

作物科学重点领域研究主题识别方法研究 ——以中美法国家级农业研究机构为例

陈露, 颜蕴, 王婷

(中国农业科学院农业信息研究所, 北京 100081)

摘要: 本文以Web of Science SCI引文数据库与CABI Abstract数据库为数据源, 以中、美、法三个国家级农业科研机构为例, 采用内容词分析、聚类分析以及关键词机构归属度等方法, 着重分析作物遗传与育种的研究主题。结果显示三大机构在重点研究主题上呈现相似的分布格局, 又各具特色: 中国农业科学院在遗传与育种领域更注重基础性研究, 在非生物胁迫等方面研究较为突出; 美国农业研究局青睐于对作物抗性育种、抗虫性等方面的研究; 法国农业科学院注重遗传工程研究。

关键词: 农业科研机构; 作物科学; 内容词分析; 主题分析

中图分类号: TP391

DOI: 10.3772/j.issn.1673-2286.2016.4.005

1 引言

当前阶段世界农业发展局势巨变, 作为农业学科体系中的支柱性学科, 作物科学也将面临机遇与挑战^[1]。因此, 在现今环境下, 为加快推进我国作物科学研究进程, 提升国内作物科学研究水平, 我们不仅需要洞悉和熟知国际作物科学研究的走向与趋势, 而且应认真总结国内作物科学研究工作以及研究成果。本文基于文献计量学理论, 以中、美、法三国的国家级农业科研机构为例, 对作物科学重点研究领域研究主题进行对比分析, 了解作物科学研究热点及前沿动态; 通过对重点研究主题以及研究方向的分析, 揭示研究领域的共性与个性特征以及优劣差距, 为我国作物科学布局、前沿发展等科技政策的制定提供数据支撑。

2 研究方法

为识别3所机构在作物科学重点研究领域的研究主题情况, 通过对比分析, 本文最终选用内容词分析法以及聚类分析法作为学科领域研究主题识别的主要方法。此外, 当前大部分研究都是借助高频词来代表机构研究主题。实际上, 高频词语义通常较为宽泛, 可以在

多个机构同时出现, 但一般情况下科研机构通常会侧重于某些研究方向, 同一关键词对于不同的机构可能侧重点有所不同, 在此类情况下, 继续使用高频关键词对研究机构主题进行对比, 难以发现机构的特色, 也不易区分机构间研究主题的差异。因此, 本文借鉴武汉大学陈果关于机构特色关键词的计算方法^[2], 在统计文献关键词的同时, 引入对每个关键词的机构归属度计算, 最终辅助确定不同机构的研究主题与研究重点。具体的主题分析如图1所示。首先需要提取目标机构关键词, 对关键词词频进行统计分析, 并计算出每个关键词的机构标准归属度; 然后确定高频关键词阈值, 选取高频关键词构建共现矩阵, 基于该矩阵进行聚类分析, 结合之前计算的关键词机构标准归属度, 最终识别出目标机构研究主题情况。

2.1 内容词分析法

内容词是文献中能直接反映文献主题的一类词^[3]。所谓内容词分析, 是指通过对文献的内容词汇进行定量分析, 发现其间关系, 从而找出文献间的内在联系, 并揭示学科结构以及演化态势。内容词分析法优化了引文分析法, 在揭示主题相似度方面, 其针对性更强, 精

