基于特征测度的领域分析文献数据集构建方法研究*

孙巍，黄政，张学福
（中国农业科学院农业信息研究所，北京 100081）

摘要：为构建高度概括学科领域综合特征的领域分析文献数据集，提出一种基于特征测度的领域分析文献数据集构建方法，阐述其实现步骤。以动物资源育种领域主题演化分析为例，考察方法有效性。结果表明，该方法能够立足分析目标与需求，在不影响分析效果的前提下，缩减分析数据量，降低分析成本，提高领域分析效率。

关键词：特征测度；随机抽样；领域分析；文献集构建
中图分类号：TP391
DOI：10.3772/j.issn.1673-2286.2015.12.002

1 引言

随着科学技术的迅猛发展，一方面，科研产出中科技文献量呈级数倍增长[1]；另一方面，各学科领域知识相互交叉、汇聚、融合，导致领域分析用户需求逐步从宏观向中观、微观层次深化。领域分析数据集位于整个领域分析流程的上游，因此，探讨大数据时代领域分析文献数据集构建方法，对于情报研究工作的顺利开展具有一定的实践意义。

现有的领域分析数据集构建方法主要是依赖SCI、INSPEC、MEDLINE、中国学术期刊网等网络学术数据库，利用数据库提供的词语、机构、作者、期刊等检索途径，采用单一或者混合检索策略进行数据遴选[2-6]。也有学者在初次检索的基础上，采用词频统计分析、网络分析等方法扩充检索入口，以此扩大数据集[7-8]。鉴于领域分析需求的日益复杂，数据来源无法提供直接的、更明确细化的领域分析数据选取入口，使得领域分析需求与典型的检索提问存在差异，诸如非成熟或者新兴交叉学科领域的属性无法利用现有检索途径描绘、扩检与缩检方法没能系统地从领域分析需求角度考虑等原因，现有的领域分析数据集构建方法已经无法适应分析对象和目标的变化。因此，如何面向多层次领域分析需求，在不影响领域分析效果的前提下，从海量的数据集中遴选并构建定量文献数据集是当前领域分析研究重点关注和急需解决的问题之一。本文基于文献特征测度手段研究面向领域分析文献数据集的动态构建方法，力图降低领域分析成本，提高分析效率。

2 研究内容与方法

2.1 数据集的界定

冯璐提出了基于需求与目标的领域分析数据集界域思想，阐明了数据集、领域分析需求与分析目标三者之间的互动关系。需求是领域分析数据集构建的起点，直接对构建的数据集提出要求；领域分析数据集直接作用于目标，依据目标来明确数据集的内容；目标是对需求的具体回答，根据实现的目标可以反馈新的需求，由此形成互动循环[9]。因此，构建数据集的首要任务是明确领域分析需求，对分析需求进行具体目标解答，建立目

* 本研究得到中国农业科学院科技创新工程“农业知识组织与知识挖掘团队项目”资助。

表1 领域分析需求、目标、数据集要素、分析内容对应表

领域分析需求		领域分析目标	领域数据集涵盖要素、分析内容	
宏观	从横向角度揭示领域间渗透、扩散、转移等关系，从整体角度剖析领域间的差异	领域产出及影响力分析	要素	可从时间维度宏观揭示文献主题及研究主体的产出数据
			内容	核心机构、人员分析，Top机构、人员分析，核心关键词分析，Top关键词分析等
中观	从纵向角度揭示领域内的产生、发展、现状、趋势等，是对领域内宏观状态的一种描绘	领域内部结构布局描绘与分析	要素	可揭示静动态主题关联、文献关联，静动态国家（机构、作者）合作关系的文献集
			内容	前沿、热点、突发等特定主题探测，合作结构布局分析，主题结构分析等
微观	从微观层面对领域产生、发展、未来趋势等运作机理的分析	领域微观现象描绘及形成发展机理分析	要素	可揭示时序主题关联、时序合作关系等的文献集，强调主题的量及关联关系的变化程度揭示
			内容	主题发展演化现象、规律及机理分析，合作关系发展演化现象、规律及机理分析等

