基金项目对科学研究的关联影响分析*

吕晶[1] 郭思月[1] 滕广青[1] 马卓[2]
(1. 东北师范大学信息科学与技术学院，长春 130117；2. 吉林省科学技术信息研究所，长春 130033)

摘要：基金项目与科学研究的多维度关联影响分析能够提供细粒度洞察，有助于国家科学战略的科学规划及科技策略制定。本文结合国家自然科学基金数据以及相关领域的论文数据，采用LDA进行主题建模，基于主题相似度识别基金项目与论文之间的关联影响。研究结果表明，较小领域的基金项目与所隶属的更大领域科学研究之间存在关联影响关系，基金项目对科学论文的引导与促进作用更大，且影响作用的持续时间更长。

关键词：基金项目；科学论文；主题建模；主题关联；影响周期
中图分类号：G250.2
DOI：10.3772/j.issn.1673-2286.2019.12.003

随着人类社会科技发展水平的不断提高，各国政府越来越意识到科学技术的发展对社会、经济、军事、教育等领域的重要作用，并通过设立科学基金项目资助那些重点需求、亟待创新的科学问题研究。科学界的基金项目对于相关领域科学研究的发展具有积极的支持与促进作用，甚至能够引导科学创新的发展方向。由于科学论文是科学研究进展的最直接体现，图书情报学界在以往的研究中，大多借助论文中标注的基金信息对其中的关联影响进行判识。然而，论文中的基金标注信息与论文中的关键词一样只是外在的形式特征，单纯依靠外在形式特征的分析仅能够获得有限的低维度信息，无法对其进行全景式及细粒度的洞察。近年来，随着数据科学思维引入图书情报学界，科技信息的多维复合分析逐渐引起学术界关注。科技信息多维复合分析，能够通过跨维度的交叉关联，挖掘与发现不同维度间隐含的模式信息。

鉴于此，本研究通过研究领域、数据对象、研究方法等多维度结合的综合分析，探索基金项目与科学研究之间的关联影响，尝试揭示其中潜在的模式与规律，以期为国家规划科技发展战略、制定科研资助策略等提供有益的支持。

1 相关研究工作

随着科学技术在社会发展中的作用愈发凸显，人类社会对于科学研究的重视已经提到前所未有的高度。各国政府通过科学基金项目重点扶持与资助那些对国家发展和社会进步产生重要支撑与推动作用的研究领域。图书情报学界也对基金项目对科学研究产生的影响作用展开研究。Butler[1]、Wang等[2]通过期刊论文中的基金标注信息筛选出获基金资助的论文，基于基金资助数据与科研总产出数据，探究资金资助对科研产出的影响。Boyack等[3]通过基金资助的论文，分析政府资助对科研出版物数量和被引频次的影响。国内学者陈秋怡等[4]基于Web of Science核心集引文索引数据库中发表论文最多的6个国家的科研基金资助与论文产出的整体分析，探测科研基金资助投入与高水平国际论文产出之间的关系。许鑫等[5]以自然科学领域的代表学科作为微观层面的研究对象，通过SCI论文中的基金标识区分基金论文和非基金论文，从引用和使用两个角度分析科学基金资助对论文的即时影响力与内容影响力的影响。上述研究都是通过论文中的基金标注信息识别基金项目与论文之间的关联关系，这种基于外在形式特征的筛选方法对于衡量基金项目的直接产出是有效的，但在探查基金项目对更大领域科学研究的影

*本研究得到国家社会科学基金项目“基于复合数据的科技信息跨维度挖掘与推荐研究”（编号：19BTQ063）资助。

响方面则显得力不从心。

随着数据科学范式的兴起，大数据思维已经被学术界普遍接受。现有的研究表明，利用多维度数据结合的研究方法可以识别出基于单一维度数据不能识别的研究前沿，研究人员在科技情报分析中逐渐有意识地将论文数据、专利数据等不同来源的数据进行整合[6-7]，即使基于单一数据源也在分析工作中包容了文献、作者、关键词、机构、时间等更多的数据维度[8]。相对于以往基于单一维度数据的研究而言，多维数据的整合为科技情报分析工作提供了更好的说服力[9]。然而，科技文献中明确标识的外在形式特征尚不足以支持更细粒度的分析工作。自然语言处理（Natural Language Processing，NLP）技术与方法逐渐被应用到多维度数据的科技信息分析中，通过对科学文献文本的语义分析获取其中蕴含的更细粒度的语义关联。此类研究工作包括基于多文献数据集的主题挖掘[10]，期刊论文、学位论文、专利文献间的主题差异[11]，科学规划与基金项目的主题识别[12]，基于基金项目与论文的前沿探测[13]，论文与专利之间的主题关联演化[14]，以及基金项目到论文的知识扩散效应[15]等诸多领域的研究。

综上所述，基于多维度数据的科技信息分析已经成为学术界的共识，而且自然语言处理技术的成熟为更细粒度的跨维度分析提供了方法支持。本研究在现有研究成果的基础上，采用基金项目与论文、下位学科与上位学科、关键词与文本主题等多维度数据结合的综合分析方法，对基金项目与科学论文之间的关联影响进行识别与分析，探测和揭示其中潜在模式与规律。

