基于LDA-WLC的图情领域学科交叉主题演化分析*

石栖[1,2] 胡正银[1,2,3] 王莉晓[1,4]
(1. 中国科学院大学经济与管理学院，北京 100190；2. 中国科学院成都文献情报中心，成都 610041；
3. 四川省科技信息智能挖掘与应用工程研究中心，成都 610041；4. 中国科学院武汉文献情报中心，武汉 430071)

摘要：学科交叉主题演化分析可深入揭示交叉主题产生和发展的过程、动力与未来发展趋势，推动学科建设与发展。首先，以图情领域中文核心期刊论文为数据来源，通过引文分析遴选该领域所有的热点交叉学科。其次，利用LDA主题模型和WLC决策规则进行主题挖掘与主题过滤，从主题强度与主题内容两个方面进行学科交叉主题演化分析，研究发现：图情领域学科交叉主题的主题强度与技术更新频率呈现正相关关系；主题内容演化均以特定年份为分界点，呈现出复杂化、多样化的特点。最后，提出对图情领域学科交叉融合发展的几点建议。

关键词：学科交叉；主题挖掘；主题过滤；主题演化；LDA
中图分类号：G353.1 DOI：10.3772/j.issn.1673-2286.2023.05.007
引文格式：石栖，胡正银，王莉晓. 基于LDA-WLC的图情领域学科交叉主题演化分析[J]. 数字图书馆论坛，2023 (5)：54-63.

随着开放科学与数据密集型科学研究范式的兴起，要解决和突破重大科技问题，通常需要跨越传统学科的界限，融合多学科领域知识，学科交叉成为推动原始创新的重要方式[1]。学科交叉容易产生新的科学生长点、新的科学前沿，通过学科交叉主题识别，可从微观角度更加准确、清晰地挖掘生长点，反映不同学科共同关注的科学问题和内在关联[1-2]。同时，学科交叉主题演化分析可深入揭示交叉主题产生和发展的过程、动力与未来发展趋势，服务科学决策，推动学科建设与发展[2]。随着学科交叉发展，知识结构、知识脉络与知识关联也越来越复杂，如何识别学科交叉主题，追踪其演化过程，已成为图书情报学科的重要研究内容。

本文对图情领域的学科交叉主题进行识别与演化分析，首先通过统计引文来源期刊所属学科类别遴选该领域所有的热点交叉学科，其次利用LDA（Latent Dirichlet Allocation）主题模型和WLC（Weighted Linear Combination）决策规则进行主题挖掘与主题过滤，最后结合关联规则从主题强度与主题内容两个方面进行多学科交叉主题演化分析。本文研究意义有以下三点。①揭示图情领域的学科交叉分布。通过学科交叉主题识别，可以归纳总结图情领域的交叉学科及其分布状况，分析促进图情领域发展的热点交叉学科[3]。②有助于识别图情领域潜在的跨学科合作机会。在当前的学科分类体系下，不同学科之间依然存在一定的界限，而通过学科交叉主题分析，可以识别可能的跨学科合作机会，促进学科之间的交流[3-4]。③有助于揭示图情领域学科交叉特点。学科交叉研究可以揭示图情领域的横向与纵向发展规律，通过吸收学科交叉知识体系与研究方法，可以全面分析图情领域学科交叉现状，推动学科交叉融合发展[5]。

收稿日期：2023-04-13
*本研究得到中国科学院“十四五”文献情报能力建设专项任务“建设多学科特色数据关联服务系统”（编号：E1290208）资助。

1 相关研究介绍

1.1 图情领域学科交叉研究

图情领域的学科交叉研究主要包括识别领域的学科交叉主题[5-7]、分析主题演化规律[7-8]以及预测主题[9-10]。在学科交叉主题识别上，王连喜等[6]基于共词聚类和LDA主题模型对图情学科和新闻传播学科热点主题进行识别，研究了这2个学科在网络舆情方面的研究共性和差异性。在主题演化规律分析上，隗玲等[7]获取了情报学论文的高频关键词，通过构造共词矩阵，利用网络社区演化工具得到学科主题演化网络图，对学科主题演化过程进行分析。Chang等[8]利用3种以引文分析为基础的文献计量方法研究了图书情报学科的学科交叉演变过程。在主题预测上，Figuerola等[9]利用LDA模型识别图情学科交叉主题，并分析了这些主题代表的交叉学科的演变过程及相互作用，预测了图情学科未来的新技术和创新研究方向。李长玲等[10]通过关键词时序聚类，识别情报学与计算机科学的学科交叉主题并预测2个学科的潜在交叉研究主题。

虽然图情领域学科交叉研究较为成熟，但是大多数研究仅探究图情学科本身包含的交叉主题以及图情学科与2个或3个学科间的交叉主题，缺少对多学科领域交叉主题识别的研究，并且对特定学科领域内多学科交叉演化的规律以及促进学科交叉融合发展的因素的探索尚不充分。