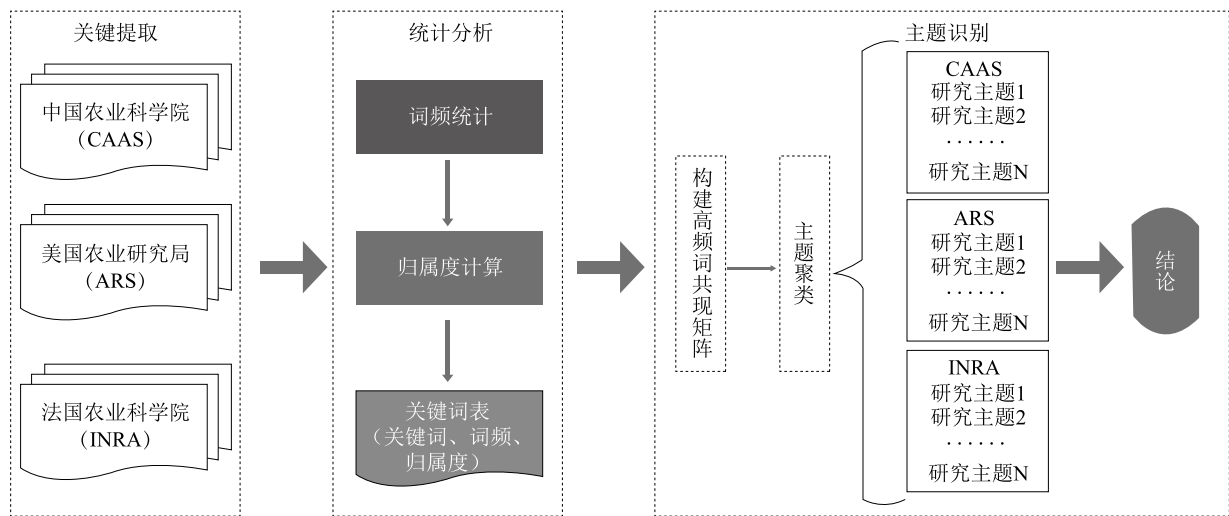


图1 主题分析流程

确性更高^[4]。

共词分析是内容词分析法中被广泛运用的一种^[5]。本文主要将该方法引入作物学科重点领域的研究主题分析,借助经过专业信息标引的文献数据,最终完成对目标机构研究主题的揭示。在对目标机构在作物科学重点领域的文献关键词词频统计分析的基础上,结合其词频的变化规律,确定高频词频阈值。这里根据Donohue J

C提出的高频词低频词界分公式 $T=\frac{(-1+\sqrt{1+8\times I_1})}{8}$

(I_1 指关键词词频为1的个数,T是高频词低频词的边界值)进行确定,或者根据经验判断人工选定,从而划分出高频关键词。本文主要选取3所目标机构每年发文中词频TOP100的关键词作为高频关键词,然后基于高频关键词构建关键词共现矩阵,借助聚类分析方法,将关键词进行划分归类,从而揭示学科领域的研究主题情况。此外,还借助信息可视化技术^[6],更加形象直观地展现分析结果^[7],从而便于研究人员理解该学科领域的主题或者结构情况。

2.2 聚类分析法

聚类分析的目的是将数据集合中各个样本点,按照一定的规则,聚集为不同的类别,使得类与类之间达到最大间距,类内部实现每个样本最小化间距,也就是最相似^[8]。聚类分析属于非监督学习^[9],也就是在不进行样本标注的前提下仅仅依靠统计学原理,建立数学物理模型,优化目标函数,通过迭代算法最终达到寻找最

优解的目的。正因为聚类分析不需要前提知识库的积累即可达到最优化,所以应用范围很广,并且根据聚类的目标函数及优化方法的不同,可分为划分聚类、层次聚类、密度聚类、网格聚类分析等。

本文选定划分方法中的K-Means算法^[10]作为本文关键词聚类的主要方法。该算法简洁、快速,面对大数据集时,可伸缩性强并且高效;面对密集数据时,每个簇形状是凸型的,不同簇之间的差异性较大,聚类效果更明显。因此,本文在构建关键词共现矩阵的基础上,基于K-Means对关键词共现矩阵进行聚类分析,最终识别出3所机构在作物科学重点研究领域的研究主题情况。

2.3 关键词机构归属度

所谓关键词的机构归属度^[2]是指在一定研究范畴的文献集合中,某一关键词总词频内由特定机构所贡献的比例。机构归属度越大,则关键词越能揭示该机构的研究特色。具体公式如下:

$$I(i,j)=\frac{freq(i,j)}{freq(i,all)}$$

其中 $I(i,j)$ 是指关键词*i*在机构*j*中的归属度,而 $freq(i,j)$ 指机构*j*中关键词*i*的词频, $freq(i,all)$ 指在所有机构中关键词*i*的词频。考虑到不同机构的产出规模不同,如果使用上述方法计算关键词归属进行比较,将有失公平。例如,机构A与机构B,A机构的发文规模为100,B为10,关键词K在A中出现10次,在B中出现5次,则K在A中的归属度为 $I_A=10/15>I_B=5/15$,这一结果与实际情况不符,事实是机构B更专注于关键词K。因此,为

了更为公平地比较不同机构的情况,可以根据机构的产出规模,对原先的计算归属度进行标准化处理,具体公式如下:

$$SI(i,j)=\frac{I(i,j)}{doc(j)/doc(all)}$$

其中, $doc(all)$ 指所有机构全部发文量, $doc(j)$ 指机构j的发文量。

$SI(i,j)=1$ 时,说明机构对该主题的研究产出与机构自身规模相应,即该机构在研究中既没有侧重于此主题,也没有削弱此主题。同理, $SI(i,j)>1$ 或 $SI(i,j)<1$,则分别表明机构在该主题的研究上有所强化或弱化。

