学科交叉热点主题发现与演化分析方法研究
——以动物资源与育种领域为例*

吴蕾，孙巍
（中国农业科学院农业信息研究所，北京 100081）

摘要：为挖掘多学科、跨学科合作交叉点，揭示其发展演化规律，本文以动物资源与育种领域为例，使用学科标识提取农学和遗传学交叉文献集，并进行词语共现分析，得到静态主题聚类；随后综合各个时间片段交叉性主题，对其进行主题动态演化分析，并结合领域专家的分析解读给出解释。研究结果表明，该分析有效挖掘动物资源与育种领域农学和遗传学学科交叉融合所产生的主题及其演化规律，具有一定的可扩展性，可辅助预测新学科领域交叉融合点，并为科学研究人员和高层管理者提供战略决策帮助。

关键词：学科交叉；热点主题发现；演化分析；动物资源与育种
中图分类号：TP391
DOI：10.3772/j.issn.1673-2286.2015.12.003

1 引言

随着现代跨领域科学知识的不断相互渗透，交叉学科的重要性也越来越受到研究者的重视。如何在学科之间寻找重要合作点，从而有效促进多学科、跨学科合作交流，从而解决客观世界复杂问题，已成为科学管理和科技政策制定的一个重要问题[1]。

近年学科交叉研究的热点问题主要集中在环境生态、生命科学、新兴交叉学科的介绍和探索、对跨学科研究本身以及对大科学时代的反思等几个方面[2]。其中生命科学最初是为了解决在物理学和生物学学科交叉发展过程中遇到的问题而产生的一门科学。随着生命科学的发展以及面临问题复杂性的增加，也逐渐融入了数学、化学等多种学科。在动物资源与育种领域问题上，其交叉性主要体现在农学和遗传学上。

文献情报方法是一种发现新知识以及其演化发展规律的有效方法[3]。以领域文献数据集为基础，通过计量学、统计学、数据挖掘与机器学习等方法构建引文、文献耦合、合作者、合作机构等网络，继而可以从宏观和微观两方面对交叉学科知识结构和知识演化给出合理解释。

本文旨在分析交叉学科的热点主题，对其演化发展规律进行分时段分析。在实验阶段以“动物资源与育种”领域农学和遗传学Web of Science核心合集为数据基础，采用词共现网络从静态热点主题和动态演化两个方面进行分析。

2 方法框架介绍

本文提出的研究方法框架分为3个部分，即交叉学科数据选择、静态热点主题发现分析和动态热点主题演化分析。图1为研究方法框架图。

由于不同学科之间的主题划分的颗粒度不同，在分别提取这些交叉学科热点主题时往往无法以相同层级匹配不同学科的主题。因此本文在选择文献数据阶段，从Web of Science核心合集数据库里提取专家挑选的29个动物资源与育种领域期刊的数据，并且以Web of Science中表示学科交叉与融合的学科分类标

* 本研究得到中国农业科学院科技创新工程项目“农业知识组织与知识挖掘团队”和中国博士后科学基金第57批面上资助项目(编号：2015M571183)资助。

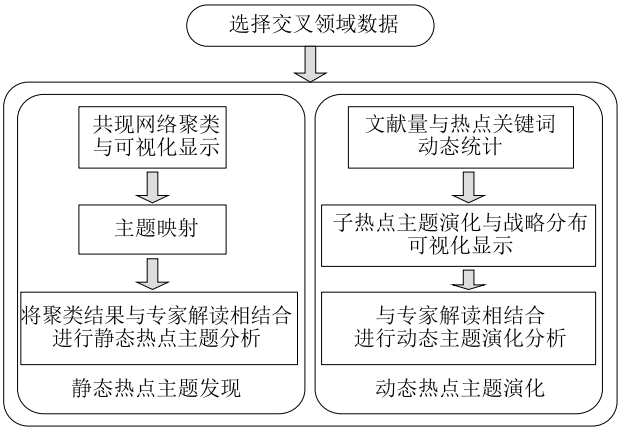

图1 研究路线图

识（Subject Categories，SC）作为提取交叉学科数据的依据。每篇文章拥有一个或多个学科分类标识。如果某两个学科标识在多篇文献中共现，则可以认为这两个学科存在交叉融合。本文根据学科分类标识，选取农学（Agriculture）与遗传学（Genetics & Heredity）的学科交叉文献共2355篇进行分析。