2 相关理论基础

2.1 文本主题模型

本研究除采用传统科学计量中的关键词分析方法外，还将对基金项目与科学论文进行主题建模，通过主题相似性识别基金项目与科学论文之间关联影响。研究工作采用目前成熟且流行的隐狄利克雷分布（Latent Dirichlet Allocation，LDA）主题模型。LDA主题模型是由Blei等[16]提出的用于识别大规模文档中潜在主题信息的三层贝叶斯概率模型，包括单词层、主题层和文档层。通过对目标文本集进行建模分析，可以通过中间层主题得到文档中出现的词的概率，其公式如下。

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{(n=1)}^{N} p(Z_n \mid \theta)\, p(w_n \mid Z_n, \beta) \quad (1)$$

公式（1）中，α、β为超参数，θ为“文本-主题”概率分布，z为词的主题分布，Z表示主题，N为词的数量，w表示词，w_n是序列中单词的第n个单词，Z_n是文档中第n个单词的主题。研究工作进一步引入吉布斯采样（Gibbs Sampling）对主题模型求解[17]。吉布斯采样假设文本中出现的词汇连成一串且不重复，在LDA迭代过程中，Gibbs为这个串中的每一个词分配一个主题，然后不断地更新其状态直到收敛到一个较为稳定的数据集，从而计算出LDA的概率分布的近似值，是目前概率分布计算中采用较多且准确程度较高的方法[18]。在使用LDA模型进行核心主题识别过程中，一个重要的问题是最佳主题数目的确定。主题数目会影响到主题模型的效果，以往采用困惑度[16]确定的主题数，往往结果冗长[19-20]且主题过于分散。现有的研究表明，主题一致性是衡量主题质量最有效的方法[21]，且一致性得分与人类判断的主题连贯性非常相似[22-23]。因此，本研究采用一致性分数对主题数目进行判定，以此提高主题词聚类的效果。其具体公式如下[24]。

$$C(t; V^t) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(V_m^t, V_l^t) + 1}{D(V_l^{(t)})} \quad (2)$$

公式（2）中$V^t = (V_1^t, \cdots, V_m^t)$代表列表中$M$个词最可能属于主题$t$，$M$表示主题中概率值最大的主题词数目，加1的目的在于避免计算0的对数。

2.2 跨维度关联影响

基金项目与科学论文之间的关联跨越两个不同的维度。以往的研究工作多通过论文中标注的基金项目信息建立二者之间的关联关系。这种外在的形式特征只能获得基金项目与其直接产出成果之间的关联信息。但是科学知识体系是一个关联错综复杂的动态系统，一处细小的改变可能会引起更大范围的变化。近年来，深度神经网络的突破，使得人工智能对多个行业或领域产生的巨大影响是这方面的一个典型代表。加之本研究重点关注一个较小领域的基金项目对其所隶属更大领域的科学研究产生的影响，因此采用文本主题建模的方法获取更细粒度的关联。

本研究通过主题相似度识别基金项目与科学论文之间的关联关系。科学研究中，基金项目立项前需要一定的研究积累，立项后会产出相应的研究成果。因此在已识别的关联关系基础上，结合时间先后顺序对关联影响的作用方向进行判定。即基金项目A与论文B主题相似

时，如果基金项目A的时间在论文B之前，则视为基金项目A对论文B产生影响。反之，则视为论文B对基金项目A产生影响。研究工作将通过多维度信息的综合分析，对基金项目与相关领域科学研究的关联影响进行分析。

3 研究方法与流程

3.1 数据采集

本研究共涉及3个数据源。①通过自主研发的爬虫工具，从国家自然科学基金共享服务网（科技成果信息系统）爬取2008—2014年立项的“交通土建工程”（E0807）所有类型的基金项目数据（截至2019年11月，科学基金共享服务网仅提供2014年及以前立项的项目信息）。②从科学网人工获取2015—2017年该领域自然科学基金立项信息。③依托Web of Science核心合集数据库，检索式为WC=“Transportation”，文献类型限定为“Article”，时间跨度选择2008—2017年，获取国际学术界在交通运输领域的科学论文数据。原则上讲，“交通土建工程”应该是“Transportation”（交通运输）的下位词，如此选择的目的在于从一个较小领域的基金项目探测其对更大学科领域的影响作用。对国家自然科学基金项目和论文数据进行相应的预处理后，最终得到有效的国家自然科学基金项目数据集和论文数据集（以下简称“基金集与论文集”），分别包含1 140项基金项目和34 825篇论文。

为了能够对基金项目和科学论文之间的内在关联关系的特征变化进行跟踪分析，研究工作以每2个自然年份为1个时间刻度，将基金集与论文集中的数据划分为5个时间窗口。获得基金项目数量与论文数量的时间序列如图1所示。