本文利用该方法对目标机构所有关键词的机构标准归属度进行计算,结合前期聚类结果,挑选出每个聚类簇中归属度较大的一些特色词汇,经过人工判断并结合专家辅助,最终对每一个聚类簇打上主题标签,从而实现研究主题的识别。

3 数据采集与处理

3.1 数据采集

数据采集的完整性与准确性是计量分析的重要前提。由于Web of Science引文数据库自身分类体系并不支持将作物学科相关数据直接提取,因此作物学科数据采集范围的确定是一个难点。本研究数据主要来源于两个数据库: Web of Science (SCI) 数据库与CABI Abstract数据库。Web of Science (SCI) 数据库是三个国家级农业科研机构数据的主要采集来源,考虑到该数据库没有作物科学分类标识,而CABI Abstract数据库是世界权威农业文摘数据库,有全面完善的农业学科分类体系(CABI Classification System),其中,

作物学分类如表1所示,每篇文献都经过专业的信息标引,主题词专业、规范,所以,选择CABI Abstract数据库作为作物科学数据采集的辅助来源。

3.2 数据处理

前期数据处理,主要包括3个阶段:(1)数据检索,从Web of Science数据库采集了2008—2014年3所目标机构发文数据58 006条(中国农业科学院8 570条,美国农业研究局25 636条,法国农业科学院23 800条),以及从CABI Abstract数据库采集了2008—2014年分类代码为FF类的作物科学全部发文数据701 746条;(2)数据预处理,对需要比对的字段进行前期处理,主要包括WOS号提取、DOI/题名特殊字符处理、构建期刊信息唯一标识码等;(3)数据匹配,通过WOS号、DOI号、发文期刊信息(ISSN、卷、期、页码)、人工处理等方式,将上述两个数据库数据进行匹配,并将CABI学科分类代码与叙词表补充进入SCI数据,抽样验证该方法匹配度约为79.6%,远超50%,对于整体的数据具有较高代表性,因此,该方法可行,具体流程如图2所示。通过该匹配方法,最终获得3所目标机构作物科学发文数据23 775条(中国农业科学院3 667条,美国农业研究局12 760条,法国农业科学院7 348条)。

4 分析结果

从作物科学研究领域分布看,全球作物科学研究领域分布(见图3右边)与三国农业研究机构作物科学研究领域分布(见图3左边)大体相似。作物成分(FF040)、作物遗传与育种学(FF020)、作物生产(FF100)、作物虫害(FF620)、作物病毒、细菌和真

表1 作物科学分类表

代码	研究方向	代码	研究方向	代码	研究方向
FF000	作物科学一般问题	FF060	作物生理生化	FF500	杂草和有害植物
FF003	园艺作物	FF061	作物营养学	FF610	作物病毒、细菌和真菌疾病
FF005	大田作物	FF062	作物水分生理	FF620	作物虫害
FF007	饲草和饲料作物	FF100	作物生产	FF700	作物生理病害和伤害
FF020	作物遗传与育种学	FF150	作物耕作制度	FF800	作物毒理学
FF030	作物形态结构	FF160	作物繁殖	FF900	作物环境耐受性
FF040	作物成分	FF170	作物组织培养		