聚焦农学和遗传学两学科展开“动物资源与育种”领域交叉热点主题分析的原因有二：一是，农学作为一种研究范围广泛的学科，与“动物资源与育种”关系十分紧密，据统计，在Web of Science核心文献集中，“动物资源与育种”领域的文章有87.5%的文献涉及农学学科；二是，遗传学研究基因的结构、功能及其变异、传递和表达规律[4]作为生物科学的核心科学，其中的遗传工程和分析遗传学都与农学有着广泛的交叉融合。如基因库、标记基因、连锁遗传等遗传学中的先进科学、技术与专业名词等也广泛存在于“动物资源与育种”领域农学与遗传学的交叉文献中，在“动物资源与育种”领域中扮演着重要的角色[5]。

在静态热点主题发现分析阶段，本文使用连接强度作为衡量共现词相似性的指标，然后将相似度作为词语共现矩阵的元素对高相关性关键词进行聚类[6-7]。连接强度可以表示为下式的形式：

$$S_{ij} = \frac{w_{ij}}{w_i w_j}$$

其中S_{ij}表示单词i和单词j的连接强度，w_{ij}表示两个单词在文章中共现的次数，w_i和w_j分别表示两个词单独出现的次数。本文使用可视化分析工具VOSviewer对静态热点主题进行可视化展示[8]。在主题映射阶段，本文在每个聚类中提取最高频词作为该聚类的标签。最后，本文利用专家判断法对本文的静态热点主题聚类结果以及分析框架进行了有效性和可行性分析。

在动态热点主题演化分析阶段，本文将数据按照时间分为2000-2003、2004-2007、2008-2011、2012-2015四个时间段的数据子集，并对文献量和热点关键词进行分时段动态统计。然后使用简单中心聚类方法对共现矩阵进行聚类，这里同样将连接强度作为共现矩阵的元素[9]，并使用SciMAT对子热点主题动态演化和战略布局进行可视化展示[10]。最后，利用专家判断法对动态主题演化以及分析框架进行有效性和可行性分析。

3 学科交叉静态热点主题发现

3.1 共现网络聚类与可视化显示

主题发现也被称为主题识别、主题抽取。通过对海量、独立的数据对象进行聚类，从中识别有用的语义信息并发现特定的主题领域，是主题发现的目标[11-12]。图2所示的网络是由VOSviewer产生的表示词语共现关系的网络。网络中的节点表示提取的关键词或关键短语，节点越大表示该关键词的频率越大。例如，聚类1中节点polymorphism出现597次，节点sequence出现326次。节点之间的连线表示共现关系，节点之间的距离越近表示共现关系越强。由节点之间的位置关系可见，热点关键词可以被划分为如图的4个聚类。

对比了100多个遗传学高频词[5]，共现网络中的polymorphism（多态性）、gene（基因）、mitochondria（线粒体）、cell（细胞）、receptor（受体）、chromosome

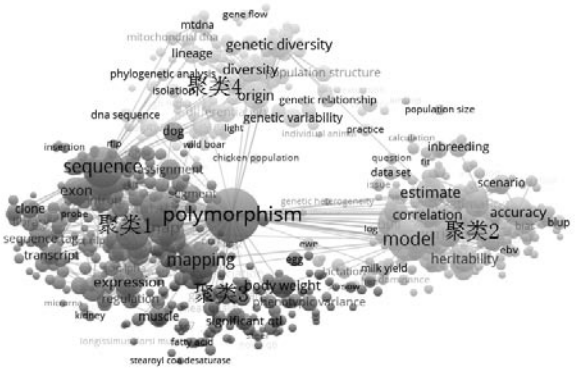

图2 “动物资源与育种”领域农学与遗传学交叉文献的热点关键词共现网络

（染色体）等词同样出现在遗传学领域的高频词中。可见“动物资源与育种”领域的农学与遗传学文献的研究者们用这些遗传学领域的科学方法、研究对象、实验技术等来分析农学领域的科学问题。

