MIANXIANG XUEKE
XINXI JICHENG DE LINGYU
FENXI SHUJUJI GOUJIAN

面向学科信息集成的领域分析数据集构建

冯 璐 著

北京邮电大学出版社
www.buptpress.com

面向学科信息集成的领域分析数据集构建

冯　璐　著

北京邮电大学出版社
www. buptpress. com

内 容 简 介

本书创新性地研究了以“领域”为分析对象的数据集构建的理论与方法。从领域分析和领域分析数据集的含义入手，探讨了领域分析数据集构建中涉及的相关理论以及它们在数据集构建中的应用价值。以此为基础，提出领域分析数据集界域理论，旨在从领域范式、分析需求和分析目标三个角度界定领域分析数据集的边界和疆域。结合领域分析数据集界域理论和现有来源数据组织状态，提出了基于界域理论的典型数据映射方法。为保证构建的数据集具有核心性，还研究了在构建各环节中的数据质量控制问题，从定性和定量的角度提出了数据质量控制要点。最后，按照领域分析构建流程，选取具体领域对构建各环节进行了模拟实践验证。

本书适合以文献型数据为分析对象的信息分析和情报研究工作的科研人员阅读，可作为高等院校信息管理与信息系统专业、图书馆学专业、情报学专业、档案学专业等教学参考书，也可供从事战略管理、政策咨询等相关工作的专业人员学习参考。

图书在版编目（CIP）数据

面向学科信息集成的领域分析数据集构建/冯璐著．--北京：北京邮电大学出版社，2013.3
ISBN 978-7-5635-2838-7

Ⅰ.①面… Ⅱ.①冯… Ⅲ.①情报分析—研究 Ⅳ.①G353.1

中国版本图书馆 CIP 数据核字(2011)第 246738 号

书　　名：面向学科信息集成的领域分析数据集构建
著作责任者：冯　璐　著
责任编辑：何芯逸
出版发行：北京邮电大学出版社
社　　址：北京市海淀区西土城路 10 号(邮编：100876)
发 行 部：电话：010-62282185　传真：010-62283578
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：北京源海印刷有限责任公司
开　　本：787 mm×960 mm　1/16
印　　张：9.75
字　　数：184 千字
版　　次：2013 年 3 月第 1 版　2013 年 3 月第 1 次印刷

ISBN 978-7-5635-2838-7　　定　价：29.00 元

前　言

学科情报研究对象随着科学发展的交叉、汇聚、融合，从传统的边界清晰的如物理、化学、数学、地理等领域逐步向新能源、海洋、纳米、人口与环境、现代农业等综合化、横向化、交叉化科技领域发展。这些领域有的是因为学科自身发展导致研究内容的相互交叉融合；有的是因为某项重大研究课题的开展形成专门的研究领域；有的则是由于某个大型设备的研制，触发一个新的研究领域的产生。基于此，学科情报研究必须响应学科发展变化，开展由于学科信息集成而形成的领域的情报研究（以下简称领域分析）。

在整个领域分析过程中，领域分析数据集位于整个领域分析流程的上游，因此，探讨领域分析数据集构建的理论与方法，对于情报研究工作的顺利开展和分析结果具有一定的实际意义。传统的学科情报工作主要是依赖检索类数据库资源进行学科情报分析，这种方式在一定程度上满足了情报分析的数据提供要求，但它不适应分析对象和分析目标的变化，主要表现为数据来源未提供直接的领域分析数据选取途径（入口），领域分析需求与典型的检索提问存在差异，领域属性无法利用现有检索途径描绘等。于是，位于情报分析工作上游的数据集的获取成为领域分析中重点关注和需要解决的问题之一，领域分析数据集构建问题成为领域分析工作的难点。

本书共分 8 章，包含 5 个主体部分，内容包括领域分析数据集构建基本理论、领域分析数据集界域理论、基于界域理论的典型数据的映射、领域分析数据集数据质量控制、领域分析数据集构建的实证分析。