标与数据集间的对应关系，确定数据集所涵盖的要素及内容。从表1可知，宏观、中观、微观分析均需要从时间、主体和主题三个维度提供领域文献数据集，只是分析层次不同，对文献数据集涵盖要素的侧重不同。其中，时间维度通常指构建数据集的时间跨度；主体维度通常指文献的研究主体，如国家、机构、作者等；主题维度通常指文献的主题，一般是基于文献、关键词等来表示。宏观分析数据集侧重领域文献数据量的客观呈现，能够客观揭示领域产出；中观分析数据集侧重于重要特征主体的体现，能够揭示特定主体，完整描绘主题布局、合作布局等；微观分析更侧重重要特征主体属性的细化，用于深度计量分析领域形成与发展的规律、运作机理等。

2.2 文献特征测度

（1）文献特征项与特征

从数据集的界定阐述部分可以看出，领域分析的主要手段是揭示文献特征间的关联，而关联的强度又离不开权威性外部特征的量化。文献特征项是指从结构化数据库中提取的元素，归纳提炼现有相关研究[10-12]，得出领域分析文献数据集特征项主要围绕三类内容，既包括与文献资源、研究主体、资助方或出版方相关的外部特征项，又包括与文献内容相关的内部特征项，详见图1。其中与相关性有关的内容特征项包括：题名、摘要、关键词、学科类别、主题、由于文献的引用关系产生的参考文献的题名、摘要、关键词等；与权威性有关的外部特征项包括：作者、作者所属机构、国家，以及来源期刊、发表时间、基金资助情况、被引频次等。结构化数据库中的文献特征项并不能完整描述文献的某些特征，如：文献的主题特征可以是文献多个特征项的语义提炼。可见，文献的特征源自结构化数据库特征项，但不局限于此特征项。

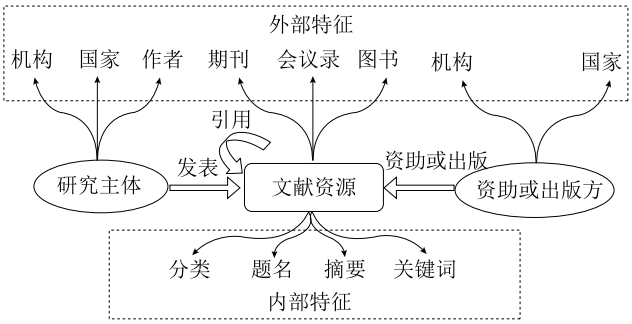

图1 领域分析文献数据集特征项

（2）文献特征测度

表1中，中微观领域分析通常是基于文献的特征来构建领域知识网络，并采用网络计量及可视分析手段对知识网络中特征节点的分布及群聚性、相关性、权威性等进行测度与分析，从而得出领域的发展历史、现状及趋势等结论的过程，是一种定量数据分析。因此，确定构建领域知识网络文献集的数量和内容同等重要，在按需遴选文献特征的基础上，适度缩减文献数据量

是构建领域分析数据集的关键。

确定了文献的主要特征后，需要进一步遴选出主要特征相对突出的文献数据集，才能实现数据集的缩减。文献的外部特征是可确定的，如文献的作者特征、篇名特征等，在给定文献的情况下，外部特征是明确的；而文献的内容特征具有不确定性，如文献的主题特征，应根据主观分析并量化来对其加以测度，一般只能描述一篇文献在相应特征上的近似情况。因此，要遴选出主题特征相对突出的文献子集，需要对主要特征加以测度。

