情报检索语言与智能信息处理丛书

丛书主编／侯汉清

基于引文分析可视化的知识图谱构建研究

李运景／著

東南大學出版社
SOUTHEAST UNIVERSITY PRESS

情报检索语言与智能信息处理丛书(侯汉清主编)

基于引文分析可视化的知识图谱构建研究

李运景著

东南大学出版社
·南京·

图书在版编目(CIP)数据

基于引文分析可视化的知识图谱构建研究/ 李运景著. —南京:东南大学出版社,2009.12

(情报检索语言与智能信息处理丛书/侯汉清主编)

ISBN 978-7-5641-1913-3

Ⅰ.基… Ⅱ.李… Ⅲ.引文—分析—图谱—研究

Ⅳ.G353.1

中国版本图书馆 CIP 数据核字(2009)第 200920 号

情报检索语言与智能信息处理丛书(侯汉清主编)

基于引文分析可视化的知识图谱构建研究

出版发行	东南大学出版社
出 版 人	江 汉
社 址	南京市四牌楼 2 号(邮编:210096)
印 刷	南京玉河印刷厂
责任编辑	李 正
	(电话:025-83790887;E-mail:leezheng1978@sina.com)
经 销	新华书店
开 本	880 mm×1 230 mm 1/32
总 印 张	50.625(本册 5.5 印张)
总 字 数	1 310 千字(本册 142 千字)
版 次	2009 年 12 月第 1 版 2009 年 12 月第 1 次印刷
总 定 价	200.00 元(共 8 本)

丛书总序

这部丛书包括下列八本专著：

（1）薛春香著《网络环境中知识组织系统构建与应用研究》；

（2）陆勇著《面向信息检索的汉语同义词自动识别》；

（3）杜慧平、仲云云著《自然语言叙词表自动构建研究》；

（4）章成志、白振田著《文本自动标引与自动分类研究》；

（5）张雪英著《情报检索语言的兼容转换》；

（6）刘华梅、戴剑波著《受控词表的互操作研究》；

（7）何琳著《领域本体的半自动构建及检索研究》；

（8）李运景著《基于引文分析可视化的知识图谱构建研究》。

这八本专著是侯汉清教授多年来指导博士生、硕士生们进行科学研究（有些是同他们合作研究）的具体成果的一部分。这些著作的主题内容，可以归结为“情报检索语言的自动化”和“自然语言检索”两个相关的问题，或者更概括地说，就是“信息检索自动化的升级问题”，属于当前信息检索学术研究的前沿课题。

这些专著，如果将其分散来看，或许不觉得分量之重；但如果把八本专著放到一起，就可以看出其成果之丰硕。侯汉清教授在带研究生中看准一个方向不断开拓、持之以恒的精神，可以出大成果，值得我们效法。南京农业大学在侯汉清教授领导下进行的有

益的研究工作，我想一定会成为我国信息检索自动化发展史册之中浓浓的一笔。

这一类项目，本质上都是情报语言学的研究课题。所以，在研究中必须遵循情报语言学的理论，吸取情报语言学的已有成果，其结论应切合情报语言学的要求。它们只是利用计算机技术作为方法手段来达到研究目的而已，不能过分强调网络环境的特殊性而置情报语言学关于检索效率的基本要求于不顾。计算机技术应当与情报语言学密切结合。侯汉清教授和他的弟子们同时具备这两方面的知识，是顺利地较好地完成这些研究项目的关键。

这八个研究项目，大多采取实验研究法，故其成果具有较大的可信度和易理解性。其中有些项目，难度较大，甚至极难，专著只是作了认真、有益的探索；有些项目，虽然尚有一些不足，但作为中间成果，可在当前信息检索工作中推广应用，在应用中进一步完善。

信息检索自动化的初级阶段已在我国普遍实现。但要晋升一级，扩大自动化过程的范围和提高自动化的水平，当前的研究还属起步，发表的科研成果尚少见，学术研究有待扩大和深入。这部丛书起了很好的开拓作用，为继续研究打下了基础，是研究者很好的学习和参考用书，希望对此感兴趣的读者能从中获益。

张琪玉
2009 年 7 月

序 言

近年来“知识图谱”这一概念在国内悄然兴起，有兴趣者渐多。尽管目前尚没有一个确定的、公认的与“知识图谱”相对应的英文术语（常见英文有 mapping knowledge domain，knowledge mapping，mapping of scientometrics 等），但是，实际上此类研究在国外已有相当的基础和可观的成就。国内学者正是根据国外的这些研究加以总结、抽象后提出了这个概念。