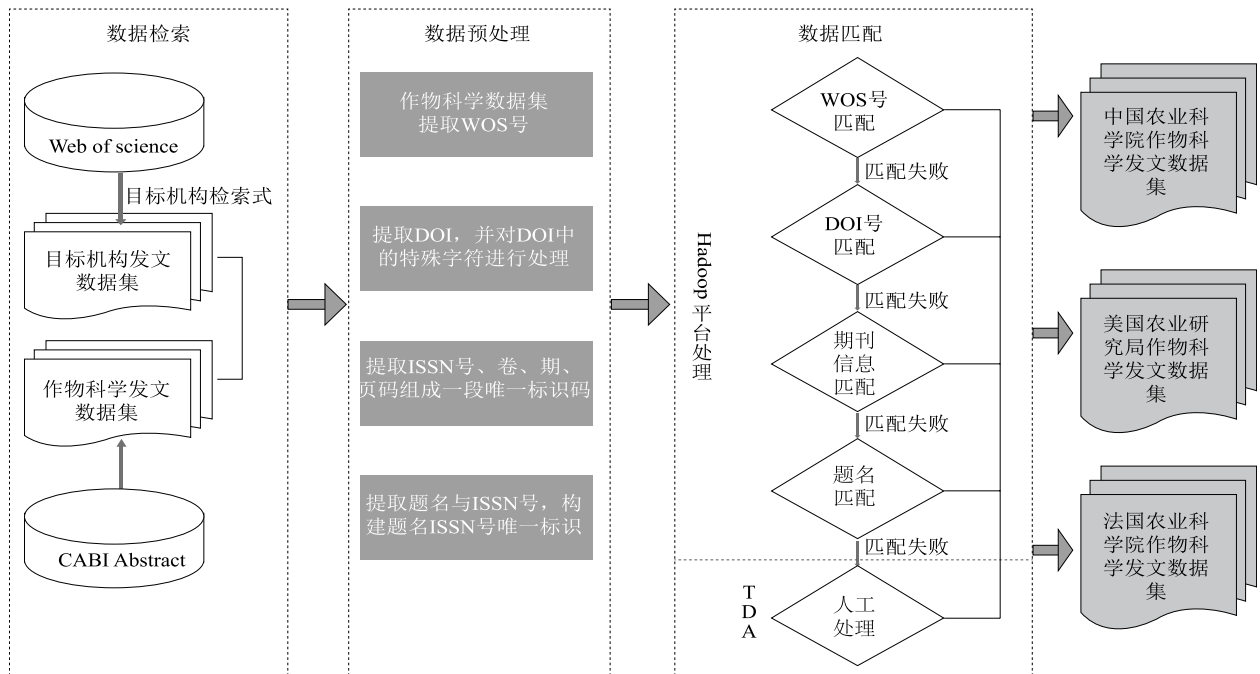


图2 前期数据处理流程图

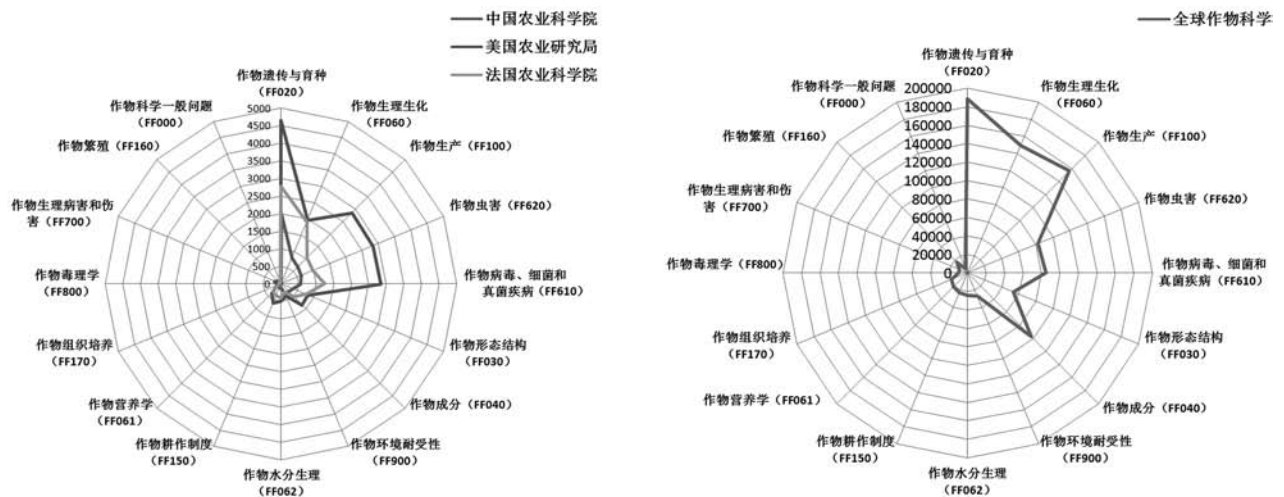


图3 作物科学研究领域分布 (2008—2014年)

菌疾病 (FF610) 以及作物生理生化 (FF060) 领域发文总和分别占中国农业科学院、美国农业研究局和法国农业科学院作物科学全部论文量的76.1%、78.9%和72.8%，因此，这6个领域是三大农业研究机构的主要研究领域。其中，作物遗传与育种学 (FF020) 领域占比最大，既是全球作物科学也是三大农业研究机构的重点研究领域。