从测度的角度考察文献特征所具有的性质，如图2所示，大致可将文献特征分为两类，第一类是具备序列性的特征，即其特征所具有的性质可以与数字系统相对应，如：文献的出版时间、文献的相关性等；第二类是不严格具备序列性的特征，这类文献特征有学科分类、主题、作者、出版形式等。针对第一类特征中的确定型文献特征，可以直接建立特征集测度函数，测度方法应具备可操作性；而针对不确定型文献特征，可以采用概率测度或模糊测度，例如：文献对主题的相关性，要计算该主题在该文中的频次。针对第二类特征，需要依据特征的局部序列性或采用人为规定方式来建立对应法则，进而建立特征集测度函数[13]。

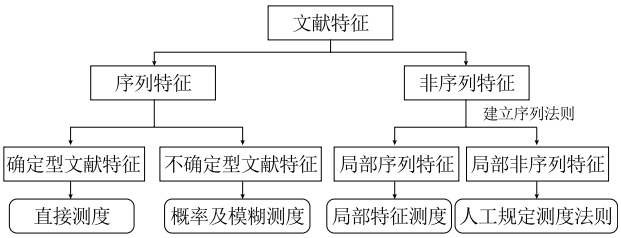

图2 文献特征分类及对应测度方法示意图

按照上述方法，如果选定了符合某种领域分析需求的文献特征C_i(i=1,2,3,…, m)，并基于此建立测度函数，根据这种测度函数便建立了特征与文献间的映射关系，就可以给任意一篇文献指派对应于此C_i特征的值（数字或符号），符合所有特征测度函数的文献就构成了满足领域分析需求的文献集合。

2.3 领域分析文献数据集构建流程

（1）明确领域分析需求，制定领域文献集检索策略

不同层次的领域分析需求，其分析目标存在差异，所依赖的领域分析数据集也各有侧重，不同的数据源也会导致获取的数据集存在差异。此阶段的主要任务

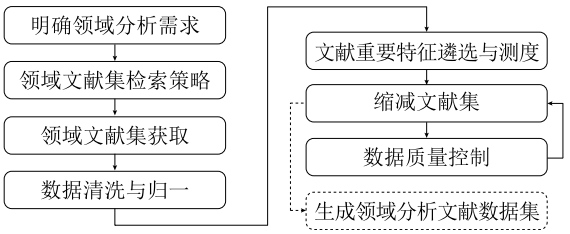

图3 领域分析文献数据集构建流程

是具体解答领域分析需求，明确分析目标和对象，确定领域分析数据集应涵盖的要素，针对不同的数据源制定与涵盖要素相关的检索策略，尽量保证领域数据的全面性。

（2）数据获取及预处理

依据不同的检索策略，分别下载或获取多源数据集。首先对各数据集内数据进行清洗与归一，包括对各文献特征数据的书写错误与内容错误的修正、对重复数据的归一；其次对数据集间数据加以整合，主要指针对多数据库的元数据整合。清洗整合后的数据集用于后续基于文献特征测度的数据集遴选。

（3）文献重要特征遴选与测度，缩减文献数据集

依据领域特点及数据集需要涵盖的要素，按需求描述领域分析文献数据集的特征，明确和突出主要特征。针对各文献特征的类型分别建立测度函数，搭建特征与文献间的映射关系；对所有符合特征测度函数的文献进行去重与归一，依据文献信息分布的集中与离散规律[13]，并结合数据分析处理能力与分析成本，按比例抽样缩减文献数据集，初步形成领域分析文献数据集。

（4）数据集质量控制

可从数据源、遴选的文献特征以及结果数据集三方面考虑，具体包括：数据源要具有权威性、高覆盖面、客观性、准确性等特点[14-15]；重要特征应具备“优先性”，被优先选取的、富含重要特征的文献集应具备核心性，即运用少量文献去涵盖大量信息；结果数据应具备典型性、主题相关性、国家、机构、作者影响度[10]等特征。