本书由冯璐著。在编写过程中，得到了北京城市学院、中国科学院文献情报中心的大力支持，在此表示衷心的感谢！由于作者水平有限，书中纰漏在所难免，恳请广大读者批评指正。

作　者

目　录

第 1 章　引言 …… 1

1.1　研究背景 …… 2

1.1.1　学科情报研究对象的变化 …… 2

1.1.2　领域分析的必要性 …… 3

1.1.3　领域分析中的数据获取 …… 5

1.2　国内外研究现状 …… 6

1.2.1　国内外研究现状 …… 6

1.2.2　现有问题 …… 14

1.3　研究意义 …… 16

1.4　研究内容及其组织结构 …… 16

1.4.1　研究范围界定 …… 16

1.4.2　研究的主要问题 …… 17

1.4.3　研究的思路与内容 …… 17

1.4.4　研究内容的组织结构 …… 18

第 2 章　领域分析数据集构建相关理论 …… 20

2.1　领域分析 …… 20

2.1.1　软件工程中的“领域分析” …… 20

2.1.2　情报科学中的“领域分析” …… 20

2.1.3　情报研究中的“领域分析” …… 22

2.2　领域分析数据集 …… 24

2.2.1 领域分析数据集的含义 …… 24
2.2.2 领域分析数据集构建研究的关键问题 …… 25
2.3 相关知识在领域分析数据集构建中的应用分析 …… 26
2.3.1 信息检索 …… 26
2.3.2 文献计量 …… 28
2.3.3 资源评价 …… 31
2.4 小结 …… 32

第 3 章 领域分析数据集界域理论构建 …… 33

3.1 基于领域范式演化的领域分析数据集界域 …… 33
3.1.1 领域范式演化 …… 33
3.1.2 基于领域范式演化的领域分析数据集界域 …… 37
3.2 基于领域分析需求和分析目标的领域分析数据集界域 …… 40
3.2.1 领域分析数据集界域思想 …… 40
3.2.2 基于分析需求的领域分析数据集界域 …… 41
3.2.3 基于分析目标的领域分析数据集界域 …… 44
3.3 小结 …… 46

第 4 章 领域分析来源数据组织状态分析 …… 47

4.1 来源数据分布状态 …… 47
4.1.1 来源数据分布状态的类别 …… 47
4.1.2 领域分析来源数据限定 …… 48
4.2 领域分析来源数据提取方式 …… 49
4.2.1 领域分析来源数据提取限定 …… 49
4.2.2 领域分析来源数据提取与信息检索的差异 …… 51
4.2.3 领域分析数据提取来源 …… 54
4.2.4 领域分析来源数据提取方式 …… 56
4.3 领域分析来源数据组织状态分析 …… 61
4.3.1 基于研究主体的分类列表法 …… 61
4.3.2 基于研究资源的组织方法 …… 62
4.3.3 基于主题内容的组织方法 …… 69

4.3.4　基于引用关系的组织方法 …… 77
4.4　小结 …… 77

第 5 章　基于界域理论的典型数据映射方法 …… 79

5.1　典型数据组织状态 …… 79
5.1.1　典型数据分布规律 …… 79
5.1.2　典型数据映射方法 …… 80
5.2　基于领域范式演化的典型数据映射方法 …… 84
5.2.1　单向移植领域的基底领域析出方法 …… 84
5.2.2　双科交融和多元综合领域的融合要素汇聚方法 …… 85
5.3　基于领域分析需求和分析目标的典型数据映射方法 …… 88
5.3.1　时间跨度 …… 88
5.3.2　文献类型 …… 89
5.3.3　被引文献 …… 91
5.4　人工定性判断法 …… 92
5.5　小结 …… 93

第 6 章　领域分析数据集数据质量控制 …… 94

6.1　数据来源质量控制 …… 94
6.1.1　权威性 …… 94
6.1.2　覆盖性 …… 95
6.1.3　客观性 …… 98
6.1.4　准确性 …… 98
6.1.5　接近性 …… 99
6.2　数据映射质量控制 …… 99
6.3　结果数据质量控制 …… 102
6.3.1　典型性 …… 102
6.3.2　主题相关度 …… 103
6.3.3　机构影响度 …… 105
6.4　小结 …… 107