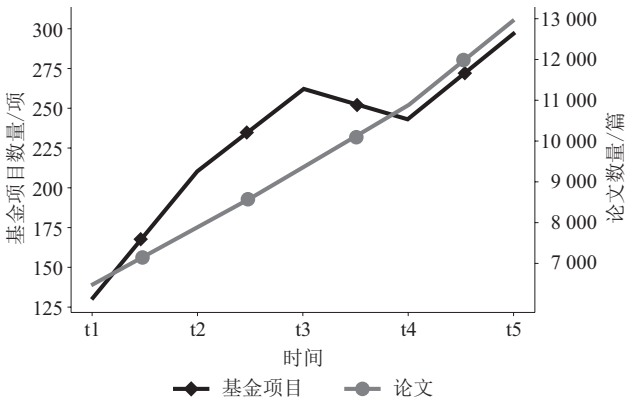

图1 基金项目与论文数量时间序列

图1反映了t1～t5时间序列（2008—2017年）上国家自然科学基金立项与科学论文发表数量的情况。可以发现，“交通土建工程”领域国家自然科学基金项目与“交通运输”领域论文数量呈总体增长趋势。其中，基金项目在t4时间窗口有所回落（242项），其后继续增长。而图中代表论文发表数量的折线则在时间序列上呈现出持续增长的趋势，且论文发表数量并未受到t4时间段基金项目数量回落的影响。

3.2 关键词网络构建

每项（篇）基金项目（论文）中的关键词是由科研人员（作者）所赋予的，旨在突出该科研工作的核心思想与主题。研究工作基于所获取的t1～t5全部时间段内的基金集和论文集数据，构建基金项目和论文两种类型的关键词网络。从基金集中提取“关键词”字段用于构建基金关键词网络；从论文集中提取“DE”字段用于构建论文的关键词网络。两种类型的关键词网络构建方式相同，均是基于关键词的共现关系构建。研究中构建的关键词网络为无向2值网络，即网络中各个关键词之间仅考虑是否存在关联。基金项目和论文在时间序列上的关键词网络的特征指标如表1所示。

表1中的数据显示，基金项目关键词网络的节点数量、连边数量与基金项目立项数量表现出相同的变化趋势，即总体上呈现逐期增加的趋势，仅在t4时间窗口略有回落。论文关键词网络的节点数量和连边数量则在时间序列上单调递增，与该领域发文数量的变化相契合。此外，基金项目关键词的网络密度、聚类系数、特征路径长度均在时间序列上存在波动起伏，而论文关键词网络的相应指标则呈现单调递减的趋势。这一现象可能由于基金项目数量的波动导致，但同时也在一定程度上反映了基金项目关键词更体现项目本身的创新性，而论文关键词网络则更好地表现出领域知识之间的关联性。

3.3 核心主题识别

基金项目或论文中标注的关键词是科学文献的显性外在形式特征，而作为基金项目或论文重要组成部分的摘要则包含大量隐含的语义信息。研究工作分别提取基金项目和论文的标题（TI）、关键词（DE）、摘要（AB），进行主题建模，挖掘基金项目与论文文本中潜

表1 基金项目与论文关键词网络特征指标

	Time	N.F	Node	Edge	Den	Clu	APL
基金项目	t1	130	526	1 101	0.007 97	0.615 72	2.820 48
	t2	210	697	1 511	0.006 23	0.609 04	5.429 97
	t3	262	1 052	2 318	0.004 19	0.508 10	4.756 54
	t4	242	927	2 152	0.005 01	0.363 34	5.392 92
	t5	296	1 179	2 780	0.004 00	0.411 99	5.493 61
	Time	N.P	Node	Edge	Den	Clu	APL
论文	t1	4 180	12 970	41 302	0.000 49	0.250 20	5.068 01
	t2	5 294	15 960	53 373	0.000 42	0.229 75	4.812 98
	t3	6 694	19 758	72 316	0.000 37	0.201 25	4.577 68
	t4	8 223	24 298	100 835	0.000 34	0.196 67	4.385 13
	t5	10 434	30 569	130 766	0.000 28	0.170 90	4.357 81

注：Time表示时间窗口；N.F表示基金项目数量；N.P表示论文数量；Node表示节点数量；Edge表示连边数量；Den表示网络密度；Clu表示聚类系数；APL表示特征路径长度

在的语义信息。

研究工作对基金集和论文集数据进行预处理，将标题、关键词和摘要进行合并构成用于构建主题的文本，通过分词、词性标注等处理把文本数据的初始信息按照语义规则以词为单位进行拆分。文本中包含的无用词对主题识别没有任何意义，对此使用自设的中英文停用词表将其过滤，从而保证研究的精准性。研究中，采用一致性指标检测基金项目和论文在t1～t5时间窗口中的最佳主题数量，选取一致性分数最高且趋于平缓的点所对应的最佳主题数量，超参数$\alpha=\frac{50}{k}$、$\beta=0.01$，迭代次数为1 000次，得到一致性结果如图2所示。