图2中聚类1、3、4存在相邻关系，因此有些词如chromosome（染色体）在这三个类中都普遍存在。只是三个聚类关注的方面不同。在聚类1中染色体主要以bacterial artificial chromosome（细菌人工染色体）、bovine chromosome/cattle chromosome（牛染色体）、ovine chromosome（绵羊染色体）等为主。在第3类中以鸡、猪染色体的形式存在。在第4类中以y染色体的形式存在。另外，图2中某些高频关键词标签被其周围拥有更高词频的节点标签掩盖，因此在可视化过程中无法显示。比如主题聚类1中，mutation（突变）一次的词频是256，但是由于其在polymorphism（多态性）一词附近，因此在全局可视图上就被polymorphism（多态性）的节点覆盖。然而，kifney（肾脏）的词频仅仅为13，但是由于其周围没有频率大的节点，所以被突出显示。

3.2 主题映射

经过对聚类主题进行映射，并经过动物资源与育种领域专家验证，可以得到4个聚类主题。表1中给出了4个聚类映射的主题以及其包含的热点关键词。图3给出了4个主题中热点关键词的3种测度指标值，分别为频次、Callon中心度和Callon密度，可表示成如下形式：

$$\text{频次}=w_i$$

$$\text{Callon中心度}=\frac{\sum_{i\in\phi_i,\,j\in(\phi-\phi_i)} w_{ij}}{N-n}$$

$$\text{Callon密度}=\frac{\sum_{i,\,j\in\phi_i,(i\neq j)} w_{ij}}{n-1}$$

其中w_i表示关键词i出现的次数，w_{ij}表示关键词i和j在共现的次数，ϕ_i表示关键词i所属的主题聚类，Φ表示主题聚类的全集。由上述等式可知频次指当前关键词在文献数据集中出现的次数，Callon中心度描述了当前关键词与其非同一主题关键词共现的比重。Callon密度描述了当前关键词与其同主题关键词共现的比重。本文给出的Callon中心度值和密度值已经做归一化处理，即最大值为1，最小值为0，值越大表示其与其他关键词的共现情况越多。Callon密度较大的关键词其临近节点较多，当前关键词与这些临近节点共现也较多。Callon中心度较大具有更大的全局重要性，说明其与其他关键词的关系也很紧密。只有3个指标都比较大的关键词才能够更有代表性的体现当前聚类的主题。图3横轴给出的关键词顺序是按照3个指标的综合排序排列的，即在每个主题中最左边关键词的3个指标综合排序最高，越往右其关键词的综合排序越低。

表1 4个热点主题及其关键词

资源类型	资源内容
1.农业动物基因结构与其对应生物性状之间关系的分子遗传学研究	polymorphism（多态性）、expression（表达）、sequence（序列）、protein（蛋白质）、mapping（图谱、定位）、exon（外显子）、linkage map（连锁图谱）、intron（内含子）、receptor（受体）、regulation（调控）
2. 使用统计模型等计算方法对经济性状育种进行估计	heritability（遗传力）、genetic correlation（遗传相关）、model（模型）、breeding value（育种价值）、Gibbs sampling（吉布斯抽样）、variance（方差分量）
3. 经济性状与基因遗传关联的研究	linkage analysis（连锁分析）、body weight（体重）、significant association（显著关联）、mutation（突变）、significant qtl（重要qtl）、genome scan（基因组扫描）、muscle（肌肉）、meat quality（肉质）
4. 从时间、地区等多种维度对遗传起源、分类、地域差异等进行研究群体遗传学研究	heterozygosity（杂合性）、genetic diversity（遗传多样性）、genetic variability（遗传变异性）、differentiation（分化）、genetic distance（遗传距离）、origin（起源）

