]. *Physica A: Statistical Mechanics and its Applications*, 2002, 311: 590-614.

⑦ 席运江, 党延忠, 廖开际. 组织知识系统的知识超网络模型及应用[J]. 管理科学学报, 2009, 12(3): 12-21.



识研究热点、知识创新趋势和知识发展脉络的一般办法；基于知识网络研究人类的知识创造与知识学习行为，探讨促成这些行为的规则。本书的意义概括为以下两个方面：

(1)在理论上，研究知识网络的形成机制及演化规律可以展现知识单元的静态和动态分布，研究情报学和知识管理领域的基本定律；揭示科学体系的结构和内部关联，把握科学发展过程中知识创新和突破的涌现机制；揭示人类获取和吸收知识信息，生产新知识的行为和过程。

(2)在实践上，研究知识网络的形成机制及演化规律可以为有效应对情报危机、提高知识信息利用效率提供理论支持；为制定科技发展战略和有效的科技资助政策提供依据；为知识发展和创新趋势预测提供科学方法和相应的工具。

这些都是科学和知识发展与利用中的重要理论课题和实践课题，对人类有效管理知识、把握知识创新、促进科学发展具有重要的意义。

► 1.2 国内外研究现状

研究知识网络的拓扑结构及演化过程是厘清知识的发展脉络、探测和追踪创新领域及发展趋势的基础。然而目前此方面的研究却较少，即使有，也多是在统计物理学、复杂网络、动力系统类期刊中零散的涉及，如 *Physical Review Letter*、*Physical Review E*、*Nature*、*Science*、*Physica A*、*Networks* 等，情报和知识管理学者对此方面的研究还是较少的。基于这样一种现状，本节对知识网络的结构及演化方面内容进行梳理和总结，分析当前研究中的不足，并提出后续的研究重点。

1.2.1 理论来源与发展思路

1.2.1.1 两种研究思路

物理史学和情报学家 Price 在 1965 年发表于 *Science*、情报学家

Brookes 在 1981 年发表于 *Journal of Information Science* 的著名文献^{①②}反映和奠定了当前知识网络领域上的两种最主要研究思路。Price 的研究基于物理统计的方法,侧重于对实际知识网络的拓扑结构和演化特征进行客观描述与分析,其统计对象通常是实际存在的知识载体,例如文献、书刊等;而 Brookes 的认知地图构想则立足于认知的角度,他从关系索引中获得灵感,将知识对象的粒度由文献定位到更微观、更抽象的情报单元,以及由情报单元组成的情报空间,虽然他进行的实证研究也是利用引文数据,然而更重要的是指明了基于认知,对微观情报单元进行考察的研究方向。

立足点和研究对象的不同导致了上述差异,也带来了研究方法上的不同。上述两种工作思路各有所长,前者对于大规模网络的统计分析具有很好的优势,且因其研究对象的实在性和唯一性,也便于从客观的角度分析知识网络的演化过程;后一思路与人类的认知行为、使用语言和思想表达相关,虽然能更好地刻画人类的认知活动和过程,但是由于自然语言在语义上的含混性、多义性和交叉性,使得依此思路的知识网络结构与演化的研究困难较大。此后,在知识网络及相关问题上的研究也基本上延续了这些差别。

1.2.1.2 理论发展

Price 的研究思路立足于对知识网络的统计分析,避免了对自然语言的语义进行判断处理,只是关注网络的拓扑结构及生成过程,其研究对象具有客观性和可比性,目前在网络的结构及生成机制方面取得了一些成果。大多数统计物理学、力学、动力系统、系统工程背景的专家学者均采用这一研究思路。其中代表性的研究有以下几种:

1) 表现知识的前承后接关系的引证网络^③,这是最早提出的一

① Price D. J. Networks of Scientific Papers[J]. *Science*, 1965, 149: 510-515.

② Brookes B. C. The Foundations of Information Science (Part IV) [J]. *Journal of Information Science*, 1981(3): 3-12.

③ Price D. J. Networks of Scientific Papers[J]. *Science*, 1965, 149: 510-515.