第 7 章　领域分析数据集构建实证分析 …… 108

7.1　实证对象背景 …… 109
7.2　构建策略分析 …… 110
7.2.1　分析对象 …… 110
7.2.2　数据来源 …… 111
7.2.3　选取方法 …… 111
7.3　结果数据质量分析 …… 113
7.3.1　结果定性分析 …… 113
7.3.2　主题相关度 …… 113
7.3.3　机构影响度 …… 116
7.4　数据集扩展与评价 …… 119
7.5　小结 …… 122

第 8 章　结语 …… 123

8.1　本研究的主要工作 …… 123
8.2　本研究的创新之处 …… 124
8.3　本研究的不足 …… 125
8.4　有待进一步研究的问题 …… 125

附录 …… 127

参考文献 …… 141

第1章 引 言

《国家中长期科学和技术发展规划纲要(2006—2020年)》中明确指出要“全面推进中国特色国家创新体系建设”。科技创新已然成为国家发展的基本战略,组成国家创新体系的企业、大学和科研机构等创新单元都在努力提高自身的科技创新能力。

“科技创新,情报先行。”科技情报无疑在科技创新过程中担负着越来越重要的职责,从科研人员的科技知识创新,到产业部门的技术开发,再到国家科技政策战略的制定,都要搜集和掌握国内外有关的科技情报。

当前,由于现代信息技术的飞速发展,科技情报工作充满机遇与挑战。第一,信息技术的发展、信息环境的变化为情报工作提供了许多工具与资源的便利,如数据挖掘与可视化技术为计量学的发展提供了新的思路和灵活的方法,提高了数据处理的效率和质量。第二,面对信息爆炸和情报不足并存的现象,如何有效利用扑面而来的海量信息,解决“信息爆炸更显情报匮乏”之困惑和需求问题。第三,科学知识体系逐步庞大、细化,各类学科之间相互交叉渗透、汇聚融合的频率增加,联系不断增强,这些都使得科技情报研究的对象和需求日益复杂,更凸显了科技情报工作的必要和重要。基于这样的信息环境变化,专业情报人员通过收集、提炼、分析和研究,为政府、企业等决策提供高增值的情报支撑将更有难度。

科技情报工作决策支持作用的有效发挥,有赖于先进的情报研究理论、方法的创新与完善,帮助科技创新活动主体真正做到“知己知彼、百战不殆”。因此,情报研究工作是科技情报工作的重要环节。

情报研究是一项内容广泛的信息加工处理和情报提炼活动。它以大量相关的原生信息为处理对象,通过对原生信息内容的分析、综合或评价,以提炼出对管理、决策等活动有支持作用的情报,为管理、决策等活动服务。情报研究以其针对性、系统性、科学性、预测性等特点在科学决策、研究与开发、市场开拓中发挥着非常重要的作用。

随着科技、经济和信息工作发展到一定阶段,情报研究最早应用于科技领域,科技情

报研究应运而生。自15世纪下半叶起，随着欧洲资本主义制度的纷纷建立，科学领域相继发生了一系列革命。16世纪，文艺复兴运动在欧洲兴起，在创造了资产阶级文学和艺术的同时，也孕育了近代自然科学，陆续出现一些学术组织和学术期刊。但这些科技信息的传播基本上是自发进行的，是科技工作中微不足道的组成部分。18世纪，伴随蒸汽机的发明，迎来了近代史上的第一次技术革命。这一革命的出现密切了自然科学与生产技术间的联系，科技图书、科技期刊等文献日益增多，一些专业性的科技情报机构纷纷建立，但这些机构的主要工作仅仅是停留在初级信息的浓缩加工和编写报道上。

真正意义上的科技情报研究的产生是在第二次世界大战之后，随着科技的深层次发展，学科的微分化和积分化双重发展趋势日益加剧，科技文献呈指数规律急速增长，科技信息的生产与利用之间出现了严重的脱节和矛盾。为解决这些问题，以对科技信息资料内容的深入分析、综合、评价为特色的科技情报研究工作开始脱颖而出。