3.3 静态热点主题分析

通过对表1、图2和图3所示的热点关键词进行分析并进行热点主题映射，可以得出以下对各热点主题的客观解释。热点主题1对农业动物基因结构与其对应生物性状之间关系的分子遗传学研究，反映了基因分子层面的特性，即哪些基因与哪些性状相关。动物资源与育种技术研究的发展经历群体遗传学、数量遗传学、分子遗传学等研究，分子遗传学时当前研究的热点问题。

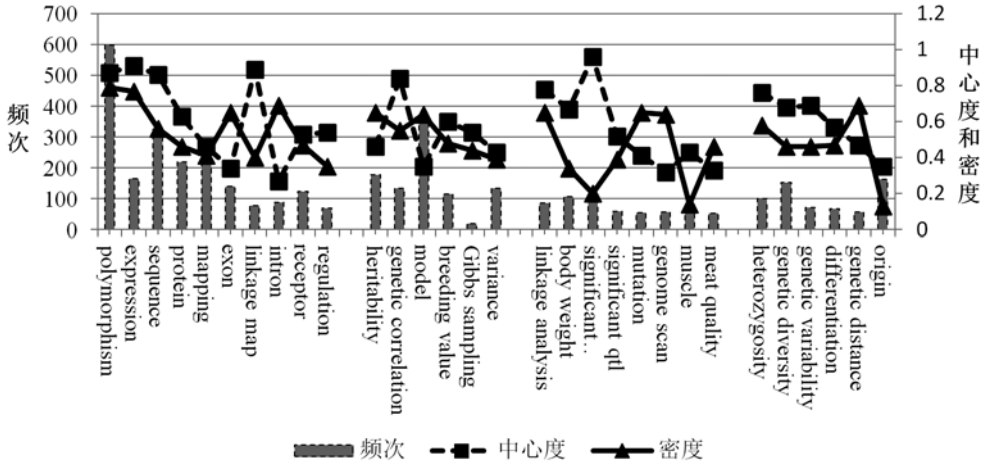

图3 4个主题热点关键词测度指标值

对于热点主题2使用统计模型等计算方法对经济性状育种进行估计，经济性状育种估计是动物资源与育种研究目的和应用基础，而在当前的大数据背景下准确测定遗传性能变得更加重要。用统计模型对遗传性能进行研究是重要的育种手段。本聚类热点主题包括遗传学模型与分子遗传学结合的育种值估计。热点主题3经济性状与基因遗传关联的研究，侧重关注经济性状。农业动物的经济性状是重要的研究对象，只有知道控制性状的遗传基础才能进行有目的的育种，并且获得经济性状的遗传基础差异是开展育种工作的前提。针对热点主题4从时间、地区等多种维度对遗传起源、分类、地域差异等进行研究群体遗传学研究，由于高通量测序技术的发展，从全基因组角度来研究不同纬度和地理分布的家畜驯化和品种的形成成为可能，也可以利用其来研究不同品种家畜的地域和风俗差异的产出品种特征。而从群体遗传学的层面来研究全球范围内动物遗传资源的多样性，既能为种质资源的保存和利用提供依据，也能为优良经济性状的聚合育种发现素材。通过动物资源与育种领域专家验证，以上分析具有可靠性和有效性。

4 学科交叉动态热点主题演化

4.1 文献量与热点关键词动态统计

图4分别给出了4个时间段上动物资源与育种领域农学与遗传学的交叉文献发文量。除了2004-2007年发文量略有下降，其余时间段上的发文量均呈现上升趋势。可见动物资源与育种领域农学与遗传学的交叉研究正处于发展上升的阶段，仍有大量农业遗传研究热点问题等待研究者们发掘。