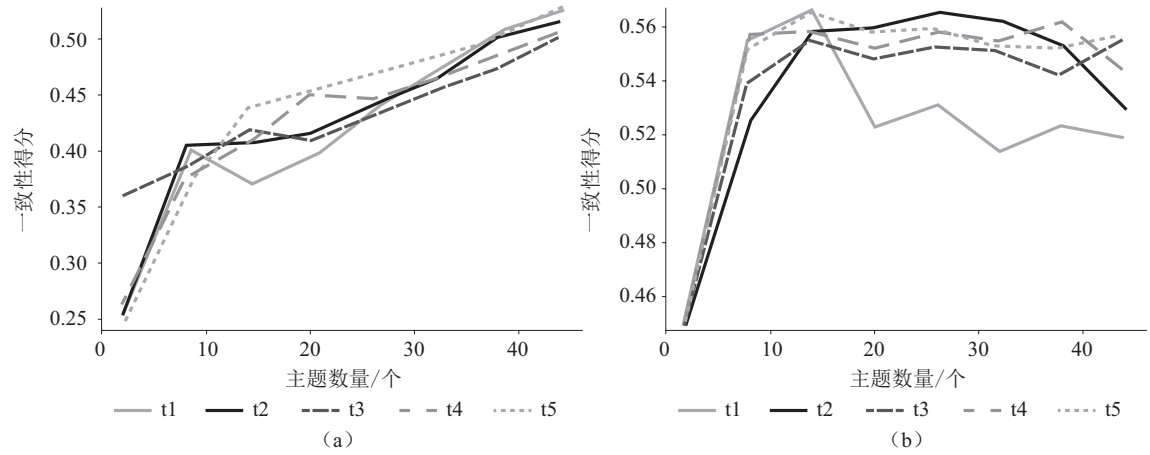

图2 基金项目与论文一致性得分

图2为基金项目与论文在t1～t5时间窗口中，不同主题数量所得到一致性得分情况。图2（a）和图2（b）分别为基金项目和论文对应的图示。一致性分数代表文本主题的连续性程度，分数最高且趋于平稳的点所对应的主题数目表示文本主题连续性最好。结合一致性分数以及多次实验，经人工判读最终确定的基金集和论文集的主题数量如表2所示。

表2列示了本研究中基金项目与相关领域论文所选取主题数量。主题数量的选取与文本数量没有关联，在

表2 基金项目与论文的最佳主题数量

Time	Number of topic (F)	Number of topic (P)
t1	19	14
t2	19	26
t3	14	20
t4	24	26
t5	14	27

注：Time表示时间窗口；Number of topic (F)表示基金项目主题数量；Number of topic (P)表示论文主题数量

t1～t5全部时间窗口中，基金项目共涉及90个主题，论文涉及113个主题。

3.4 主题关联构建

为了能够进一步探析基金项目与科学论文之间细粒度的关联信息，研究工作依托主题之间的相似程度，构建基金项目与论文之间的关联关系。由于本研究的数据来自不同的数据源，基金集中的文本语言为中文，而论文集中的文本语言为英文。因此，为了能够更加准确地将两类数据集内的主题进行相似性测算，研究工作基于中国规范术语数据库将论文核心主题词与数据库内相应领域的词进行对齐，将论文主题词转换成中文进行计算。进而使用余弦相似度[25]在同一语种空间计算两个数据集中各个主题之间的相似程度。主题A和主题B之间的余弦相似度公式如（3）所示。

$$CS=\cos(\theta)=\frac{A\cdot B}{\|A\|\|B\|}=\frac{\sum_{i=1}^{n}A_i\times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (3)$$

公式（3）中，A_i、B_i分别代表两个主题向量A和B的各分量；CS取值范围为[0，1]。据此可以得到t1～t5时间窗口中基金项目与论文的所有主题间关联，包括基金项目与基金项目之间、基金项目与论文之间、论文与论文之间的3种主题关联。由于本研究旨在探索较小领域基金项目对所隶属的较大学科领域科学研究的影响作用，故只保留基金项目与论文之间的主题关联。主题间相似度的大小表示两个主题之间的关联强度，由于学科相关性原因，计算结果显示多数主题之间均存在一定的相似性。为保证研究的精准性，考虑到数据采集范围以及文本类型的异质性，研究工作参考文献[13]的阈值设定原则，过滤掉0值以及相似度低于0.1的关联，从而保留具有分析意义的关联强度较大的关联关系。在此基础上，结合时间序列上先后顺序确定基金项目与论文之间的关联影响关系。

4 结果分析

4.1 基于关键词的关联影响分析

通过对关键词进行分析可以掌握学科领域的研究热点。关键词网络的节点度值体现了知识之间的关联程度，关键词的度值越大，意味着与越多的关键词存在关联，越能代表领域的核心知识。而关键词的词频可以反映领域的研究热点。因此，研究工作基于前文构建的t1～t5时间窗口下基金项目与论文的关键词网络，分别提取度值和词频排名前10位的关键词加以分析，结果如表3、表4所示。