“知识图谱”是在以引文分析为主要内容的文献计量学、科学计量学和信息可视化技术的基础上发展起来的新兴研究领域。其特点是利用各种分析手段，尤其是引文分析、共引分析可视化以图形的方式来直观展示某一学科或某一专业领域的知识结构、发展历史或研究的前沿等状况，以便认识和掌握学科发展方向和运行规律，为科学学、人才学、管理学的研究和决策管理提供依据和参考。

近年来国外的有关研究取得了很大进展，诸如其所提供的图形已经从黑白平面图发展到彩色的立体图形，且有的已能够进行三维动画，引文分析可视化集成系统也开始出现，等等。引文索引的创建者美国著名学者加费尔德博士的 HistCite，目前活跃在国际知识图谱界的美国 Drexel 大学的陈超美博士的 CiteSpace 系统软件等都是有着很大影响的成果。

最近几年随着国内中文引文数据库的建立和不断完善，这方面的研究也取得了不小的进步，除了介绍、引进国外相关成果之外，已有结合国内学科发展绘制知识图谱的尝试和成果出现，但是，毋庸讳言，这方面的研究还不系统、深入，某些本土化的研究成

果还散见于各期刊论文中，专门探讨“知识图谱”和“引文分析可视化”的专著还很鲜见。因此，李运景博士《基于引文分析可视化的知识图谱构建研究》一书的正式出版，可谓正逢其时，难能可贵，恰好能够弥补这一研究领域目前研究之不足，可喜可贺。

李运景博士师从一向治学严谨的侯汉清教授，在读书期间就对知识图谱等研究非常感兴趣，作为她的博士论文的答辩委员，我对她研究态度的认真；论文资料的翔实、论证的充分、结论的合理等有着深刻的印象。本书作为一本论述知识图谱构建的专著，有许多值得一提的鲜明特点。

首先，本书对构建知识图谱的理论与方法进行了系统的阐述，并对其中的步骤与关键技术进行了详细的分析与对比，对有意开展相关研究的读者而言具有很强的指导性。

其次，书中采用实证研究的方法，利用中文引文数据，构建了中国杂交水稻研究这一专业领域的历史发展图和学科结构图，在结合中国某专业领域描述知识结构和发展规律方面具有创新意义。

再次，本书的实证研究采用可视性较强的寻径网络算法和技术进行著者同被引分析，以构建专业学科知识图谱，这在国内也尚属首次。

另外，本书提供了利用三种不同计算方式获得著者同被引数据而生成的三种专业学科知识图谱，并就它们的差别进行了对比分析，从而证实由不同数据统计方法所得到的分析结果存在着差异，为以后的同被引分析提供了有益的借鉴。

信息社会和知识经济的发展，迫切需要利用新的研究方法对知识的生产、组织、利用、管理等进行深入的研究，知识图谱可以为我们提供新的研究视角和新的研究方法。利用引文分析可视化构建知识图谱，把复杂的知识领域通过数据挖掘、信息处理、知识计

量和图形绘制显示出来，以揭示知识领域的动态发展规律，为学科研究提供切实的有价值的参考。一个好的学科知识图谱不仅可以提供对知识状态的透视（显示内部结构），而且还可以帮助我们在科学发现中有所突破。该项研究涉及应用数学、图形学、信息计量学、图书馆学、情报学、计算机科学、科学哲学、科学社会学等多学科理论与技术，需要多个专业领域的研究人员的合作和持久探索。

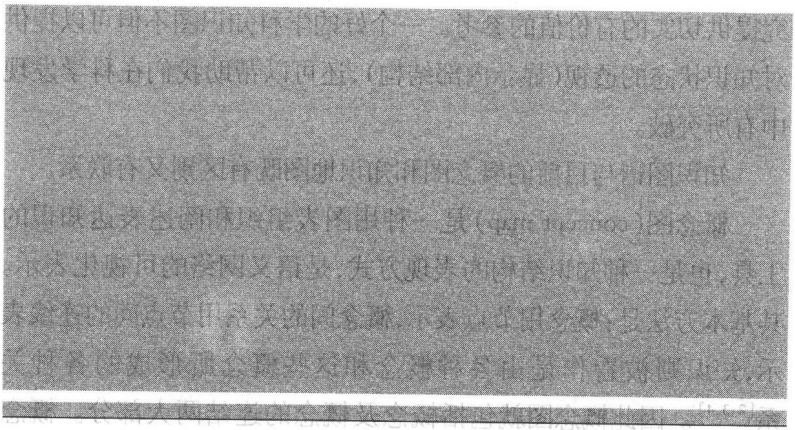
相信本书的出版能够吸引更多的学者关注引文分析可视化和知识图谱的研究，有助于提高人们对各学科知识的有序增长、学科交流规律的认识。同时我也希望李运景博士能够再接再厉，继续关注国际有关研究的发展，密切结合中国学科发展的实际，将知识图谱的研究引向深入，为做出更多有新意的成果，尤其注重产出具有本土原始创新的成果而不断努力。是为序。