20世纪中后期以来，信息技术的发展和社会环境的改变，使得科技、经济和社会间的联系更加紧密，情报研究课题日益综合化、复杂化，为了满足社会需求，情报研究服务从传统的科技领域向多样化方向发展。

1.1 研究背景

1.1.1 学科情报研究对象的变化

学科情报研究工作是针对学科，运用科学方法，收集、整理、加工和分析有用的社会信息、科学知识和新的科研成果，有计划、有目的、准确、及时地为学科科研活动提供服务的一种情报活动。学科情报研究对象既有针对单一学科的分析，如物理、化学、数学、地理等，也有随着科学发展的交叉、汇聚、融合，从传统的边界清晰的领域逐步向新能源、海洋、纳米、人口与环境、现代农业等综合化、横向化、交叉化科技领域发展的。

特别是，随着科学的不断发展，科学上的重大突破、新的增长点乃至新学科的产生常常是由于不同学科的彼此交叉、相互渗透而产生的。学科交叉导致知识融合，各种思路相互启迪，多种方法彼此借鉴；多科学研究的协同催生了新知识突变；跨学科研究的群体效应可以为各学科和工程技术的发展提供综合的方法，产生新的知识论、方法论和价值观。学科交叉交融已成为当代科学发展主要特点之一，在这种形势下，世界各国的政府和资助机构、科学组织和大学都把资助学科交叉放在一个重要的战略位置。

从交叉科学的角度，学科的变化包括学科内交叉学科、学科间交叉学科、领域间交叉学科。

学科内交叉主要是指自然科学和社会科学内的各个一级学科中，其下属学科相互作用、相互结合形成的一种领域。例如，数学中的代数几何学、微分几何学等，物理学中的电磁学、电流体力学等，生物学中的微生物遗传学、植物生理学等。这类领域特点是学科交叉出现较早，是学科内分化综合的初级结果。

学科间交叉主要是指自然科学或社会科学中不同的部门学科通过相互作用、相互结合而形成的一种领域。例如，自然科学中的天体物理学、生物化学、光电化学、量子力学、晶体物理学化学等，社会科学中的政治经济学、社会心理学、教育史学等。这类领域特点是学科内众多分支交叉，是学科内进一步分化综合的产物，其中部分较大分支交叉形成的领域分析有据可依，部分较小分支交叉形成的领域就显得比较模糊。

领域间交叉学科主要是指不同领域学科间交叉形成的，这类领域发展迅速，且交叉点前沿性明显高于其他类型的领域，是知识时代迫切需要解决的重大问题，如中国科学院设置的“1＋10”创新基地。“1＋10”科技创新基地的“1”是加强具有明确目标导向的交叉和重大科学前沿基础研究；“10”是建设信息科技创新基地，空间科技创新基地，先进能源科技创新基地，纳米、先进制造与新材料创新基地，人口健康与医药创新基地，先进工业生物技术创新基地，现代农业科技创新基地，生态与环境科技创新基地，资源与海洋科技创新基地。

针对这些极具综合性和前沿性的学科情报研究对象，如何准确把握领域分析需求、描述领域范围、收集整理领域信息等都是在科技情报研究工作中亟待探寻的问题。

1.1.2 领域分析的必要性

长久以来，辅助国家确立科技发展战略一直是情报分析人员的重要任务。实践证明，情报研究在科学决策、研究与开发、市场开拓活动、创新体系建设中，发挥着非常重要的作用。

具体而言，情报研究的作用主要表现在以下几个方面[1]。

第一，2004年，胡锦涛总书记在中国科学院第十二次院士大会、中国工程院第七次院士大会上发表讲话强调：“未来科学技术引发的重大创新，将会推动世界范围内生产力、生产方式以及人们生活方式进一步发生深刻变革，也将会进一步引起全球经济格局的深刻变化和利益格局的重大调整。这个发展趋势，必然对世界经济、科技发展和国际综合国力竞争带来重大影响。在这样的大背景下，如果看清世界科技进步的大势，能够制定