3 实证分析

3.1 数据获取

选取动物资源与育种领域作为分析对象，考察基于特征测度的领域分析文献数据集构建方法对该领域

主题演化分析的效果。考虑到动物资源与育种领域是一个相对较小的农业领域研究范畴，同时又考虑实验的复杂度，为了获取全面覆盖该领域综合特征的文献集，我们选择从专业期刊中初步遴选文献集的数据获取策略。以animal、livestock、poultry、zoology、breeding、swine、dairy、sheep、goat、ruminant、poultry、horse、rabbit为检索词，仅对Web of Science的6423种期刊名称及期刊分类进行检索，命中32个期刊（截至2015年9月10日），经语种“English”及专家分析筛选，得到29个“动物资源与育种领域”期刊，批量下载期刊文献，初步获取59,911篇文献。通过数据清洗与归一，进一步析出2000年至2014年间的58,294篇文献。

3.2 动物资源与育种领域主题演化分析数据集构建

（1）数据集要素识别与提取

以动物资源与育种领域主题演化分析作为实例，通过领域时序主题结构分析，从宏观层面了解动物资源与育种领域内主题的产生、发展、变化过程，属于中观领域分析。数据集要素应涵盖一定时间维度的主题关联内容，文献特征具体体现在题名、关键词、摘要、引文信息、出版时间等。相对于题名、摘要等文献的自然语言特征项，关键词特征项的形式较为规范，且具备代表性。因此，本研究优先选取关键词特征项作为面向领域分析的文献集重要特征，结合作者关键词与辅助关键词（机器自动标注关键词），对58,294篇文献的主题特征加以表示，共得出98,615个主题词，分析时间跨度定为15年（2000-2014年）。

（2）文献特征测度

从文献主题特征覆盖面、时间维度的主题分布度、分析成本三方面综合考虑，建立文献主题特征测度函数，具体公式如下：

$$f(K_x)=\text{Round}\left(\frac{S_{K_x}^2}{\sum_1^n K_i}\right)$$

其中，K_x表示文献主题特征K的某一特征值，K_i表示文献主题特征的任一特征值，n为文献集主题特征值总数，S_{K_x}表示具备特征值K_x的文献总量，$S_{K_x}/\sum_1^n K_i$表示具备特征值K_x的文献总量占总文献量的比重。当$f(K_x)\geq 1$时，说明依据特征值K_x能够抽取相应比例的文献子集，也说明该特征值具有一定的核心性，由此共得到32,248个文献特征值，即主题词集。

（3）生成文献数据集

对每个遴选特征值对应的文献集按$S_{K_x}/\sum_1^n K_i$比例随机抽取文献子集，得到与特征值对应的32,248个文献子集，并对其进行去重处理，最终得到包含11,540篇文献的领域分析结果文献数据集。

3.3 数据集质量评价

（1）数据源评价

ISI公司的Web of Science（WOS）数据收录了理工类文献，且其收录的化学、生物、农业、物理领域等文献数据在各类综合数据库中也享有一定的优势，本研究以WOS数据作为数据源，可以从权威性、覆盖面、客观性、准确性几方面保证数据的质量。

（2）文献集特征评价

考虑到关键词的形式及内容的规范性，在具有主题特征的题名、关键词、摘要、引文等文献特征中优先选取作者关键词和机构关键词表示主题特征，数据集的主题特征具备一定的“优先性”。

按文献特征映射函数，最终从98,615个主题词中遴选了前32,248个高频主题词，占总主题词量的32%。按文献信息分布的集中与离散规律，所提取的文献特征值可以充分代表数据集的主题特征，具备一定的“核心性”。

（3）结果数据评价

对于用户的信息需求而言，没有所谓“最好”的数据来源，而是要找到最具有“典型性”的资源。对领域主题演化分析的结果数据而言，考察其典型性主要是指构建的数据集在“广度”和“深度”层面上的主题代表性，即以“少量资源实现大量产出”，形成领域分析数据集。