叶继元

2009 年 7 月于南京大学

目　次

图表目次



第1章
引　言

1.1　知识图谱、概念图和知识地图

知识图谱(mapping knowledge domain)在图书情报界也称为知识域可视化(knowledge domain visualization)[1]，是显示知识发展进程与结构关系的一系列各种不同的图形。具体来说，它是把应用数学、图形学、信息科学等学科的理论和方法与计量学引文分析、共现分析等方法结合，用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构的多学科融合的一种研究方法。它把复杂的知识领域通过数据挖掘、信息处理、知识计量和图形绘制显示出来，揭示知识领域的动态发展规律，为学科研

究提供切实的有价值的参考。一个好的学科知识图不但可以提供对知识状态的透视(显示内部结构),还可以帮助我们在科学发现中有所突破。

知识图谱与目前的概念图和知识地图既有区别又有联系。

概念图(concept map)是一种用图表组织和阐述表达知识的工具,也是一种知识结构的表现方式,是语义网络的可视化表示。其基本方法是:概念用节点表示,概念间的关系用节点间的连线表示,知识则被看作是由各种概念和这些概念所形成的各种关系[2,3,4]。因此概念图就包括概念及概念的连结两大部分。概念图的理论基础主要是奥苏伯尔(David Ausubel)的认知心理学和有意义学习理论。它的基本思想是由学习者自发地学习概念的属性,通过将新的概念和命题纳入到自己固有的概念中,从而产生学习。概念图可看作是个体认知结构某一层面的形象化表示,它有助于将形象化地表示外物联系的知识结构整合内化到自身的认知结构中。由于概念图能够促进学习,因此概念图主要被尝试用于教学中,目前化学、生物、医学等学科都有应用概念图教学的成功实例。

知识地图(knowledge map),在早期,就是表达科学技术知识或一般知识资源地理分布状况的地图。美国捷运公司最早的知识地图是一张展示知识资源地理分布的美国地图[6],这就是知识地图的雏形。之后,带有索引号或用其他方式表示层次关系的表格和文件,以及用来表示信息资源与各部门或人员之间关系的信息资源管理表和信息资源地理分布图,都是知识地图的早期形式。随着信息技术的迅速发展,知识地图进入了电子时代,在互联网上普遍使用的超文本链接和应用链接就是知识地图的简单形式。这时,很多绘制知识地图的工具应运而生,如 LotusNotes, IBM 的 KnowledgeX 和微软的 Visio 等,它们都是基于数据库来绘制知识

地图，有利于知识地图的动态更新和扩展，并在企业中得到一些应用。知识地图对企业的知识资源进行合理配置，企业绘制的知识地图内容包括两个方面：知识资源目录和目录内各款目之间的关系。知识资源包括企业内部和企业外部资源，一份完整的知识地图包括的内容十分丰富，它能提供知识资源的存储地点、所有权人、有效性、及时性、主题范围、检索权利、存贮媒介及使用渠道等，并能揭示所有的知识资源。随着情报学研究对知识地图的关注，知识地图的概念也就突破了局限于描述知识地理分布的知识地图界限，并逐渐在含义与内容上更加接近学科知识图[5,6,7]。如果以语词或概念为分析单元，通过语义关系构造学科知识图，那么所形成的也就是现在的知识地图。

1.2 知识图谱的发展历程

如果用“以图形的方式展示学科领域发展”来定义知识图谱的话，那么知识图谱最早要归功于 SCI 的创办人加菲尔德。1964 年加菲尔德手工完成了 DNA 领域的引文编年图，随后，1965 年普赖斯运用相同的数据完成了他的经典论文《科学论文网络》。这两个事件可谓是知识图谱绘制的开山之作。后来加菲尔德利用引文时序网络绘制学科发展引文编年图就直接由此发展而来。

1963 年加菲尔德创立的美国费城科学情报研究所（Institute of Scientific Information，简称 ISI，）编辑并正式出版了《科学引文索引》（简称 SCI）。SCI 的设计初衷本是为人们提供一个检索工具，它是从被引文献去检索引用文献的索引，它在被作为一个检索工具提供给人们的同时，由于揭示了科学文献之间、作者之间的引用与被引用的关系，从而提供了引文分析所必需的大量数据。因此，SCI 的问世不但向科学工作者提供了一个强大的追踪科技文献的工具，也极大地促进了引文分析的发展。