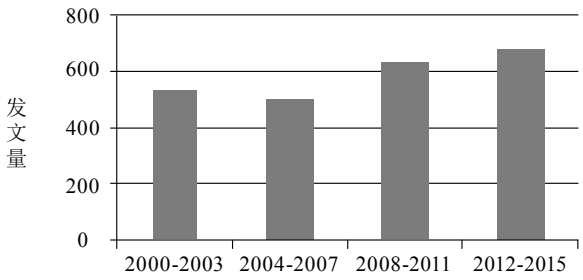

图4 4个时间段中的发表文献数量

图5给出了热点关键词的时间分布情况。从图中颜色分布情况可以看出，左边方框标出的关键词对应的时间段偏早。可以知道对于基因克隆、基因转录等研究很早就开始有研究者关注，但是近些年热度减小，因此其发表的平均时间也相对较早。图中右边方框标出的关键词对应的时间偏后，即这些关键词是近些年新兴的，可以初步断定在“动物资源与育种”领域农学和遗传学交叉融合的文献中，使用统计模型等计算方法对经济性状育种进行估计的研究近年逐年增多。

4.2 子热点主题演化与战略分布可视化显示

图6给出了SciMAT展示的子热点主题时序演化趋势。节点越大表示在该时间段出现该子热点主题的文章越多。图中的实线表示子热点主题主流的演化方向，虚线表示其支流的演化方向。由图可见越往后的时间

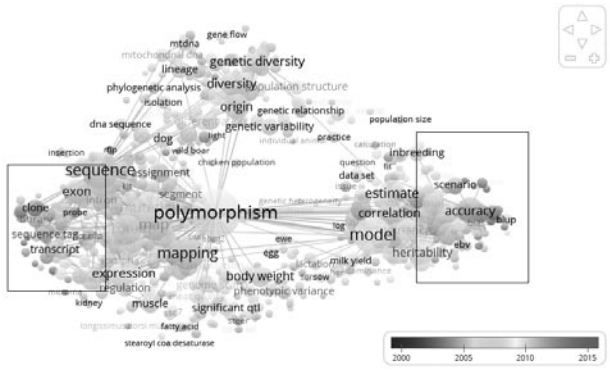

图5 热点关键词的时间分布

片段，子热点主题的个数越多。可知随着时间的变化，动物资源与育种领域的交叉研究主题日益增多并且愈加深入、细化，并呈现出主题之间联系越来越密切的现象。图7给出了SciMAT展示的4个时间段子热点主题的战略坐标分布图。图中横坐标表示Callon中心度，纵坐标表示Callon密度。节点上标的数字代表出现该主题中词语的文章数。第一象限主题的Callon中心度和密度都很高，说明这个区域的研究具有非常好的发展环境，是引擎类主题；第二象限表示Callon中心度略低，但是Callon密度较高，说明这个区域的主题具有特色性，即一般单独对其进行研究；第三象限表示Callon中心度和密度都略低，说明这个区域的主题要么属于新兴主题，要么是衰退主题或者需要与其它主题相结合研究的主题，即提到这些主题的几率相对下降；第四象限的Callon密度略低，但是Callon中心度较高，说明这个区域的主题具有较强的横向关联性，即在进行其他研究时仍然需要经常用到这些主题。