表3 基金项目中度值、词频前10位关键词

t1		t2		t3		t4		t5	
关键词	Deg	**关键词**	Deg	**关键词**	Deg	**关键词**	Deg	**关键词**	Deg
沥青路面	32	高速铁路	42	沥青路面	62	沥青路面	95	沥青路面	84
交通安全	28	沥青混合料	26	沥青混合料	45	高速铁路	74	高速铁路	64
可靠性	25	沥青路面	23	交通安全	40	交通安全	29	沥青混合料	40
高速铁路	16	路用性能	19	驾驶行为	24	无砟轨道	28	桥梁结构	40
驾驶行为	16	细观结构	18	高速铁路	23	沥青混合料	26	交通规划	31
优化	15	沥青	17	仿真	20	交通流	24	盾构隧道	28
沥青混合料	14	机理	16	细观结构	19	机理	20	无砟轨道	25
路用性能	12	驾驶行为	14	机理	17	重载铁路	18	交通设计	24
沥青混凝土	12	高速列车	12	交通流	16	城市轨道交通	17	道岔	23
水泥混凝土路面	12	无砟轨道	12	高速公路	16	交通规划	17	沥青混凝土	20
t1		t2		t3		t4		t5	
关键词	Fre	**关键词**	Fre	**关键词**	Fre	**关键词**	Fre	**关键词**	Fre
交通安全	8	高速铁路	14	沥青路面	17	沥青路面	26	沥青路面	23
沥青路面	8	沥青混合料	10	沥青混合料	13	高速铁路	21	高速铁路	18
可靠性	7	沥青路面	9	交通安全	12	无砟轨道	9	沥青混合料	12
驾驶行为	5	路用性能	8	高速铁路	7	交通安全	8	桥梁结构	10
高速铁路	4	机理	6	驾驶行为	6	交通流	7	交通规划	9

续表

t1		t2		t3		t4		t5	
关键词	Fre	关键词	Fre	关键词	Fre	关键词	Fre	关键词	Fre
沥青混合料	4	沥青	6	仿真	5	沥青混合料	7	道岔	7
优化	4	高速列车	5	高速公路	5	机理	6	盾构隧道	7
出行行为	3	细观结构	5	机理	5	交通规划	6	交通设计	7
交通仿真	3	安全	4	细观结构	5	城市轨道交通	5	无砟轨道	7
交通管理	3	高速公路	4	车联网	4	高速公路	5	沥青	5

注：t1～t5表示时间窗口；Deg表示度值；Fre表示词频

表3中，基金项目度值排名靠前的关键词与词频排名靠前的关键词大体一致。两者结合分析发现，道路材料与铁路建设是该领域基金项目长期重点支持的研究方向。从时间序列上看，“沥青路面”（t1~t5）、“沥青混合料”（t1~t5）、“沥青混凝土”（t1、t5）、“水泥混凝土路面”（t1）、“沥青”（t2、t5），以及“高速铁路”（t1~t5）、“高速列车”（t2）、“重载铁路”（t4）、“无砟轨道”（t4~t5）、“道岔”（t5）等关键词，表明我国在公路与高速铁路建设方面长期的资助倾向。此外，表3中多次出现的“交通安全”“安全”“可靠性”等关键词，反映出交通安全问题也是我国基金项目的资助重点。

表4列示的是论文中度值与词频排名前10位的关键词。整体上看，度值排名与词频排名同样表现出大体的相似性。在5个时间窗口中“安全”一直作为高度值、高词频的关键词，与安全相关的关键词如“道路安全”“交通安全”在5个时间窗口中均有出现，并且占据

表4 论文中度值、词频前10位关键词

t1		t2		t3		t4		t5	
关键词	Deg	关键词	Deg	关键词	Deg	关键词	Deg	关键词	Deg
仿真	242	仿真	361	仿真	365	道路安全	533	道路安全	708
安全	190	优化	244	道路安全	364	公共交通	444	智能交通系统	599
正交频分复用	174	安全	233	优化	332	优化	384	交通工程计算	575
优化	161	道路安全	192	安全	296	安全	375	公共交通	556
道路安全	155	交通安全	172	交通安全	288	交通工程计算	363	安全	491
损伤	154	驾驶	168	交通运输	260	仿真	362	电动汽车	432
排放	143	柴油发动机	155	公共交通	214	智能交通系统	354	优化	406
驾驶	136	流动性	151	损伤	184	道路车辆	345	交通安全	377
多入多出	122	排放	147	交通	183	道路交通	313	道路交通	367
交通	115	正交频分复用	147	可靠性	174	流动性	309	交通运输	359
t1		t2		t3		t4		t5	
关键词	Fre	关键词	Fre	关键词	Fre	关键词	Fre	关键词	Fre
正交频分复用	49	安全	48	安全	66	安全	52	安全	79
仿真	44	仿真	48	道路安全	44	流动性	50	资源配置	70
优化	30	正交频分复用	43	仿真	44	公共交通	48	中国	62
多入多出	28	优化	35	优化	40	优化	47	优化	62
损伤	24	驾驶	32	正交频分复用	39	道路安全	44	多入多出	57
排放	23	损伤	32	中断概率	38	中国	41	公共交通	56
安全	23	交通安全	28	驾驶	35	仿真	41	功率分配	55
调度	23	认知无线电	27	交通安全	34	驾驶模拟器	39	建筑环境	52
柴油发动机	21	柴油发动机	27	损伤	34	可达性	38	道路安全	49
衰落信道	20	频谱监测	27	公共交通	30	交通安全	36	中断概率	48

注：t1～t5表示时间窗口；Deg表示度值；Fre表示词频

较高的排名。论文中的关键词表明，交通安全、道路安全等问题是该领域国际学术界长期的重点研究内容。同时，“电动汽车”“排放”与“智能交通系统”以及通信传输类（正交频分复用、多入多出、衰落信道）等关键词则凸显出交通运输大学科领域对绿色交通和智能交通的关注。值得一提的是，t4～t5时间窗口中“中国”均作为高频关键词出现，表明近年来中国的交通事业发展受到国际学术界普遍关注。

