科研人员研究主题的聚焦与迁移研究*

陈立雪 郭思月 滕广青 庾锐
（东北师范大学信息科学与技术学院 长春 130117）

摘要：科研人员的主题偏好分析有助于洞悉科学知识发展的动力机制，引导学科领域科技创新方向。本文采用LDA主题模型进行主题划分，对不同类型科研人员主要研究方向的精专程度和主题迁移进行分析。研究结果表明，高发文科研人员的研究主题相对集中，且在主要研究方向上的精专程度更高，高被引科研人员研究主题的迁移性更突出。

关键词：高发文；高被引；LDA；主题强度；研究方向；主题迁移
中图分类号：G250.2
DOI：10.3772/j.issn.1673-2286.2019.12.002

科研人员的研究主题分析一直受到图书情报学领域的高度重视。科学论文作为科研人员学术成果的重要载体，凝聚了科研人员的智慧，其中包含的大量隐含信息是对科研人员研究主题进行识别的重要依据。近年来，科学知识的更新速度不断加快，各学科间的交叉融合趋势愈加明显。一方面，学科领域内的研究主题不断推陈出新，一些原有的热点主题不断强化，而另一些新的知识逐渐成为流行主题，学科知识体系更呈现复杂性。另一方面，一些科研人员在自己的主要研究方向上坚持始终，也有一些科研人员逐渐呈现研究方向的多样化，甚至还有科研人员热衷于追逐学科领域内新的流行主题。面对这些问题，高发文、高被引等不同类型的科研人员会有怎样的表现，是一个值得深入研究的问题。

本研究采用隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）模型，计算提取不同类型科研人员的研究主题。通过主题强度等指标分析，探测不同类型科研人员的研究偏好，分析其在主要研究方向上的精专程度与主题迁移，为科学发展提供决策支持与参照依据。

1 相关研究综述

图书情报学领域，关于特定学科研究主题的计量与分析由来已久。早期的相关研究主要基于科学论文的关键词等形式特征进行分析[1]，关注的重点包括领域热点识别[2]、主题聚类[3]等多个方面。随着研究工作的开展，研究者不再局限于对研究主题整体进行研究，而是基于研究主题与科研人员之间的关联关系，探索不同科研人员（群）研究主题的模式与特征。谭春辉等[4]采用词频分析和引文分析识别图书馆学领域的核心科研人员和领域研究主题。徐健等[5]基于关键词计算科研人员之间的兴趣相似度，并通过网络聚类发现了科研人员个体在研究主题上的多样性。但是科学论文的关键词只是对论文主题的高度概括和浓缩，虽然能够大致反映出所在论文的主题方向，但其带有的人为主观性难以全面地揭示论文的主题内容。并且，传统的主题划分方法多是基于对论文中的高频关键词进行统计分析之后得到的。由于未涉及文本细节内容或词语所包含的语义信息，使得主题提炼结果显得较为粗糙，而且单纯基于关键词获得的研究主题中，所伴随的数据失真也是一个难以忽略的问题。

为弥补关键词无法对主题进行完全描述这一不足，基于自然语言处理的文本主题挖掘技术被应用于科学计量中。Mane等[6]采用TF-IDF算法从PANS刊载的论文中提取特征词，选择其中突发权重指数最大的前50个词，通过共词分析对论文的主题及其变迁进行研究。尽管TF-IDF能够计算文档中词的重要性，但用

*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。

于提炼文档主题仍显不足。Deerwester等[7]提出潜在语义分析（Latent Semantic Analysis，LSA）模型用于挖掘文档与词语之间隐含的潜在语义关联，但由于其算法复杂度较高，因此并没有得到广泛应用。随后Hofmann[8]对 LSA算法进行改进，在LSA的基础上进行扩展后提出概率潜在语义分析（Probabilistic Latent Semantic Analysis，PLSA）模型。该算法尽管能够在一定程度上降低计算的复杂度，但遇到大规模文本时，同样也会使模型变得庞大以致增加处理难度。Blei等[9]在PLSA基础上提出了LDA模型，通过引入文本的主题分布极大地降低了数据的维度，同时模型的参数空间规模是固定的，这也使LDA模型更适用于大规模文本集。现有的研究成果已经证实LDA在挖掘文本主题方面比LSA和PLSA模型具有更大的优势[10-11]。此后的研究中，LDA被大量应用于科学文献的文档主题提取。Griffiths等[12]采用LDA主题模型基于PNAS文集中的论文摘要进行主题提取，识别其中的热点主题和冷门主题及其主题强度变化。Wang等[13]结合LDA模型和TOT（Topics Over Time）模型对NIPS文献数据集的隐含主题揭示后进行了主题偏移的相关研究。国内学者李湘东等[14]通过LDA的主题提取结果及JS散度来探测科技期刊的主题在强度和内容两方面的演化，并对不同时间窗口的主题稳定性做出相应分析。廖列法等[15]在LDA主题建模的基础上，引入IPC分类号度量专利文本的技术主题强度。