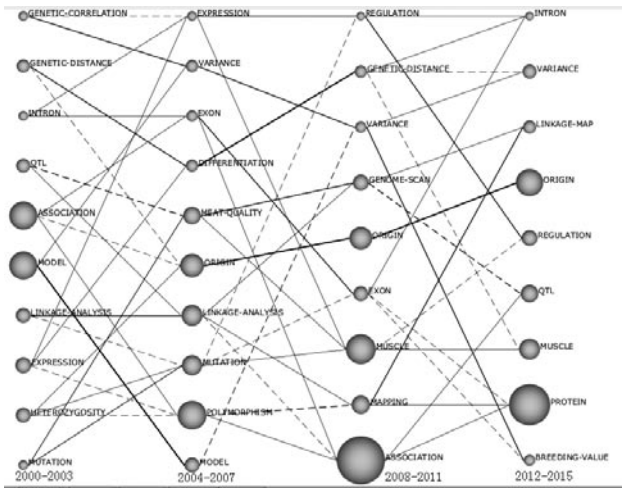

图6 子热点主题随时间演化趋势

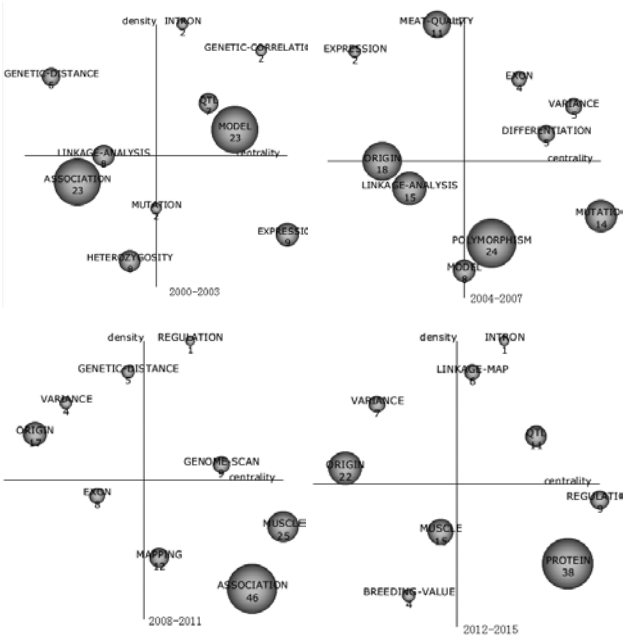

图7 2000-2015时间段子热点主题的战略分布图

4.3 动态热点主题演化分析

通过对图6和图7进行联合分析，可知随着大规模全基因组测序的普及，对于单个基因的克隆和转录分析研究逐渐转向对基因结构与其对应生物性状之间关系的分析研究；近年分子技术与数量遗传学形成新的结合点，促使通过统计模型对经济性状育种进行估计的研究成为现今重要的研究方向；获得经济性状与基因及其更精细变异的遗传关联是进行有效选育的基础。不同生物性状表现出不同特征是由基因控制的，基因和性状之间必然存在着相关。分子技术的发展促进了对经济性状与基因遗传关联的研究的发展。另外由于DNA序列较为稳定，通过遗传学技术定位出来的性状更为准确，这也促进了对经济性状与基因遗传关联研究的发展，使其越来越受到研究者们的重视；世界各国家禽选育形成了巨大而宝贵的遗传资源，对这些遗传资源系统、全面地认识是不可或缺的工作。随着分子生物学和基因组等技术的深入发展，遗传资源的挖掘越来越深。另外遗传起源和多样性的研究主要取决于采集的样本，目前大多家养动物的起源和多样性问题均被研究过，后期要结合新样品或者新地理区域样品等满足当代生产发展需求的样品进行研究；连锁分析主要是分子标记等与性状关联的研究。随着全基因组连锁分析的发展，高通量测序手段不断推陈出新，对连锁分析

的研究主要与分子生物的研究一起进行；遗传突变的发现与连锁分析的发展几乎同步，目前更注重的是找到突变之后证实这个突变的真实性，以及与性状的关联性，如何加以运用；对动物基因定位的研究形成了大量家畜禽遗传资源群体，并对一系列重要的经济性状候选基因进行了染色体定位。由于测序技术的发展和全基因组监测成本的降低，使得借助全基因组进行遗传关联分析定位重要基因逐渐升温。通过专家验证，以上分析具有可靠性和有效性。