从关键词的分析结果看，由于基金项目选择较小的“交通土建工程”领域，因此其主要资助方向为道路材料、高速铁路、交通安全；而在更大的交通运输领域，科研论文表现出以安全为最主要的研究方向，同时包括绿色交通与智能交通。值得注意的是，交通安全同样作为“交通土建工程”领域基金项目的重点资助对象之一，说明我国在“交通土建工程”小领域的基金资助倾向，一定程度上与所隶属的大学科领域的研究重点有相同之处；那么“可靠性”关键词在基金项目与论文中出现的时差（基金项目t1时间窗口，论文t3时间窗口），在一定程度上反映出，我国在较小领域中的基金资助对更大范围的国际学术界以及更大的学科领域产生了积极的引导与促进作用。在t4～t5时间窗口论文关键词中高频出现的“中国”及其排名提升趋势就是一个有力的佐证。

4.2 基于文本主题的关联影响分析

上述基于关键词的关联影响分析中可以初步看到，我国对较小领域的基金项目的资助，一定程度上对国际学术界较大学科领域的科学研究产生积极的引导与促进作用。为了获得更细粒度的证据以及更精准的判断，研究工作针对基金项目与论文的文本，采用LDA主题模型获取文本主题，并计算主题相似度。基于基金集与论文集时间序列上全部主题的相似度测量结果，分别从数量与内容的角度分析二者之间的关联影响。计算得到相同时间窗口基金项目与论文的相似性相关数据，如表5所示。

表5 相同时间窗口基金项目与论文相似性数据

指 标	t1	t2	t3	t4	t5
相似主题数量	7	12	7	12	11
论文主题数量	14	26	20	26	27
相似主题占比	50.00%	46.15%	35.00%	46.15%	40.74%

表5中的数据显示，在t1～t5时间窗口中，基金项目与论文的主题之间满足相似程度的关联共有49对。考虑到论文数据来自更大的学科领域，因此以论文主题数为分母，得到相似主题占比如表中所示。表5中较高的相似主题占比（35.00%～50.00%）表明来自于较小领域的基金项目与所隶属的较大领域的科学论文之间存在较大的相似性，二者之间的存在一定的关联关系。但是基于同一时间窗口的比较仅仅能够说明二者间存在关联，尚不能确定二者之间的影响关系。研究工作进一步结合时间序列的先后顺序，考察二者之间的影响关系。基于时间顺序的影响关系如表6所示。

表6 基于时间顺序的基金项目与论文间影响关系

跨越时间窗口数量/个	不同方向上的影响关系数量/对	
	F→P	P→F
1	26	13
2	18	15
3	13	6
4	6	2

注：F→P表示基金项目对论文的影响；P→F表示论文对基金项目的影响

表6中，整个时间序列上基金项目与论文之间的影响关系共计99对。其中基金项目对论文的影响关系（基金项目在前，论文在后）数量为63对，论文对基金项目的影响关系（论文在前，基金项目在后）36对。基金项目立项需要一定相关的前期积累，而项目研究又会产出相应的研究成果。从数量来看，即使较小领域的基金项目对所隶属的更大领域的论文产生的影响仍然要大于论文对基金项目的影响。从影响的时间长度上来看，基金项目对论文的影响关系中，跨越1～4个时间窗口的影响关系数量分别为26、18、13、6，平均时间长度为1.984个时间窗口（约3.968年）。论文对基金项目跨越1～4个时间窗口的影响关系数量分别为13、15、6、2，平均时间长度为1.917个时间窗口（约3.834年）。从这个意义上讲，基金项目对论文的影响周期更长远一些。研究工作采用桑基图将99对影响关系逐一列示，结果如图3所示。

图3中，由左向右表示基金项目或论文对另一方的影响关系，连边的宽度为主题相似程度，反映影响关系的强弱。这种影响关系在现实中对研究工作的开展产生引导或促进作用。从图中可知，t1时间窗口的基金项目对论文影响较大的主题包括#2、#4、#5、#16（图中