1.2 学科交叉研究方法

从现有研究来看，学科交叉研究方法主要包含两大类：文献计量方法和文本挖掘方法[11-13]。文献计量方法中的经典方法是共词分析法，该方法基于学科间的交叉关键词来识别学科间的交叉主题，通过分层聚类、频次分析或突发词检测等研究关键词所代表的研究领域或学科交叉主题的演化过程[1,5]。共词分析法还包括社会网络分析方法和复杂网络方法等[11]。引文网络分析是另一种重要的文献计量方法[1,11-13]：基于学科的引文网络并结合测度指标，可识别学科交叉主题，通过结合时间序列分析等方法，即可对学科交叉主题进行演化分析[11]。文本挖掘方法则借助各种文本挖掘算法，尝试从研究文献的文本内容中提取学科交叉主题，常用方法包括主题模型方法[14-16]、文本聚类方法[1,6,17]、非相关知识发现方法[1]、概念格方法[18]和知识图谱方法[19]等。作为一种文档主题生成模型，LDA主题模型具有清晰的层次结构且无须标注训练集，被广泛应用于主题发现研究[14-15]。

文献计量方法没有直接利用文献内容信息，而仅从文献题录信息中提取研究主题并进行演化分析，存在较多人工干预过程，且语义误差较大，因此适用于特定专业领域的主题演化研究。文本挖掘方法可更全面准确地发现文献的研究主题，且对领域专业度要求相对较低。因此，本文结合文本挖掘方法中的LDA主题模型与WLC决策规则，对图情领域学科交叉主题进行识别和演化分析。

2 研究方法与数据获取

研究框架（见图1）分为3个部分：图情领域热点交叉学科遴选、基于LDA与WLC的热点交叉学科主题识别和基于关联规则的主题演化分析。

图1 研究框架

2.1 研究方法

2.1.1 图情领域热点交叉学科遴选

从引文分析角度出发，通过以下过程获取图情领域的热点交叉学科：首先，统计该领域被引期刊所属学科类别，分类汇总并降序排列；其次，设定被引频次阈值，筛选满足条件的期刊；最后，定义热点交叉学科为上述期刊所属学科中被引频次排名前10的学科。

2.1.2 基于LDA与WLC的热点交叉学科主题识别

以热点交叉学科对应参考文献原文的题名、摘要、关键词等为数据源，利用LDA主题模型识别各个时间窗口文档集的潜在主题，利用WLC决策规则过滤垃圾主题，得到主题识别的最终结果。

（1）基于LDA的主题发现。在利用LDA发现主题时，准确确定主题的数量与质量非常关键[20-21]。LDA一般使用主题一致性（Topic Consistency，TC）和主题多样性（Topic Diversity，TD）作为衡量主题质量的绩效指标[8-9]。主题数量是TC和TD的综合表征，对主题发现的最终结果有直接影响。确定LDA模型主题数量的方法一般有2种[15-16]：①计算LDA模型在文档集上的困惑度；②引入狄利克雷过程以确定主题数量。考虑到算法的可行性及效率，选用困惑度来确定各时间窗口的最佳主题数量，困惑度定义如公式（1）所示[23]。

$$P(D)=\exp\left[-\frac{\sum_{d=1}^{D}\log p(w_d)}{\sum_{d=1}^{D}\log N_d}\right] \quad (1)$$

式中：D为文档集中文档的总数量；w为主题词；d为文档；$p(w_d)$为文档集中每个主题词出现的概率；N_d为文档d不排重的词项数。当$P(D)$的值最小时，主题具有较好的语义表达效果，取此时的主题数量作为文档集建模主题数量。

（2）基于WLC的冗余主题过滤。根据困惑度指标可获取最佳主题数量，但是由于LDA是典型的无监督的贝叶斯模型，可能有一部分得到的主题内容难以解读，因此需要进一步过滤。WLC决策规则主要进行多目标决策分析，可应用于冗余主题过滤，其基本原理为：主题间的相似度是多种相似计算方法共同作用的结果，而对每种方法都可赋予权重，通过加权线性组合的方式即可过滤不满足条件的冗余主题[24]。参照以往研究[24-25]，采用余弦相似度和Pearson相关系数2种相似度计算方法来衡量主题相似度，并利用相似度阈值进行过滤，再通过WLC决策规则进行多目标分析，根据相似度阈值确定各时间窗口的种子主题。

2.1.3 基于关联规则的主题演化分析

（1）主题关联建立。由LDA得到的各时间窗口的主题是彼此没有关联的，在进行主题演化分析之前需要建立相邻时间窗口之间主题的关联。主题演化在内容上存在继承关系，相邻时间窗口之间具有关联的主题在内容上有一定相似性[25]。将夹角余弦值作为主题关联的衡量指标，考虑到2个相邻时间窗口包含主题存在不一致的情况，定义在时间窗口t中未出现的主题词w对应的概率值如公式（2）所示[15]。

$$p(w|T_i^t)=\frac{\beta}{n_{T^t}^{(\cdot)}+V\times\beta}\quad(w\in V^{t+1},\ w\notin V^t) \quad (2)$$

式中：V^t与V^{t+1}分别为t和$t+1$时间窗口内的词分布，V为V^t与V^{t+1}中的词项数总和；β为主题的模型参数；T_i^t为在t时间窗口中的第i个主题；$n_{T^t}^{(\cdot)}$为T_i^t所含词项数总和。