针对动物资源与育种领域主题演化分析所构建的数据集，从广度上讲，该数据集能够在缩减数据量的条件下涵盖具有重要特征的主题词。如图 4所示，横轴表示主题词的总词频区间，纵轴表示对应词频区间主题词量。特征测度前将主题词按词频排序，并按每段10,000个主题词进行主题分段统计，便形成了横轴的主题词区间。从图4可以看出，有近10,000个词，其词频在[7,5918]区间内。进一步统计发现，特征测度之后的主题词数仅占原有主题词数的32%，但这32%的主题词却涵盖了全部大于等于1的高频词，和部分等于1的高频词，说明数据集具备很高的主题覆盖广度。从深度上讲，在高度概括文献集主题特征的基础上，该文献

集能够保持原有时序文献分布趋势。从图 5可以明显看出，特征测度前后的文献集分布趋势相似，进一步说明此文献数据集从时间维度上可以体现出高主题影响力文献的客观分布，具备一定的文献覆盖深度。

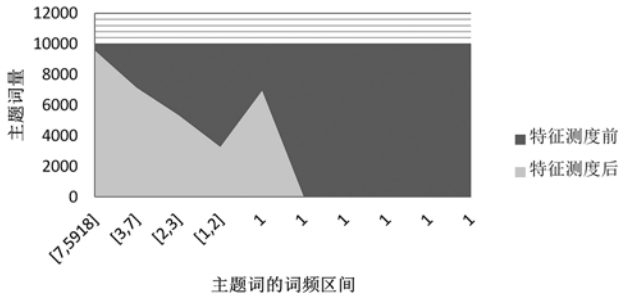

图4 特征测度前后文献主题特征分布对比

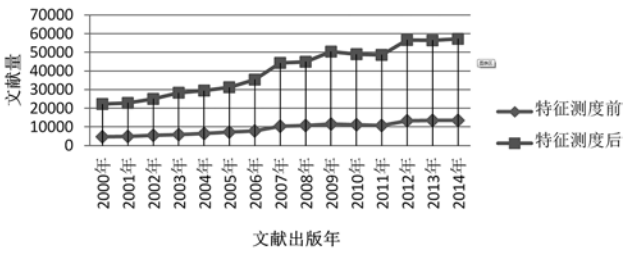

图5 特征测度前后文献时序分布对比

4 结语

本研究提出了基于特征测度的领域分析文献数据集构建方法，并以中观层次的领域主题演化分析为例，从数据源、文献集特征、结果数据三方面对该方法的有效性加以评价。从分析过程及评价结果看，该方法能够立足领域分析需求与目标，动态遴选文献的重要特征，通过特征测度使重要特征具备优先性，同时又建立了特征与文献间的关系，进而得到具备典型性特征的领域分析文献数据集。鉴于时间及资源有限，本文仅利用动物资源与育种一个领域的主题演化分析验证了方法的有效性。针对其他层次分析需求的可用性，以及在动物资源与育种之外领域的应用仍有待进一步研究验证。

参考文献

[1] 刘勘.基于科技文献的知识挖掘[J].图书情报工作,2012(4):5.

[2] Chinchilla Rodriguez.Zaida, Comra-Alvarez.Elena,Herrero-Solana,Victor. A New technique for building maps of large scientific domains based on the co-citation of classes and categories [J]. SCIENTOMETRICS, 2004 (1):129-145.

[3] Maria A.Z, Maria B. A global approach to the study of teams in multidisciplinary research areas through bibliometrics indicators [J]. RESEARCH EVALUATION, 1999(8):111-118.

[4] Boyack KW, Henry S, Richard K. Improving the accuracy of co-citation clustering using full text[J]. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE & TECHNOLOGY, 2013, 64(9): 1759-1767.

[5] Wang X, Xu S, Wang Z, et al. International scientific collaboration of China: collaborating countries, institutions and individuals [J]. SCIENTOMETRICS, 2013, 95(3): 885-894.

[6] 刘伟, 王松俊, 郝继英,等. 基于机构的军事医学文献数据集的构建及分析[C].图书情报工作杂志社第30次学术研讨会,2014.