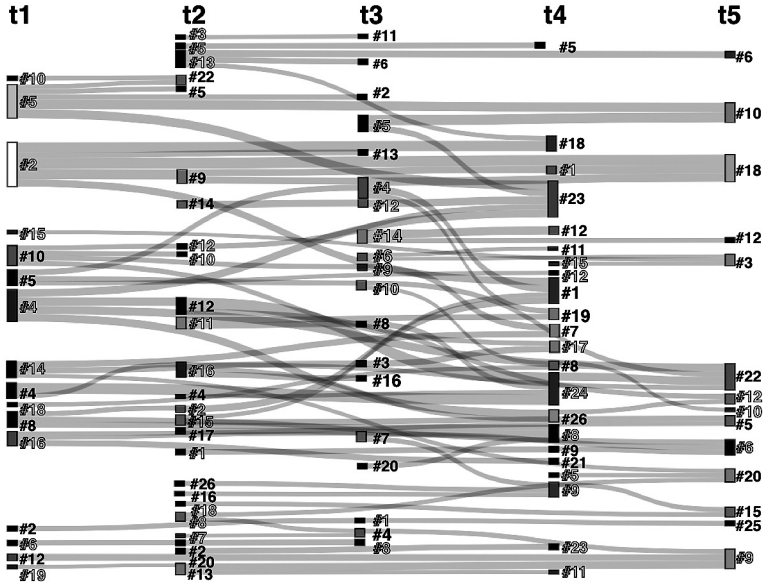

图3 基金项目与论文之间的影响关系

注：空心数字表示基金项目主题；实心数字表示论文主题；#表示主题编号

空心数字显示）等。其中，基金项目#2主题是轨道交通的可靠性方面的内容；#4主题是公路行驶安全方面的内容；#5主题主要是信号控制等智能交通方面的内容；#16主题为尾气污染等环境保护方面的内容。上述前3个基金项目主题对交通运输领域论文的影响蔓延至t2～t5时间窗口，#16主题的影响波及t3～t5时间窗口。此外，其他对后期论文产生较大影响的基金项目主题包括t2时间窗口的#13、#15、#16主题，以及t3时间窗口的#4、#5、#14主题等。其对应的主题内容分别为基础设施可靠性、道路建设材料、安全危险识别、安全风险评估、信号智能控制、轨道交通优化6个方面。影响波及范围均涉及其后至少2个时间窗口。

另外，论文对基金项目的影响主要集中于t1、t2两个时间窗口。t1时间窗口对基金项目影响较大主题为#4、#5、#10（图中实心数字显示）。#4主题是危险驾驶行为方面的内容，#5主题为行人出行安全问题，#10主题为城市公共交通政策方面的内容。3个论文主题的影响范围至少涉及其后的2个时间窗口，但最远只波及到t4时间窗口。t2时间窗口#9、#12主题对基金项目的影响较大，#9主题为可靠性评估的内容，#12主题仍为驾驶行为安全的内容。影响范围同样涉及至少2个时间窗口。

综合基金项目与论文之间的双向影响关系，显然基于文本语义的关联影响分析的识别粒度更细，而且基金项目对论文的影响更大，作用周期更长。其中，在轨道交通、交通安全、绿色交通、智能交通、道路材料等方面，我国的自然科学基金项目都对国际学术界交通运输领域的相关研究产生了引导与促进作用。

5 结论

本文通过基金项目与论文、下位学科与上位学科、关键词与文本主题、网络分析与时间序列分析多维度结合的综合分析，对国家自然科学基金项目与国际学术界科学研究成果之间的关联影响进行识别与分析。综合上述分析结果，研究工作初步得出以下结论。

（1）较小领域的基金项目与所隶属的更大领域科学研究之间存在关联影响关系。关键词网络的分析结果显示，基金项目与科学论文之间存在大量相似或相近的高度值高词频关键词（见表3、表4）；主题分析则显示相同时间窗口的主题相似程度达到35.00%～50.00%（见表5），在时间先后顺序上形成99条相互关联影响关系（见表6），其中以交通安全领域表现最为突出。这些现象与数据说明，即使较小领域的基金项目与其所隶属的更大学科领域的科学研究之间也存在关联影响关系，可为国家在对科技创新的资助上选择重点方向实现以点带面，提供了数据支持。

（2）基金项目与科学论文之间前者对后者的引导与促进作用更大。尽管基金项目立项前的研究积累与立项后的成果产出，在一定程度上决定了基金项目与论文

之间的关联影响是相互的，但是研究中也发现一些关键词在基金项目与科学论文之间存在时差，即基金项目关键词先于论文关键词（见表3、表4）。而主题分析中识别出的99条影响关系（见表6）中，基金项目对论文的影响关系（63条）要远多于论文对基金项目的影响关系（36条）。加之论文关键词中“中国”在时间轴后期的高频出现，都进一步表明我国自然科学基金项目对相关领域科学研究的引导与促进作用。

（3）基金项目与科学论文之间前者对后者的影响持续时间更长。研究中基于主题相似度与时间先后顺序识别出的99条关联影响关系中，基金项目对论文的影响关系的平均作用时间长达1.984个时间窗口（见表6），比论文对基金项目的影响周期更长；而且图3中大部分此类影响关系的作用一直蔓延到时间轴的末端（t5时间窗口）。相对而言，论文对基金项的影响作用则周期较短。这一结果为国家通过基金资助调控科技创新主攻方向，提供了有力的科学支持。

研究中采用的科技信息的多维度分析视角与方法，能够有效挖掘与发现维度间潜在的模式与规律，为科学研究提供更全景化的信息服务。研究中也存在一些局限，基于单一领域分析基金项目与科研成果的特征难免存在些许局限。未来的研究工作将纳入更广泛的学科领域，融合更加多维的信息，探索科学研究与科技创新中更细粒度的模式与规律。

参考文献

[1] BUTLER A. Revisiting bibliometric issues using new empirical data [J]. Research Evaluation, 2001, 10 (1): 59-65.

[2] WANG X W, LIU D, DING K, et al. Science funding and research output: a study on 10 countries [J]. Scientometrics, 2011, 91 (2): 591-599.

[3] BOYACK K W, BORNER K. Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers [J]. Journal of the American Society for Information Science & Technology, 2003, 54 (5): 447-461.