相邻时间窗口的主题关联规则如下：对时间窗口t中的某一主题T_i^t，将其与时间窗口$t+1$内各主题的相似度按降序排列，相似度最大的即为T_i^t的后向主题，定义如公式（3）所示。

$$T_j^{t+1}=f_{\text{post}}(T_i^t) \quad (3)$$

对时间窗口$t+1$中的某一主题T_j^{t+1}，将其与时间窗口t内各主题的相似度降序排列，相似度最大的即为T_j^{t+1}前向主题，定义如公式（4）所示。

$$T_i^t=f_{\text{prior}}(T_j^{t+1}) \quad (4)$$

（2）主题关联过滤。为提升主题演化分析的准确性，需要对主题关联进行过滤。对已有研究[25]定义的主题关联过滤规则进行完善，并利用该规则对主题关联进行过滤。

①设立主题相似度阈值$\varepsilon\in(0,1)$，若相邻时间窗口中的最大主题相似度S_{MAX,T_i^t}小于ε，则认定该关联失效。

②出现以下2种情况，则认为T_j^{t+1}与T_i^t的关联失效：

若T_j^{t+1}为主题T_i^t的后向主题，设时间窗口t内主题与T_j^{t+1}计算的相似度排名中，T_i^t的排序位置为第s（s>1），此时存在主题T_k^t所处排序位置为$\rho \in [1, s)$，且$f_{post}(T_k^t) \neq T_j^{t+1}$；若$T_i^t$是主题$T_j^{t+1}$的前向主题，设时间窗口$t$+1内主题与$T_i^t$计算的相似度排名中，$T_j^{t+1}$的排序位置为第$s$（$s$>1），此时存在主题$T_k^{t+1}$所处排序位置为$\rho \in [1, s)$，且$f_{prior}(T_j^{t+1}) \neq T_i^t$。

③设时间窗口t内的主题与T_j^{t+1}的最大主题相似度为$S_{MAX,T_j^{t+1}}$，设定阈值$\mu \in (0, 1)$，若$S(T_i^t, T_j^{t+1}) < \mu \times \max\{S_{MAX,T_j^{t+1}}, S_{MAX,T_i^t}\}$，则$T_i^t$与$T_j^{t+1}$的关联失效。

对已有关联关系的主题对应用以上关联过滤规则，若过滤后主题对仍存在关联，则可认定该主题对之间具有演化关系。

（3）主题演化分析。主要从主题强度和主题内容两方面进行主题演化分析[15-16,25]。

①主题强度演化。主题强度是主题自身的属性之一，其有助于用户对主题的趋势进行直观评价和对比分析，计算方法如公式（5）[16]所示。

$$I(T_i^t) = \frac{\sum_{j=1}^{D_t} p(T_i^t \mid d_j)}{D_t} \quad (5)$$

式中：D_t为t时间窗口内的文档总数。主题强度的变化能够直接衡量主题的稳定程度，若主题强度随时间发展幅度变化较小则说明主题很稳定，反之则不稳定。

②主题内容演化。主题的内容演化分析主要是指判断相邻时间窗口对应主题的演化关系，将过滤后具有关联的主题分别进行前向和后向推理，将其关系归纳为新生、消亡、继承、合并与分裂5类[25]，如图2所示，其中虚线表示主题间演化关系不存在。

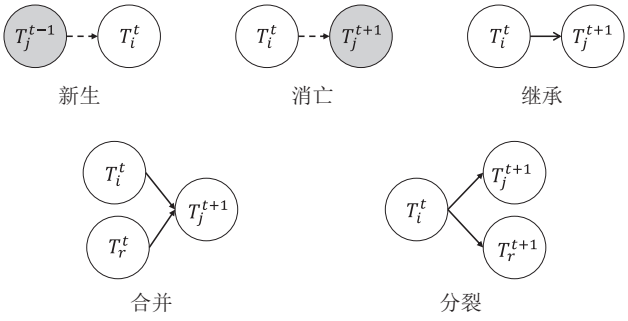

图2　主题内容演化的5种类型

2.2　数据获取

以中国知网（以下简称“知网”）为来源数据库。首先，根据期刊2021年的影响因子确定图情领域排名前10的中文期刊。其次，检索期刊2002—2021年发表的所有论文，统计每篇论文的各参考文献期刊所属学科类别，从中遴选图情领域的热点交叉学科。考虑到论文的题名、摘要和关键词能充分表达论文主题，按年份获取参考文献属于热点交叉学科的论文对应的题名、摘要、关键词等题录数据用于主题发现。

2.2.1　获取图情领域的核心期刊论文

根据期刊2021年的影响因子确定图情领域排名前10的中文核心期刊：《中国图书馆学报》《情报学报》《图书情报知识》《图书与情报》《大学图书馆学报》《图书馆工作与研究》《情报资料工作》《图书情报工作》《情报理论与实践》《现代情报》。

获取上述10种期刊2002—2021年发表的所有论文，检索日期为2022年4月9日。经过数据去重、去空、去除奇异字符等预处理后，得到51 615条记录，利用自编爬虫程序按年份获取各论文的参考文献，经过数据清洗后得到228 904篇参考文献的信息。

2.2.2　图情领域的交叉学科识别

统计所有论文的参考文献期刊所属学科类别，并分类汇总、降序排列，剔除图情领域期刊，筛选剩余期刊中被引频次大于50次的期刊，并根据知网的学科分类对其所属领域进行划分，再筛选出被引频次排名前10的交叉领域（见表1）。