出正确的科技发展战略，奋力跟上科技发展的时代潮流，就可以在未来的发展中进一步把握住机遇、赢得主动。反之，如果没有看清世界科技进步的大势，不能制定出正确的科技发展战略，在全球激烈的科技竞争中落伍了，那就会失去机遇、陷于被动。”现在我们加强科学技术的情报研究工作，就是要认清和把握当今世界科技的发展态势，为制定适合我国国情的科技发展战略提供情报信息保障和创新思维支持。

第二，在现代科学技术的引领和推动下，人类社会正经历着从工业社会向知识社会的快速演进，科学技术不断创造出新的经济增长点，在解决人类可持续发展的一系列重大问题上发挥日益重要的作用。创新越来越成为国家发展的核心驱动力，成为世界各国的战略选择。现在我们加强科学技术的情报研究工作，可以深入了解国外主要创新型国家的创新发展历程，剖析研究相关创新范例，在不同类型国家的创新战略与创新路径选择、创新体制顶层设计和创新政策协调机制、构建有特色的国家创新、卓有成效的保障措施、创新精神与创新文化等方面借鉴学习国外政府支持创新的典型经验，为建设中国特色的创新体系提供深入的科学决策咨询服务。

第三，20 世纪 90 年代以来，国际科研活动的规模与范围日益扩大，合作研究的论文日益增多，许多国家和组织机构的 R&D 计划对外开放，科技评估的国际化已成趋势，跨国公司纷纷在境外（如中国）设立了研发机构，而且投资之高、技术之超前也是空前的，还有越来越多的国家加大了吸引和利用国际化人才的力度。多边的和全球性的合作越来越重要，科技发达国家和国际性组织设置的大科学国际合作计划有 50 余项，主要集中在全球变化、生态环境、生命科学与生物工程、地球系统科学与观测系统、核聚变、空间科学与空间天文学、地面天文学等领域。全球性问题成为国际科技交流与合作研究的重要内容，全球气候变化、环境问题甚至成为国际关系的重要外交议题。在重要的知识领域，已经形成了全球性的知识网络。国际组织在国际科技合作交流中发挥着日益重要的作用。加强科学技术情报研究，将有助于我国科研机构及科技专家了解科研活动的国际化趋势，有助于设计面向全球化时代的科技政策。

第四，全球化和信息化带来的竞争条件的转变，正在迫使政府和公司不断地反思和调整他们在科技发展方面的优先顺序、投资和 R&D 管理情况。许多国家和企业逐渐察觉到未来的竞争优势将依赖高知识内涵的产业，即使是传统低附加值的产业，为了适应瞬息万变的竞争环境，也必须对经济、科技与社会发展各方面的信息进行系统的搜集分类和“情报”式的分析整理。在这样的背景下，利用情报研究作为战略战术决策的基石来促进创新和跨越发展，已经成为国家和企业获得竞争优势的必要途径。

第五，情报研究有助于科学决策能力的提高。情报研究工作主要承担类似智囊机构

的咨询工作，部分承担信息跟踪分析的工作。正确的决策来源于正确的判断，正确的判断来源于通过情报研究提供的全面的信息收集与系统分析，从而达到对客观情况全面而系统的把握。同时，情报研究工作还有助于决策者更新知识，开阔眼界，启发思路，在决策前充分了解有关领域中的已有成果、现状和发展前景，接受新的信息，使自己的思想保持先进性，增强敏锐性和判断能力。

特别是，面对学科之间日益交叉融合，诸如能源、生物、先进制造这样的综合领域，开展领域分析工作的需求日益强烈。这些领域有的是因为学科自身发展导致研究内容的相互交叉融合；有的是因为某项重大研究课题的开展形成专门的研究领域；有的是由于某个大型设备的研制触发一个新的研究领域的产生；有的则是因国家目标导向而形成的战略领域。因此，非常有必要响应学科发展变化，开展面向领域的情报研究（以下简称领域分析）工作。