综上所述，学术界在基于关键词的领域主题分析、针对作者的研究主题识别、基于LDA的主题挖掘等方面都积累了一定的成果。特别是在科研人员的研究主题分析方面，Nature旗下的*Scientific Reports*曾刊文指出，科研人员在进行科学研究时更容易被流行的主题所吸引[16]。近年来，学术界追逐流行热点的风气仍然存在。鉴于此，本研究采用LDA模型提取科研人员的研究主题，对高发文、高被引以及随机抽取的3类科研人员的主题特征进行分析，通过主题强度和主题变异系数探测不同类型科研人员在主要研究方向上的精专程度与主题迁移。

2 相关基础理论

2.1 科研人员的研究主题

针对科研人员研究主题的相关研究，研究者通常将科研人员已发表科研成果的主题划分作为科研人员研究主题的重要依据。对科研成果主题进行划分之后，将主题结果对应到每位科研人员，以此作为科研人员的研究主题。实际上，在大多数科研人员的整个科研生涯中，其研究主题并单一的。因此，本研究将科研人员所发表的科学论文进行主题划分后的结果作为科研人员的研究主题，并将出现频次最多的主题作为科研人员的主要研究方向进行重点分析。考虑到不同类型科研人员在其主要研究方向上的执着程度可能会有差别，研究工作从发文数量、被引数量与随机抽取3个维度对科研人员进行分类，探查不同类型科研人员在研究方向以及主题变换方面的模式特征。

研究中的主题划分采用目前比较主流的LDA主题模型[9]。LDA主题模型打破了传统空间向量“文档－词”的模式，将文档直接映射到主题空间上，是基于“词汇、主题、文档”的三层贝叶斯模型。其主要思想是将文档集合中的每篇文档的主题以概率的形式给出，而主题就是词汇的概率分布。采用主题描述文档有效地解决了维度灾难的问题，同时也克服了空间向量模型的缺点。由于主题是文档内容的分类聚集，因此LDA可以很好地模拟大规模语料的语义信息。

2.2 主题强度与变异系数

不同主题科学论文的出现频率反映了科研人员对不同主题的关注程度，在一定时期内科研人员所发表的论文中，如果该科研人员在某一主题下发文频率越高，则表明该科研人员对这一主题关注程度越高。本研究采用主题强度来描述某一时段科研人员所发表论文的主题热度[17]。某一时间窗口中，科研人员所发表的某一主题的论文数量越多，则这一时间窗口该主题的强度就越大，见公式（1）。

$$\theta_z^t = \frac{\sum_{m=1}^{M_t} \theta_z^m}{M_t} \quad (1)$$

公式（1）中，θ_z^t表示在t时间窗口某一科研人员在主题z的主题强度，M_t表示t时间窗口该科研人员发表的所有主题的论文数量，θ_z^m表示的是主题为z的第m篇文献，$\sum_{m=1}^{M_t} \theta_z^m$表示在$t$时间窗口该科研人员所发表的主题为$z$的论文数量的总和。

为了考察科研人员主题强度的波动情况，本研究在主题强度的基础上，借用统计学中变异系数的思想计算主题变异系数，对科研人员研究主题的稳定性进

行衡量，具体计算公式如下。

$$V = mean\left(\frac{\sigma_{\theta_z^t}}{\overline{X}_{\theta_z^t}}\right) \quad (2)$$

根据公式（1）所计算的某一科研人员在不同时间的某一主题强度值θ_z^t，基于该科研人员这一主题在不同时间窗口的主题强度值，计算其标准差$\sigma_{\theta_z^t}$和平均值$\overline{X}_{\theta_z^t}$。然后对每类科研人员计算平均值，以得到该类科研人员的主题变异系数V。较低的变异系数说明主题的波动性较小，稳定性好；相反的，主题变异系数值越大，则说明主题波动较大，稳定性差。