[4] 陈秋怡，刘海波. 科研基金资助投入与高水平国际论文产出研究——基于六国SCI论文的实证分析 [J]. 中国科技论坛，2018 (1)：158-163.

[5] 许鑫，于霜，王立梅. 科学基金对开放存取论文的影响力分析——以SCI收录的自然科学领域论文为例 [J]. 数字图书馆论坛，2019 (5)：26-36.

[6] 许晓阳，郑彦宁，刘志辉. 论文和专利相结合的研究前沿识别方法研究 [J]. 图书情报工作，2016，60 (24)：97-106.

[7] KIM Y G, SUH J H, PARK S C. Visualization of patent analysis for emerging technology [J]. Expert Systems with Applications, 2008, 34 (3): 1804-1812.

[8] BARBOSA S D J, SILVEIRA M S, GASPARINI I. What publications metadata tell us about the evolution of a scientific community: the case of the Brazilian human-computer interaction conference series [J]. Scientometrics, 2017, 110 (1): 275-300.

[9] 化柏林. 多源信息融合方法研究 [J]. 情报理论与实践，2013，36 (11)：16-19.

[10] YAN E J. Research dynamics, impact, and dissemination: a topic-level analysis [J]. Journal of the Association for Information Science and Technology, 2015, 66 (11): 2357-2372.

[11] 张子振，储煜桂，吴小兰. 基于LDA的多源文献主题及其差异研究——以“机器学习”为例 [J]. 情报科学，2019，37 (6)：108-112，150.

[12] 白如江，冷伏海，廖君华. 一种基于多数据源主题对比的科学研究前沿识别方法 [J]. 情报理论与实践，2017，40 (8)：36，43-48.

[13] 刘博文，白如江，周彦廷，等. 基金项目数据和论文数据融合视角下科学研究前沿主题识别——以碳纳米管领域为例 [J]. 数据分析与知识发现，2019，3 (8)：114-122.

[14] 徐红姣，曾文，张运良. 基于Word2vec的论文和专利主题关联演化分析方法研究 [J]. 情报杂志，2018，37 (12)：36-42.

[15] 刘自强，许海云，岳丽欣，等. 面向研究前沿预测的主题扩散演化滞后效应研究 [J]. 情报学报，2018，37 (10)：979-988.

[16] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3 (1): 993-1022.

[17] GRIFFITHS T L, STEYVERS M. Finding scientific topics [J]. Proceedings of the National Academy of Science, 2004, 101 (1): 5228-5235.

[18] 阮光册. 主题模型与文本知识发现应用研究 [M]. 上海：华东师范大学出版社，2018：70-72.

[19] 王静茹，陈震. 基于隐含狄利克雷分布的文本主题提取对比研究 [J]. 情报科学，2018，36 (1)：102-107.

[20] 关鹏，王曰芬. 科技情报分析中LDA主题模型最优主题数确定方法研究 [J]. 现代图书情报技术，2016，32 (9)：42-50.

[21] MIMNO D M，WALLACH H M，TALLEY E M，et al. Optimizing semantic coherence in topic models [C] //Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics，2011：27-31.

[22] ANJIE F，CRAIG M，IADH O，et al. Examining the coherence of the top ranked tweet topics [C] //Proceedings of the 39th InternationalACM SIGIR Conference on Research and Development in Information Retrieval，2016：825-828.

[23] LAU J H，NEWMAN D，KARIMIS，et al. Best topic word selection for topic labelling [C] //Proceedings of the 23rd International Conference on Computational Linguistics，2010：605-613.

[24] RÖDER M，BOTH A，HINNEBURG A. Exploring the space of topic coherence measures [C] //Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM，2015：399-408.

[25] JIMENEZ S，GONZALEZ F A，GELBUKH A. Mathematical properties of soft cardinality：enhancing Jaccard，Dice and cosine similarity measures with element-wise distance [J]. Information Sciences，2016（367/368）：373-389.

作者简介

吕晶，女，1995年生，硕士研究生，研究方向：知识组织与信息分析。
郭思月，女，1996年生，硕士研究生，研究方向：科技信息分析。
滕广青，男，1970年生，教授，通信作者，研究方向：知识组织与信息分析，E-mail：tengguangqing@163.com。
马卓，女，1982年生，副研究员，研究方向：科技信息分析。

Analysis on Impact of Fund Projects on Scientific Research

LV Jing[1] GUO SiYue[1] TENG GuangQing[1] MA Zhuo[2]
(1. School of Information Science and Technology, Northeast Normal University, Changchun 130117, China; 2. Institute of Scientific and Technical Information of Jilin Province, Changchun 130033, China)

Abstract: The multi-dimensional correlation impact analysis of fund projects and scientific research can provide fine-grained insights, which are helpful for the scientific planning of the country's scientific strategy and the formulation of scientific and technological strategies. This article combines data from the National Natural Science Foundation of China and paper data from related fields, and uses LDA for topic modeling. The correlation impacts between fund projects and papers are identifiedbased on topic similarity. The results show that there are correlation impacts between fund projects in smaller field and scientific research in the larger field. Fund projects have a greater role in guiding and promoting scientific papers, and the duration of the effect is longer.

Keywords: Fund Project; Scientific Paper; Topic Modeling; Topic Correlation; Impact Duration

（收稿日期：2019-11-10）