针对不同用户的需求，通过分析可以确定领域研究的价值、领域新兴热点问题、领域未来的发展方向、同领域的关联目标机构等，从而可以满足制定国家科技发展规划的需要，准确指导国家确定优先发展领域、领域内重点发展内容以及发展某一领域的关键技术。有关领域分析的理论和实践研究，既在宏观层面上满足了国家制定领域发展需求，也在微观层面上指引具体领域有的放矢地开展相应科技攻关。因此，在情报研究工作至关重要和当今科技迅速发展的大背景下，领域分析存在着很强的必要性。

1.1.3 领域分析中的数据获取

从情报研究工作的流程来看，分析使用的数据集位于分析链条的上游，数据集问题直接影响着整个分析过程的进行和分析结果的质量。传统的学科情报工作主要是依赖检索类数据库资源进行学科情报分析。这种方式在一定程度上满足了情报分析的数据提供要求，但它不适应分析对象和分析目标的变化，主要表现为数据来源未提供直接的领域分析数据选取途径（入口）、领域分析需求与典型的检索提问存在差异、领域属性无法利用现有检索途径描绘等。于是，位于情报分析工作上游的数据集的获取成为领域分析中需要重点关注和解决的问题之一，领域分析数据获取问题成为领域分析工作的一大难点。

因此，提出开展领域分析数据集的研究是有必要的。在领域分析数据集的研究中，需要明确影响数据构建的因素，以及弱化这些影响因素需要的相关知识和方法，也需要解决数据来源的选择、数据源中的数据加工等问题，这些问题直接决定着信息分析人员是否能够准确和深入地开展领域分析工作，最终实现领域分析结果质量的提升。

1.2 国内外研究现状

1.2.1 国内外研究现状

1. 领域分析数据集构建理论现状

情报学中有关领域分析的理论研究起源于美国著名情报学家、哥本哈根皇家图书情报学院的赫约兰德(Hjorland)和阿尔布莱奇森(Albrechtsen)。两人全面地阐述了领域分析的历史渊源、理论思想和方法,他们对领域分析的表述是:理解情报学中的情报的最佳途径是研究作为社会劳动分工一部分的知识领域,即话语社群。在不同的领域中,知识组织和结构、合作模式、语言和交流形式、信息系统和相关性标准都是话语社群的工作客体和社会角色的反映。个体的心理、知识、信息需求和主观相关性判断都应该从这种视角来看待[2]。

在领域分析的经典表述中,"领域"与"话语社群"是两个相关联的关键概念。作为领域分析的创始人和主要推动者,赫约兰德认为领域可以是一个学科或学术区域,也可以是与信仰、职业或惯例等相关联的话语社群。话语社群形成的标志则是在某一群体中存在着有序的、由概念结构、制度栅栏和话语场域的管控共同加以结构化的交流过程。赫约兰德的关注重点放在了领域的维度分析上,他指出,领域可以从本体论、认识论和社会学三个维度来认识,其中后两个维度是核心所在,三个维度之间也是相互作用和变化的,领域的研究应考虑本体论、认识论和社会因素之间的复杂交互。

麦(Mai)采用与赫约兰德相似的思路,明确指出领域是指分享共同目标的人类群体。领域的概念要把形式化结构和实际工作与活动结合起来考虑,后一方面跨越学科专业界限而聚焦于人的活动、合作以及共同目的。这表明领域概念与人类活动密切关联,所以领域是以活动为中心的。这种定义实质上是以活动理论作为元理论基础的反映,活动理论强调的活动这一核心概念规制了领域的本质和界限,正是以包括共享目标在内广义上的活动作为标杆,领域的轮廓才得以勾画清楚。因此,可以把领域理解为"以社会中有机联系的共同活动(包括共享目标、任务、合作、交流)为基础,结合专业的形式化结构而形成的群体"。依此,可以把学科和专业看成是具有相似结构或特征的多领域聚类而成的领域簇。