5 结语

本文提出的分析框架分析了交叉学科的热点主题及其演化进程，并以动物资源与育种领域中的农学和遗传学为例进行了实验验证。首先利用学科标示识别出具有交叉性的核心文献集；然后通过对文献集进行词语共现分析，得到静态主题聚类；最后综合各个时间片段交叉性主题，对其进行动态演化分析。从学科交叉性角度进行主题发现和演化分析，并对其进行定量和定性评价，为交叉学科领域主题分析提供支持。所有结论都经过动物资源与育种领域专家的分析和解读，具有客观合理性，并且经过领域专家验证本文提出的分析框架可靠、有效。目前有待解决的问题是对交叉性学科颗粒度的定义以及对不同颗粒度匹配，这将是我们以后研究的重点。

参考文献

[1] 许海云,刘春江,雷炳旭,等. 学科交叉的测度、可视化研究及应用——一个情报学文献计量研究案例[J]. 图书情报工作,2014,58(12):95-101.

[2] 程妍. 国外交叉学科研究现状分析——基于学术期刊的视角[J]. 学术界,2014(2): 204-211.

[3] 田丽,杨晋. 竞争情报职业教育体系构建——以辽宁省为例[J]. 图书馆学刊,2013 (11):9-10.

[4] 全国科学技术名词审定委员会. 遗传学名词[M]. 北京:科学出版社,2006.

[5] 傅松滨. 2014年主题词索引[J]. 国际遗传学杂志,2014,37(6):329-330.

[6] Van Eck N J,Waltman L. Bibliometric Mapping of The Computational Intelligence Field[J]. International Journal of Uncertainty,2007, 15(5):625–645.

[7] Van Eck N J,Waltman L,Van den Berg J,et al. Visualizing The Computational Intelligence Field[J]. IEEE Computational Intelligence Magazine,2006,1(4):6–10.

[8] Van Eck N J, Waltman L. Software Survey:VOSviewer, a Computer Program for Bibliometric Mapping[J]. Scientometrics. 2010, 84(2): 523–538.

[9] Cobo M J, López-Herrera A G, Herrera-Viedma E, et al. An Approach for Detecting, Quantifying, and Visualizing the Evolution of a Research Field: A Practical Application to The Fuzzy Sets Theory Field[J].Journal of Informetrics,2011,5(1): 146-166.

[10] Cobo M J, López-Herrera A G, Herrera-Viedma E, et al. SciMAT: A New Science Mapping Analysis Software Tool[J].Journal of the American Society for Information Science and Technology,2012,63 (8):1609-1630.

[11] Xu G, Qiu L, Liu H. Study on Hot Topic Discovery from Chinese Texts[J]. Journal of Digital Information Management, 2014,12(4):267.

[12] 叶春蕾,冷伏海. 基于社会网络分析的技术主题演化方法研究[J]. 情报理论与实践. 2014,37(1):126-130.

作者简介

吴蕾，女，1985年生，中国农业科学院农业信息研究所博士后，研究方向：知识分析、主题发现，E-mail：girlrable@126.com。
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，通讯作者，E-mail：sunwei@caas.cn。

Study of Interdisciplinary Hot Topic Discovery and Evolution Research: A Case of Animal Resources and Breeding Field

WU Lei, SUN Wei
(Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing 100081, China)

Abstract: In order to discover the interdisciplinary and multidisciplinary cooperation intersection, and reveal the law of development and evolution. The paper takes the field of animal genetics breeding as an example to extract agriculture and genetics documents dataset with the subject categories, analyses the co-occurrence of words, and clusters the static topics. At the same time, the paper analyses and explains the dynamic evolution process of topics during different periods with the help of experts in the field of animal genetics breeding. The results show that the analysis discovers cross topics and the law of evolution in the field of animal genetics breeding. The method can be extended to other researches. The interpretation can predict new intersections, and assist scientific researchers and managers in making strategic decisions.

Keywords: Interdisciplinary; Hot Topic Discovery; Evolutionary Analysis; Animal Genetics Breeding

（收稿日期：2015-12-08）