结合中美法3所机构在作物科学研究领域分布的结果，本研究着重对中美法3所农业科研机构在作物遗传与育种 (FF020) 领域的主题进行对比分析。本文主要借助文献计量分析工具bibexcel，对前期整理好的目标

机构在该领域的发文数据进行关键词提取，主要选择的是经过专业标引的叙词作为本文分析的关键词，统计目标机构在该领域的关键词词频，并计算每个关键词的机构归属度情况。基于该结果，运用信息可视化工具citespace^[11-13]，对目标机构在该领域内的高频关键词进行共词聚类，结合之前计算的关键词机构归属度，最终识别出目标机构在该研究领域的主题情况，具体结果见图4。

中国农业科学院作物遗传与育种领域的1 983篇文献，时间按照1年进行切分，按词频排序，选取每年TOP100的关键词，去重后获得131个关键词；对这131个关键词进行聚类分析，最终聚成6大研究主题，分别



图4 中美法3所国家级农业科研机构作物遗传与育种领域主题聚类图集 (2008—2014年)

为分子育种、非生物胁迫、种质资源、作物生长发育、生物进化、基因工程、抗病育种。分子育种的主要物种是水稻、玉米；非生物胁迫研究主要是关于作物的耐旱性、耐盐性方面的研究；种质资源侧重于地方品种种质资源建设；作物生长发育，主要对象是大豆，对其生长发育进行调控；生物进化基因工程主要是进行基因测序；抗病育种，主要研究对象是小麦。

美国农业研究局作物遗传与育种领域的4 653篇文献，时间按照1年进行切分，按词频排序，选取每年TOP100的关键词，去重后获得141个关键词；对这141个关键词进行聚类分析，最终被聚成5大主题，分别为分子生物学、抗性育种、分子育种、种质资源、生物进化。分子生物学的主要对象是玉米、水稻、大麦、番茄，其中，大麦的机构归属最高，即对大麦的相关研究重视度最高；抗性育种，主要对象是土豆、小麦、冬小麦，在土豆抗性研究上表现较为突出；此外，在抗虫性研究上优势明显。

法国农业科学院作物遗传与育种领域的2 775篇文献，时间按照1年进行切分，按词频排序，选取每年

TOP100的关键词，去重后获得165个关键词；对这165个关键词进行聚类分析，最终被聚成5大研究主题，分别是分子生物学、分子育种、生物进化与基因组学、抗性育种、遗传工程。分子生物学，优势在于其技术在遗传育种领域的应用；分子育种，基本与其他机构类似；生物进化与基因组学，主要对象是小麦、水稻；抗性育种，主要对象是葡萄，侧重于对其抗病性的研究；遗传工程，主要对象是油菜、玉米，侧重于杂交。

总之，如表2所示，在作物遗传与育种领域的研究主题上，三国农业科研机构既有共性又有各自不同的侧重点。其中，分子育种、抗性育种以及生物进化、基因组学是三个机构的共有研究主题；在非生物胁迫、作物生长发育研究方面，中国农业科学院表现突出；美国农业研究局抗性育种方面是强项，法国农业科学院在遗传工程方面优势明显。从研究对象上看，中国农业科学院在水稻、玉米的分子育种方面表现突出，美国农业研究局侧重于土豆、小麦等抗虫性育种，法国农业科学院倾向于葡萄的抗性育种。

表2 中美法国家级农业科研机构作物遗传与育种领域研究主题对比

研究主题	中国农业科学院	美国农业研究局	法国农业科学院
分子育种	水稻(归属度高)、玉米	无突出物种	无突出物种
分子生物学	—	玉米、水稻、大麦(归属度高)、番茄	分子生物学技术
非生物胁迫	耐旱性、耐盐性	—	—
种质资源	地方品种	无突出物种	—
抗性育种	小麦	土豆(归属度高)、小麦、冬小麦、抗虫性	葡萄
生物进化、基因组学	基因测序	生物进化	小麦、水稻
作物生长发育	大豆	—	—
遗传工程	—	—	油菜、玉米，杂交

5 结语

本文基于文献计量学的理论, 提出运用共词分析法、聚类分析法并结合关键词机构归属度计算进行学科领域研究主题识别的方法。在数据源选择上, 选用Web of Science与CABI Abstract两个数据库, 将两者的数据进行融合, 作为主要的数据来源, 提高了数据的整体质量, 一定程度上避免了由于作者主观因素所造成的偏差。结合该方法, 最终识别出目标机构在作物科学遗传与育种领域的研究主题的情况, 清晰揭示了我国农业科研机构与国外农业科研机构在作物科学重点领域的差异和特点, 为我国农业科研机构作物科学研究布局提供科学定量的依据。