表1　被引频次排名前10的交叉领域

交叉领域	被引频次/次
信息科技	18 376
经济与管理科学	12 148
社会科学II	7 709
基础科学	5 023
社会科学I	2 878
哲学与人文科学	1 145
工程科技II	775
医药卫生科技	528
农业科技	67
工程科技I	55

2.2.3 获取主题建模数据

根据被引频次排名前10的交叉领域研究现状和现今图情领域与各技术类学科的高度交叉融合现状，主要研究图情领域与信息科技和基础科学中的技术领域的交叉情况。首先，统计信息科技和基础科学相关技术领域的期刊，共得到815种与技术相关的信息科技中、英文期刊，以及1 240种基础科学中、英文期刊。其次，按年份筛选涉及这些期刊的参考文献，得到共26 679条记录。最后，分别按年份检索各参考文献对应的论文，下载其标题、摘要和关键词信息作为LDA建模的文本语料库。

3 结果分析

3.1 热点交叉学科主题识别与分析

3.1.1 LDA模型参数设置

利用困惑度指标来确定2002—2021年（20个时间窗口）的最佳主题数量。分别设置各个时间窗口文档集的主题数量K=[1，50]，步进值为1，并设置LDA的超参数$\alpha=\frac{50}{K}+1$、β=0.1。设置LDA模型抽取各主题下出现概率最大的3 500个词，将迭代频次设置为1 000次，选用Gibbs抽样方法进行主题建模，分别计算各主题数量的困惑度，选择困惑度最低的主题数量作为建模主题数量。图3显示了2014年数据集中困惑度与主题数量之间的关系。由图3可知，在主题数量为17个时困惑度最低，故2014年的最佳主题数量为17个。

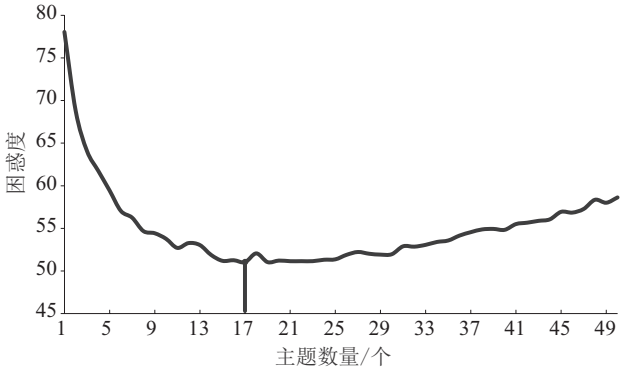

图3 不同主题数量对应的困惑度

3.1.2 WLC冗余主题过滤

进一步利用WLC进行冗余主题过滤，得到各个年份最终的主题数量，如表2所示。从表2可以看出，WLC过滤了一部分冗余主题，尤其针对2002—2006年的主题建模结果起到了较好的过滤效果。虽然过滤前后2007—2021年主题数量无变化，但是通过对2007—2021年各个年份的主题进行内容分析，发现所有主题均非冗余主题。

表2 各时间窗口的最佳主题数量

年 份	过滤前主题数量/个	过滤后主题数量/个	年 份	过滤前主题数量/个	过滤后主题数量/个
2002	7	3	2012	18	18
2003	6	3	2013	23	23
2004	11	9	2014	17	17
2005	17	14	2015	21	21
2006	15	14	2016	21	21
2007	15	15	2017	20	20
2008	13	13	2018	21	21
2009	16	16	2019	21	21
2010	13	13	2020	25	25
2011	14	14	2021	22	22

3.1.3 主题识别结果分析

得到主题建模结果后，参考已有研究[26]，并邀请专家进一步对主题识别结果进行评估，得到最终的主题识别结果，表3展示了2014年部分主题中出现概率排名前10的主题词情况。

以T4（数据挖掘）为例，数据挖掘为数据科学与图情学科的交叉主题，其借助数据分析工具发现数据与模型之间的关系，能挖掘数据隐含的特征。将数据挖掘应用到图情领域的典型案例是知识发现，可应用数据挖掘技术进行关键词抽取、引文分析、文本聚类等，获取新知识。另外随着时间积累，图书馆电子资源呈指数级增长，通过利用数据挖掘相关技术进行数据分析，可提高图书馆和读者效率。

根据主题识别结果可发现图情领域普遍与其他技术学科交叉，并且研究对象已经深入高校、科研院所及政府等。因此，进一步对图情领域的交叉主题进行主题强度和主题内容上的演化分析，以期厘清各交叉主题的演化脉络，分析图情领域与交叉领域的交叉主题发展进程，助力高校、科研院所及政府等所需技术发展。

表3　交叉领域主题中出现概率排名前10的主题词情况

T4（数据挖掘）		T7（云计算）		T11（电子政务）	
主题词	概　率	主题词	概　率	主题词	概　率
数据挖掘	0.012 09	云计算	0.036 82	政府	0.017 82
热点	0.007 09	大数据	0.035 59	电子政务	0.012 42
可视化	0.006 00	物联网	0.011 50	技术预见	0.007 97
科技人才	0.002 51	mapreduce	0.007 49	绩效评估	0.003 65
引文网络	0.002 23	hadoop	0.006 10	教育	0.003 25
多媒体	0.002 10	服务模式	0.003 89	改革	0.002 58
项目管理	0.001 96	信息系统	0.002 92	信息化	0.002 04
意见挖掘	0.001 82	数字图书馆	0.002 92	科学共同体	0.002 04
电子资源	0.001 54	用户体验	0.002 92	信息服务	0.001 50
竞争情报	0.001 40	信息安全	0.002 50	政府信息公开	0.001 50