他们对领域分析的研究,主要从认识论和社会认知的角度,强调以领域整体为关注

点，将社会因素融入到对某一个主题领域的知识整理，最终是要解决如何对特殊领域知识进行分类的问题。因此，特殊主题领域知识的分类是他们研究领域分析的主要目标，而领域知识整理只是服务于知识分类的一种获取知识的手段。在对特殊主题领域知识进行整理的过程中，涉及情报科学人员需要从哪些角度了解领域知识，这方面的研究与领域分析数据集构建相关，下面是赫兰约德整理出的11种了解领域知识的方法，具体包括[2-4]以下几种。

- 文献指南或主题式资源指引网站(literature guides and subject gateways)；
- 专门的分类法和索引典(special classification and thesauri)；
- 索引编制和检索的研究(research in indexing and retrieving specilities)；
- 实证的使用者研究(empircal user studies)；
- 文献计量学研究(bibliometrical studies)；
- 历史研究(histroical studies)；
- 文件和类型的研究(document and genre studies)；
- 认识论和批判研究(epistemological and critical studies)；
- 术语研究、专用语言和论述研究(terminologies studies，LSP，discourse studies)；
- 科学传播中结构与机构的研究(studies in structure and institutions in scientific communication)；
- 专业认知与人工智能的领域分析(domain analysis in professional cognition and artifical intelligence)。

这11种领域分析方法在实际的领域分析工作中作为领域分析数据的获取来源或途径已经被广泛地应用，这些方法要么单独使用，要么组合搭配，各个方法之间可以兼容并蓄，不具有排他性。此外，赫兰约德总结的这些方法还将组织、机构、研究团体等“人”的因子加入进来，这也恰恰符合领域分析数据遴选中的同行评议机制。虽然上述11种方法为领域分析数据集构建在理论层面上提供了一种参考，将领域分析实践工作上升到了一定的理论归纳层面，加强了实践与理论的关联。但是，赫兰约德并没有明确给出领域的定义，也没有研究领域的特质，而在实际的领域分析工作中，理解领域分析对象正是领域分析数据集构建的关键，这一点可以在Palmer的文章中得以证实。Palmer认为，领域的动态、成长和演变会导致信息流、信息策略和信息需求的改变[5]。因此，构建代表领域状态的数据集就应该从领域的演化入手，通过揭示领域的演化，进而确定相应改变的信息。而目前有关领域演变与信息之间关系的研究尚显薄弱。

2. 领域分析数据集构建实践方法

由于领域分析是一个实践性和应用性很强的工作，因此，有关领域分析数据集构建

的专题理论研究文章并不多。于是，作者从现有科技情报研究分析报告和学科评估报告入手，对国内外领域分析中数据集构建的具体操作方法进行综述。

从具体的领域分析对象来看，涉及的领域内容既包括传统学科领域，也包括新兴交叉学科领域。如荷兰莱顿大学的科学技术研究中心(Centre for Scientific and Technology Studies，University of Leiden，以下简称 CWTS)关注的主要领域有物理领域[6]、生命科学领域[7]、食品领域[8]等；英国苏塞克斯大学科技政策研究机构(Science and Technology Policy，University of Sussex，以下简称 SPRU)关注的主要领域有能源动力、信息通信、生命科学技术等[9]；加拿大国家研究理事会(National Research Council of Canada，NRC)涉及的领域有能源领域、纳米领域、燃料电池领域[10]等；中国科学院文献情报中心涉及的主要领域包括物理领域[11]、化学领域[12]、能源领域[13]、纳米科技领域[14]、大科学装置领域[15]等。

从现有领域分析数据集构建的数据遴选过程来看，主要是依赖科学文献数据库及其检索途径，同时配合“同行评议”来遴选领域分析的数据。

(1) 数据库及其检索途径是数据集资源和构建方法的基础

长期以来，文献作为人类文明成果记录与传播的重要载体，一直是科学研究工作最直接的体现。因此，在情报研究工作中，一般也都是选取文献作为重要的信息源，把文献作为开展情报研究工作的基础。数据库及其检索途径正是选取文献及其构件作为分析单元，并且不断吸收计算机技术深化基于传统的印刷型文献的情报研究方法，使这些方法向自动化、智能化发展。