尽管该方法在主题的揭示上准确性高、科学性强, 但其自身仍存在一定的局限性, 一定程度上依赖于对文献数据的标引, 标引准确性越高, 最终主题识别的结果也越好。考虑到当前情况下大部分的文献都未经过专业的文献标引, 因此准确性上将有所降低。此外, 本文选择的词都是高频关键词, 而一些低频关键词可能也是未来的研究重点, 在本研究中没有考虑。最后, 本文在对研究主题分析方面, 并未考虑时间因素对研究主题的影响情况, 而在实际情况下, 学科领域的研究主题在一定程度上会随着时间而改变。因此, 在后续研究中, 可以考虑引入时间因素, 从更全面的角度对研究主题进行探测。

参考文献

- [1] 万建民. 中国作物科学“十一五”的发展方向[J]. 作物杂志, 2006(1): 1-4.
- [2] 陈果. 基于特色关键词的科研机构研究主题揭示: 方法与实证[J]. 图书情报工作, 2014(16): 110-115.
- [3] Braam R R, Moed H F. Mapping of Science by Combined Co-Citation and Word Analysis: Structure Aspects [J]. Journal of The American Society for Information Science, 1991, 42(4): 233-251.
- [4] 葛菲, 谭宗颖. 基于文献计量学的科学结构及其演化的研究方法述评[J]. 情报杂志, 2012, 31(12): 34-39, 50.
- [5] 冯璐, 冷伏海. 共词分析方法理论进展[J]. 中国图书馆学报, 2006, 32(2): 88-92.
- [6] 侯剑华, 陈悦, 王贤文. 基于信息可视化的组织行为领域前沿演进分析[J]. 情报学报, 2009(3): 422-430.
- [7] 侯海燕, 刘则渊, 陈悦. 当代国际科学学研究热点演进趋势知识图谱[J]. 科研管理, 2006, 27(3): 90-96.
- [8] 钟伟金, 李佳, 杨兴菊. 共词分析法研究(三): 共词聚类分析法的原理与特点[J]. 情报杂志, 2008(7): 118-120.
- [9] 王光宏, 蒋平. 数据挖掘综述[J]. 同济大学学报(自然科学版), 2004, 32(2): 246-252.
- [10] 仇新玲. K-均值聚类算法改进及应用[D]. 北京: 北京邮电大学, 2012.
- [11] 李长玲, 翟雪梅. 基于硕士学位论文的我国图书馆学与情报学研究热点分析[J]. 情报科学, 2008(7): 1056-1060.
- [12] 皇甫青红, 华薇娜, 刘艳华等. 国际数字图书馆领域研究热点及作者团体分析: 基于共词分析和社会网络分析[J]. 情报杂志, 2013, 23(1): 118-123.
- [13] Chen Chaomei, Kuljis J. The rising landscape: A visual exploration of superstring revolution in physics [J]. Journal of the American Society for information Science and Technology, 2003, 54(5): 435-446.

作者简介

陈露, 女, 1991年生, 硕士研究生, 研究方向: 信息资源管理, E-mail: luchen0730@sina.com。

颜蕴, 女, 1964年生, 研究馆员, 研究方向: 文献资源建设, 通讯作者, E-mail: yanyun@caas.cn。

王婷, 女, 1968年生, 研究馆员, 研究方向: 文献资源建设、信息计量学, E-mail: wangting@caas.cn。

Method Study on Topic Identification of Main Research Field of Crop Science: A Case Study on National Agricultural Research Institutions of China, USA and France

CHEN Lu, YAN Yun, WANG Ting
(Agricultural Information Institute of CAAS, Beijing 100081, China)

Abstract: Based on Web of Science SCI citation database and CABI Abstract database as the major data source, this paper takes three national agricultural research institutions in China, USA, France as examples, with the method of content words analysis, cluster analysis, institution relevancy of keywords and so on, emphatically analyzes the research topics of the field of plant breeding and genetics. Results show that in the main research topics, this three institutions presents similar distribution, but each has its own characteristics. In the field of plant breeding and genetics, CAAS pays more attention to basic research, especially the abiotic stress research; ARS favors resistance breeding of crops, such as insect resistance, INRA focuses on genetic engineering research.

Keywords: Agricultural Research Institution; Crop Science; Content Words Analysis; Thematic Analysis

(收稿日期: 2016-04-12)