3.2　主题演化分析

3.2.1　主题强度演化

以2004年的T3（图像领域信息检索）、T4（个性化用户信息服务）、T5（数据库信息安全）、T6（电子商务）为例，依据主题强度演化分析方法得到上述4个主题在2012—2021年10年间的主题强度变化，如图4所示。

图4　2012—2021年4个主题对应的主题强度演化

图4中4个主题的强度均有明显的起伏，以T6（电子商务）为例：其在2012—2013年主要包括复杂网络结构研究，主题强度有所下降，说明该主题受关注度呈下降趋势；2014年发展为社会网络聚类算法的相关研究，主题强度大幅上升；2015年该主题进一步发展为文本聚类相关算法的研究，主题强度下降；2017年演化为中文语义抽取相关算法的研究，相关研究一直持续到2020年，在此期间，该主题的主题强度整体呈下降趋势；2021年该主题进一步演化为中文文本情感分类的相关研究，且其主题强度陡然上升，说明在未来中文文本情感分类可能成为受关注的研究方向。对其余主题的强度演化情况进行总结，发现不同技术领域的交叉主题的主题强度存在差别，可见不同技术领域的交叉研究主题的研究热度与学科交叉程度可能存在正相关关系，并且若某一交叉主题长时间不改变，其主题强度会逐渐下降，而产生新技术或新应用时，主题强度会在短时间内陡然上升。这与现实情况相符合：某一主题若想一直保持热度就要不断进行创新，若长时间没有新发展则其关注度会下降。

3.2.2　主题内容演化

以2002年的T1（基于agent的信息化建设）为例，进行主题内容演化分析，该主题中出现概率排名前20的主题词及其概率分布如表4所示。根据2.1.3节中的主题内容演化分析方法得到T1（基于agent的信息化建设）的内容演化情况，如图5所示。

从图5可以发现，在2014年前，主题演化类型以分裂为主。例如，2002年的T1（基于agent的信息化建设）在2007年由T1（智能信息采集）分裂为2008年的T10（管理系统设计）和T11（文本分类算法）2个主题，其中：T11（文本分类算法）继承发展至2011年消亡；2009年，2008年的T10（管理系统设计）由T5（功能系统网络设计）分裂为2010年的T5（复杂网络结构研究）和T13（信息系统发现）2个主题，T13（信息系统发现）

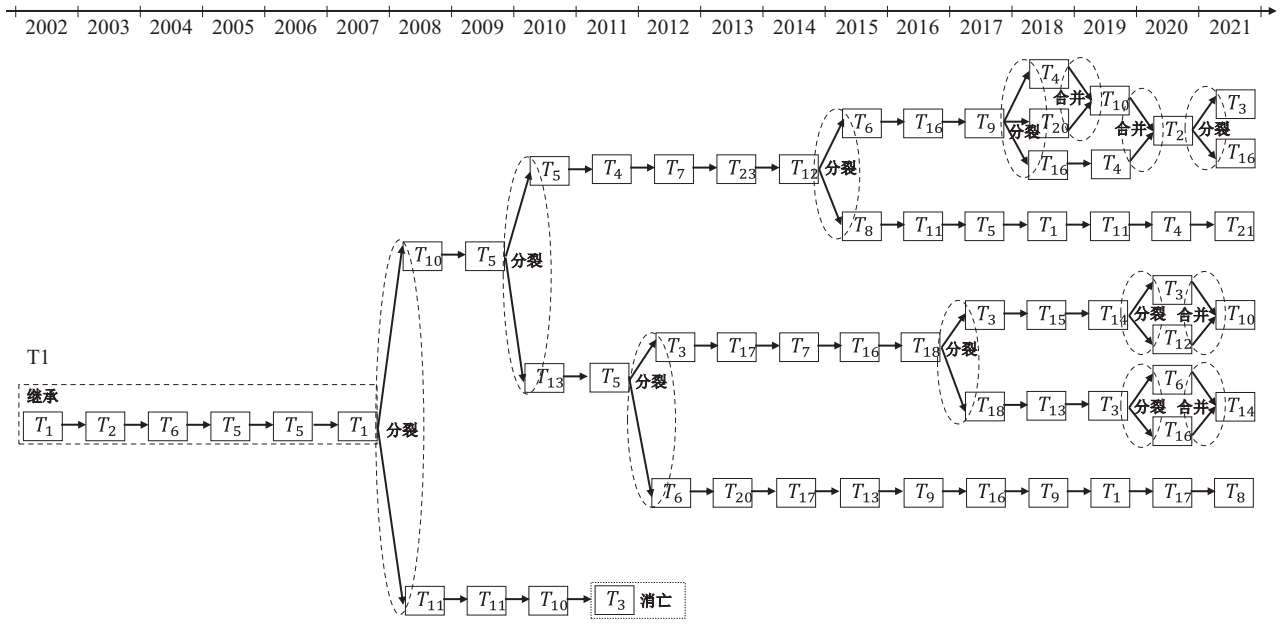

图5 T1主题内容演化图

表4 T1的主题词概率分布

主题词	概 率	主题词	概 率
系统	0.045 90	分布式	0.005 03
计算机	0.015 00	通信	0.004 54
agent	0.014 50	交互	0.004 54
模型	0.012 51	检测	0.004 54
图像	0.009 52	犯罪	0.004 54
环境	0.009 02	开放	0.004 04
入侵检测	0.007 53	超媒体	0.003 54
个性化	0.005 53	信息化建设	0.003 54
智能	0.005 53	三维	0.003 54
协作	0.005 53	数据仓库	0.003 04