目前信息分析人员主要依靠网络学术数据库开展学科情报研究工作。一些大型的网络学术数据库提供了丰富的检索途径和各种分析工具，方便信息分析人员遴选数据。

此方法具有强烈的情报学特色，是情报学中开展情报研究工作中专门的分析方法，主要包括文献计量学方法、引文分析法和内容分析法等。这些方法的研究对象可以是整篇文献、期刊、报纸或专著，也可以是标志文献的外部特征(如篇名、作者、引文、出版社、网站、借阅与复制的情况等)，或是标志文献的内容特征(如概念、词语、关键词等)。图1-1是与数据库检索途径对应的文献构件。获取这些数据的途径可以借助书目、索引、文摘、百科全书、数据库等二次或三次文献，也可以从报纸、期刊或网络上获取发表的原始文献。

以上述文献及其构件为数据源，利用“共现”原理，即相同或不同的文献特征项共同出现的现象，例如，共词、共篇、共引等，均被广泛应用于科技领域的情报研究中。利用这些方法可以描述情报研究对象的现状，概括情报研究对象的发展规律，分析和评价研究

对象，预测他们的发展趋势，并且利用文献之间明显的相关性挖掘更为重要的隐性信息。

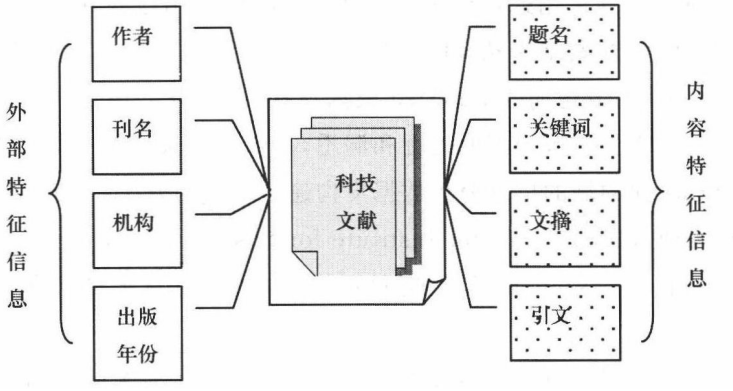

图 1-1 与数据库检索途径对应的文献构件

目前，国内外分析人员通常使用的数据库有 SCI 数据库、INSPEC 数据库、MEDLINE 数据库、中国学术期刊网数据库等。一些相对成熟的领域，如物理、化学、数学等可以借助数据库提供的分类检索途径进行相应的数据遴选和下载。信息分析人员经常利用 SCI 数据库提供的 22 个类目直接下载各类目中的数据构建数据集，但它并未提供更细分的学科类目的检索功能，因而不能用来构建低层级学科领域的数据集。况且对于新兴交叉性领域，涉及的内容相对复杂，就更不能直接采取分类的检索途径来构建数据集，因此，遴选、锁定有代表性的词语、机构和期刊作为数据集构建入口途径是领域分析人员面临的主要问题。

① 基于词语的构建方法。基于词语的构建方法主要是以词语为入口途径在数据库中遴选数据，词语的来源主要有两种。

一种是基于对领域的了解，由研究人员根据数据库的检索要求自行选择代表领域的相关词语。如 Glanzel、Meyer、Noyons 等人都采用 Braun(1997)使用的简单术语描绘纳米领域，“nanoscience”用“nano ＊”表示进行检索。Michael Zitt 等人在进行“纳米”领域数据遴选时，还将纳米的扩展名称、实验方法、关键技术等属性附加到主题中[16]。在对“环境医学”进行领域分析的时候，CWTS 指出将“环境健康”、“职业健康”等词条附加到“环境医学”上进行检索能有效界定目标领域[17]。哥伦比亚大学医学信息学系的 James Cimino 提出利用主题词和非主题词的关系进行文献检索，以扩大检索范围[18]。他们通过分析下载记录的主题词、副主题词，得到主题词和副主题词的频次及组配的关系，使用户可以重新组配检索，利用这样的数据集可以分析某一主题在不同时间段的分布情况，还可以发现相同类型的不同主题间的相似度。之后有很多学者都采用挑选主题词和副