[7] 丁洁, 王曰芬. 基于特征项的文献共现网络在学术信息检索中的应用[J].图书情报工作, 2014(15):135-141.

[8] 邱均平,马凤.中国高校在建设世界一流大学过程中的进步和问题——基于2011年《世界一流大学与科研机构学科竞争力评价》的分析[J].中国高教研究,2012(01):17-22.

[9] 冯璐, 冷伏海. 基于领域分析需求和目标的领域分析数据集界域研究[J].图书情报工作, 2009,53(24):51-54.

[10] 冯璐.领域分析数据集构建的理论与方法[D].北京:中国科学院文献情报中心, 2007.

[11] 曹艺.面向学术影响力评价的网络学术交流中文献的下载与引用研究[D].南京:南京理工大学, 2012.

[12] 吴云标.计算机文献组织中的文献内容特征测度[J]. 情报科学,2006,18(7):631-632.

[13] 邱均平.信息计量学第七讲:文献信息分布的集中与离散规律——布-齐-洛分布系及理论[J].情报理论与实践,2001(1):77-80.

[14] Tanahashi Y. Developing Web Site Evaluation Criteria: Selection Criteria of Current Web Contents. (Evaluation of Digital Resources)[J]. JOURNAL OF INFORMATION SCIENCE & TECHNOLOGY ASSOCIATION, 2000, 50: 297-300.

[15] Janke RG. Current Contents Connect and PubMed--a comparison of content and currency [J]. HEALTH INFO LIBR J, 2002, 19(4): 230–232.

作者简介

孙巍，女，1978年生，中国农业科学院农业信息研究所副研究员，研究方向：农业知识组织与可视化分析， E-mail：sunwei@caas.cn。
张学福，男，1966年生，中国农业科学院农业信息研究所研究员，研究方向：农业知识组织与可视化分析，通讯作者，E-mail：zhangxf@caas.cn。

Research on Method of Literature Dataset Construction for Domain Analysis Based on Feature Measure

SUN Wei, HUANG Zheng, ZHANG XueFu
(Agricultural Information Institute of CAAS, Beijing 100081, China)

Abstract: In order to extract and construct a reasonable scale of dataset from big literature dataset for domain analysis, a method of literature dataset construction based on the feature measure is proposed, the detail implementation steps are described. The topic evolution analysis of animal breeding and resources is taken as an example and the effectiveness of the method is examined. The evaluation results show that the proposed method can be based on the goals and needs of the analysis, reduce the quantity and cost of data analysis, improve analysis efficiency without affecting the analysis results.

Keywords: Feature Measure; Random Sampling; Domain Analysis; Dataset Construction

（收稿日期：2015-12-08）

书 讯

《科技报告体系构建研究》

为推进我国科技报告制度建设，强化科技报告资源共享服务，贺德方研究员率领中国科学技术信息研究所科技报告研究团队，进行了国家社会科学基金重点项目“中国科技报告资源体系构建”（11ATQ006）研究，并对20多年来中国科学技术信息研究所相关研究和实践进行了归纳、凝练、整理和补充，最终形成了《科技报告体系构建研究》一书。

本书作为国家社会科学基金重点项目的主要研究成果，总结了科技报告产生发展的管理历程、凝练了科技报告制度的建设路径、制订了科技报告资源的整合方案，提出了科技报告体系的构建模式，归纳了科技报告实践的操作过程。本书对各级科技计划管理人员强化科技计划项目过程管理具有借鉴作用，对科研人员撰写高质量科技报告具有指导作用，对各类科研机构做好科技报告呈交、推进科技项目的规范管理和机构知识库建设具有参考价值，对图书信息机构做好科技报告深层次加工和收藏利用具有引导作用，也可供高校信息管理、科技政策与管理等专业研究生学习参考。

《科技报告体系构建研究》于2014年12月由科学技术文献出版社出版，定价78.00元。