在2011年由T5（web信息平台设计）分裂为2012年的T3（面向产品的系统设计）和T6（电子政务技术）。

在2014年后，主题之间的演化形式趋于复杂化、多样化。例如，2002年的T1（基于agent的信息化建设）演变为2014年的T12（社会网络聚类），并分裂为2015年的T6（文本聚类）和T8（社会网络发现），在2017年又由T9（中文语义抽取）分裂为2018年的T4（文本分类）、T16（图像分类）和T20（目标评价方法），其中T4（文本分类）和T20（图像分类）在2019年又合并为T10（文本自动抽取），T10（文本自动抽取）和T4（深度学习在中文分类中的应用）在2020年合并为T2（中文事件抽取），继而分裂为2021年的T3（文本信息自动挖掘）和T16（中文文本情感分类）。

通过对交叉主题的主题内容进行演化分析，发现随着科学技术的发展，学科交叉越来越复杂化、多样化。在对T1（基于agent的信息化建设）的内容演化情况分析中，发现该主题的演化以2014年为分界点，2014年后，主题之间的分裂和合并的频率明显增加。通过对其余主题进行类似的演化分析，因篇幅所限仅展示2002年以来T2（搜索引擎信息数据分析）和T3（网络信息开发）的主题内容演化结果（见图6），也可以得到在特定年份后，主题分裂和合并的频率明显增加，主题内容演化复杂化、多样化。这也说明，图情相关交叉技术学科的发展促进了科学领域整体的发展，进一步推动了学科交叉创新点的提出和发展。

4 结论与建议

本文融合LDA主题模型、WLC决策规则和关联规则对图情领域的多学科交叉主题进行分析，开展了对该领域内多学科综合演化发展规律的探索。不同于以往研究分析2个或3个学科交叉情况，本文全面分析了图情领域与其他热点技术领域的交叉现状，揭示了各交叉主题演化发展历程，得出如下结论。

其一，从图情交叉领域分布来看，图情交叉领域主

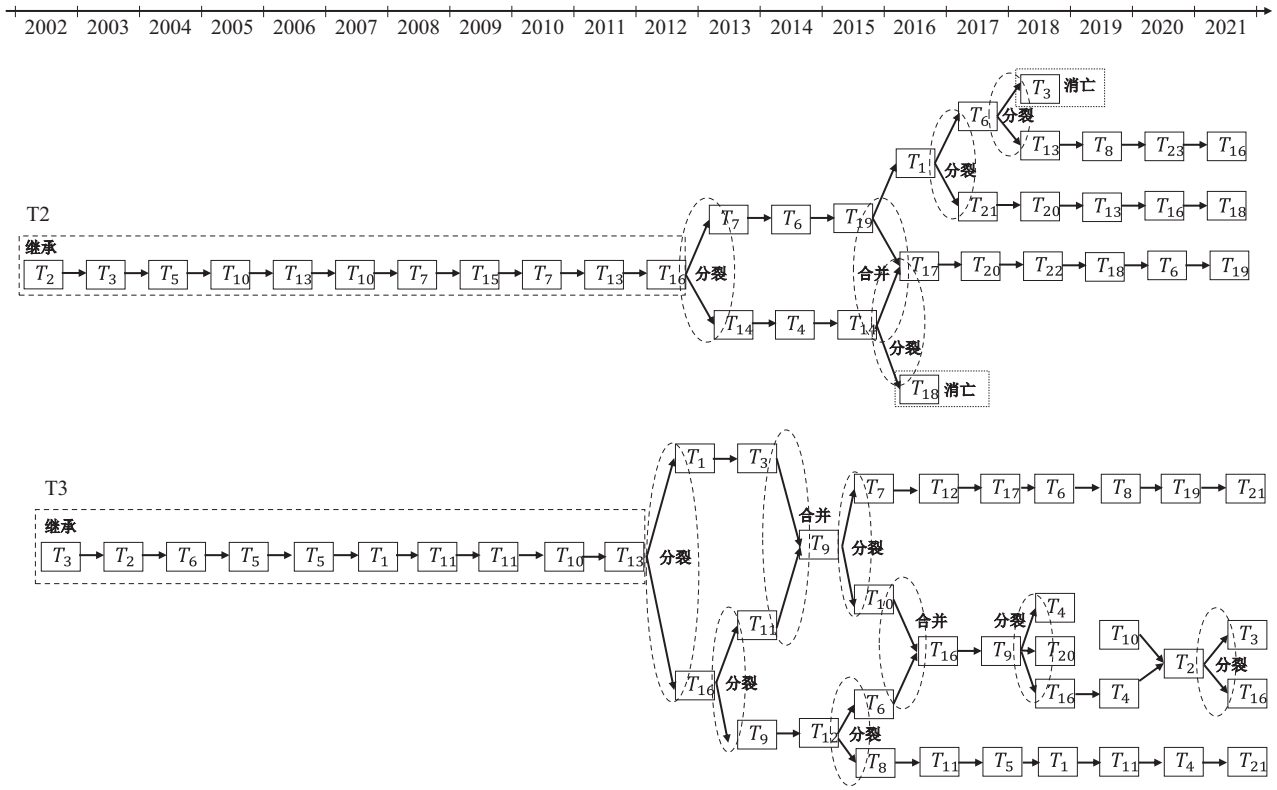

图6　T2和T3主题内容演化图

要集中在信息科技、经济与管理科学和社会科学，具体的学科领域包括数据科学、计算机软件及计算机应用（含人工智能）、科学研究管理等，涵盖方法、技术和应用领域，它们在图情领域的学科交叉发展中起到了关键作用。

其二，从图情领域和相关热点技术学科的交叉主题演化来看，首先通过主题强度演化分析，发现交叉主题的主题强度与技术更新的频率呈现正相关关系；其次通过对主题内容演化进行分析，发现图情领域的交叉主题演化均会以特定年份为分界点，呈现出复杂化、多样化的特点。

基于此，本文提出以下两点促进图情学科交叉融合发展的建议。

其一，从图情交叉领域分布来看，应促进图情与数据科学、信息科学，特别是人工智能技术、管理科学等交叉融合发展。科技文献蕴含海量专业、可信、规范的科技知识与数据，图情学科应加强与上述学科的交叉融合，加强科技文献大数据深层次开发利用，为人工智能发展提供高质量数据支撑与知识发现服务[27-28]。

其二，从图情领域的学科交叉主题演化分析结果来看，一方面，在图情学科与其他学科交叉融合发展过程中，学科交叉主题逐渐复杂化、多样化，从而使本学科面临新的挑战，因此在与其他学科的理论知识、技术方法交叉融合过程中，要牢记本学科的特质，坚定学科内核，提升学科影响力和地位[29-30]；另一方面，信息科技、计算机等学科知识已经被广泛应用于图情领域，促进了本领域往大数据、人工智能等领域纵深发展，同时促进了本领域的创新发展，在此背景下图情领域的人才培养要注重数学和计算机等相关技术学科知识，培养满足时代需求的高质量人才，促进学科发展和社会进步[3,30-31]。

需要指出的是，本文尚存在一定的局限。①数据源不够丰富。本文未考虑文献涉及的一些外部特征信息，例如期刊排名等，在后续工作中可以考虑加入外部特征信息以完善主题演化分析；另外可以进一步考虑引入专著、专利等信息，使得数据源更多样化。②主题词不够规范化。本文仅将已有文献的关键词导入用户词表，但未考虑不同文献中同一主题关键词表达的多样性，未来可以考虑引入规范化词表，提升学科主题下对应主题词表达的准确性和规范度。③学科主题内容定义不够准确。由于缺乏学科领域知识，对于学科主题的定义仅来源于主题中已有的主题词，导致主题定义可能

存在偏差，在未来的研究中，应该进一步思考如何更准确地表达主题的含义。学科交叉尚有诸多未解之谜，未来研究宜针对上述不足展开，提高学科交叉主题演化分析的准确度，促进不同学科之间的合作和交叉融合发展。

参考文献

[1] 李佳蕾，安培浚，肖仙桃. 学科交叉主题识别方法研究综述 [J/OL]. 数据分析与知识发现：1-19 [2022-10-26]. http://kns.cnki.net/kcms/detail/10.1478.G2.20220909.0823.002.html.

[2] 温芳芳，杨倩倩，李翔宇. 我国人文社会科学学科交叉性的测度及其演化规律研究：基于国家社科基金关键词耦合分析 [J]. 现代情报，2022，42（3）：157-167.

[3] 魏来. 师范类院校图情档学科建设与发展思考 [J]. 图书情报工作，2022，66（1）：75-82.

[4] 王思茗. 图书馆学情报学领域学科交叉及其地区差异的演化分析 [D]. 长春：东北师范大学，2021.

[5] 崔斌. 基于共现网络的我国图书情报学跨学科合作研究 [D]. 淄博：山东理工大学，2017.

[6] 王连喜，曹树金. 学科交叉视角下的网络舆情研究主题比较分析：以国内图书情报学和新闻传播学为例 [J]. 情报学报，2017，36（2）：159-169.

[7] 隗玲，许海云，胡正银，等. 学科主题演化路径的多模式识别与预测：一个情报学学科主题演化案例 [J]. 图书情报工作，2016，60（13）：71-81.

[8] CHANG Y W，HUANG M H. A study of the evolution of interdisciplinarity in library and information science：using three bibliometric methods [J]. Journal of the American Society for Information Science and Technology，2012，63（1）：22-33.

[9] FIGUEROLA C G，MARCO F J G，PINTO M. Mapping the evolution of library and information science（1978—2014）using topic modeling on LISA [J]. Scientometrics，2017，112（3）：1507-1535.

[10] 李长玲，郭凤娇，魏绪秋. 基于时序关键词的学科交叉研究主题分析：以情报学与计算机科学为例 [J]. 情报资料工作，2014（6）：44-48.

[11] 商宪丽. 基于潜在主题的交叉学科知识组合与知识传播研究 [D]. 武汉：华中师范大学，2017.

[12] 许海云，董坤，隗玲. 学科交叉主题识别与预测方法研究 [M]. 北京：科学技术文献出版社，2019.

[13] 张琳，黄颖. 交叉科学：测度、评价与应用 [M]. 北京：科学出版社，2019.

[14] BLEI D M，NG A Y，JORDAN M I. Latent dirichlet allocation [J]. The Journal of Machine Learning Research，2003，3（4/5）：993-1022.

[15] 徐传舰. 基于LDA的国内图书情报学学科交叉及演化研究 [D]. 曲阜：曲阜师范大学，2020.

[16] 茅利锋. 基于主题模型的主题演化分析及预测 [D]. 南京：南京邮电大学，2016.

[17] 吴蕾，孙巍. 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 [J]. 数字图书馆论坛，2015（12）：15-20.

[18] 邵作运，李秀霞. 基于引文耦合和概念格的学科交叉知识结构探测 [J]. 图书情报工作，2015，59（8）：78-86.

[19] 王思茗，魏玉梅，滕广青，等. 图书情报学领域中的学科交叉现象及其地区差异 [J]. 情报理论与实践，2019，42（12）：8-15.

[20] HU Z Y，FANG S，LIANG T. Empirical study of constructing a knowledge organization system of patent documents using topic modeling [J]. Scientometrics，2014，100（3）：787-799.

[21] 范云满，马建霞. 基于LDA与新兴主题特征分析的新兴主题探测研究 [J]. 情报学报，2014，33（7）：698-711.

[22] GURCAN F，CAGILTAY N E. Exploratory analysis of topic interests and their evolution in bioinformatics research using semantic text mining and probabilistic topic modeling [J]. IEEE Access，2022，10：31480-31493.

[23] MAO J，LIANG Z T，CAO Y J，et al. Quantifying cross-disciplinary knowledge flow from the perspective of content：introducing an approach based on knowledge memes [J]. Journal of Informetrics，2020，14（4）：101092.

[24] 李保利，杨星. 基于LDA模型和话题过滤的研究主题演化分析 [J]. 小型微型计算机系统，2012，33（12）：2738-2743.

[25] 秦晓慧，乐小虬. 基于LDA主题关联过滤的领域主题演化研究 [J]. 现代图书情报技术，2015（3）：18-25.

[26] ZHANG Y，PORTER A L，HU Z Y，et al. “Term clumping”for technical intelligence：a case study on dye-sensitized solar cells [J]. Technological Forecasting and Social Change，2014，85：26-39.

[27] CHO S M，PARK C U，SONG M. The evolution of social health research topics：a data-driven analysis [J]. Social Science & Medicine，2020，265：113299.

[28] 胡正银. 建设学科领域知识发现系统，助推知识服务智能化升级转型 [C] //2022年中国图书馆学会专业图书馆分会学术年会. 中国图书馆学会专业图书馆分会，2022.

[29] 陆瑶，卢超，董克，等. 从幕后到台前：数据要素化带来图情学科发展机遇与挑战 [J]. 图书情报知识，2021，38（6）：123-133.
[30] 吴蝶，曹如中，熊鸿军，等. 新文科建设背景下图书情报档案学科数字化转型发展研究 [J]. 图书馆理论与实践，2023（2）：12-16，26.
[31] 孙曙光，张帆，郝爽. 学科评估视域下图书情报与档案管理学科建设的问题与策略 [J]. 情报科学，2021，39（5）：163-168.

作者简介

石栖，女，硕士研究生，研究方向：大数据情报分析方法与技术。
胡正银，男，博士，研究员，通信作者，研究方向：科技大数据分析方法与技术、科技情报知识挖掘与知识发现，E-mail：huzy@clas.ac.cn。
王莉晓，女，硕士研究生，研究方向：颠覆性技术识别。

Analysis of Interdisciplinary Topic Evolution in the Field of Library and Information Science Based on LDA-WLC

SHI Xi[1,2]　HU ZhengYin[1,2,3]　WANG LiXiao[1,4]
(1. School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, P. R. China;
2. Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu 610041, P. R. China; 3. Sichuan Science and Technology Information Intelligent Mining and Application Engineering Research Center, Chengdu 610041, P. R. China;
4. Wuhan Library, Chinese Academy of Sciences, Wuhan 430071, P. R. China)

Abstract: The analysis of the evolution of interdisciplinary topics can reveal the process, power, and future development trend of the generation and development of interdisciplinary topics, and promote the construction and development of disciplines. Firstly, we take the papers of Chinese core journals in the field of library and information science as the data source and select all the hot interdisciplinary subjects in this field through citation analysis. Secondly, we use LDA topic model and WLC decision rules to mine and filter topics. The evolution of interdisciplinary topics is analyzed from the aspects of subject intensity and subject content. The results show that there is a positive correlation between the intensity of the topic and the frequency of technology updating. The evolution of theme content takes a particular year as the demarcation point, showing the characteristics of complexity and diversification. Finally, we put forward some suggestions on the development of interdisciplinary integration of library and information science.

Keywords: Interdisciplinarity; Topic Mining; Topic Filtering; Topic Evolution; LDA

（责任编辑：王玮）