3 研究方法和流程

3.1 数据的采集

本研究以Web of Science核心数据库作为基础数据来源，根据《期刊引证报告》（JCR，2018），选定“INFORMATION SCIENCE & LIBRARY SCIENCE”学科中影响因子排名前十位的权威期刊进行文献检索，检索日期为2019年12月5日，检索时间段为2006—2019年，将文献类别限定为“Article”，语种限定为“English”，最终得到8 644篇有效文献，相关信息如表1所示。

表1 期刊论文数据汇总

序 号	期刊名称	影响因子	论文数量/篇
1	*International Journal of Information Management*	5.063	1 041
2	*Journal of Computer Mediated Communication*	4.896	493
3	*Journal of Knowledge Management*	4.604	715
4	*Mis Quarterly*	4.373	606
5	*Government Information Quarterly*	4.311	736
6	*Journal of the American Medical Informatics Association*	4.292	1 947
7	*Information Management*	4.120	885
8	*Journal of Strategic Information Systems*	4.000	245
9	*Information Processing Management*	3.892	1 130
10	*Journal of Informetrics*	3.879	846

表1列示了数据集合中的期刊名称、期刊的影响因子以及各个期刊所刊发的论文数量。从中可以大致看出，尽管各个期刊的发文量相差较大，各刊的影响因子却是图书情报学领域排位靠前的，能够较好地代表该学科的发展。表中的8 644篇论文共由1 898位科研人员完成，单个人员的最高发文量为72篇，最低发文量为1篇。研究工作将表1中所有期刊论文作为本研究中科研人员研究主题建模的原始数据。

3.2 代表性科研人员提取

不同的科研人员在学术界的表现存在很大的差异。传统文献计量学往往采用发文量、被引量等指标衡量科研人员的学术贡献水平。科研人员在某一时间段的总发文量在一定程度上反映了其研究的活跃程度[18]；而所著论文的被引用频次在一定程度上反映该科研人员成果的学术价值[19]。因此，研究工作选择发文量和文献被引频次两种指标，从两个不同的考核维度分别筛选两组不同的科研人员，即高发文科研人员和高被引科研人员。同时，采用有放回随机抽样的方法在1 898位科研人员中随机抽取10组科研人员作为参照。由于本研究更关注科研人员研究主题的迁移变化情况，为了确保随机抽取科研人员的可参照性，在进行抽样时舍去了发文量低于2篇的科研人员。据此得到“INFORMATION SCIENCE & LIBRARY SCIENCE”学科领域影响因子排名前十位的期刊中高发文科研人员、高被引科研人员、随机科研人员及其对应的文献数量，结果如表2所示。

限于篇幅的原因，表2仅列示了一组随机抽取的科研人员及其发文数量。从表2中的数据可以发现，科研人员的发文数量较少并不意味其被引频次越少，如作者ELLISON N B并不属于本研究中的高发文作者，但其所撰写论文的被引频次最多。并且，不同类型科研人员的重合率很低，具有良好的可比较性。

3.3 主题模型构建

研究工作将所获取的8 644篇有效科学论文中每一篇论文的标题、关键词和摘要经过一系列的分词、去停用词等预处理之后作为一个文档d，构成训练文档集合D作为LDA模型输入的语料，由此获得整体文献集的研究主题。在这个过程中，对于主题数的确定是主题划分的一个关键步骤，不同学者给出了一些不同的主题数选取指标，Blei等[9]曾提出用基于困惑度的方法来确定

表2 不同类型作者与文献数量

高发文科研人员姓名	文献数量/篇	高被引科研人员姓名	文献数量/篇	随机科研人员姓名	文献数量/篇
BATES D W	72	ELLISON N B	3	DELEGER L	6
BORNMANN L	52	BOYD D M	1	ZHANG Y	20
THELWALL M	48	BATES D W	72	MOON Y J	2
XU H	44	VENKATESH V	24	ROSE J	3
ROUSSEAU R	42	LAMPE C	3	GUSMAO D E	3
WRIGHT A	41	STEINFIELD C	3	BOH W F	4
HRIPCSAK G	39	RAI A	14	MAMYKINA L	4
ABRAMO G	37	CHUTE C G	29	GUPTA A	18
DENNY J C	35	WETZELS M	3	MUTZ R	11
SITTIG D F	34	PAVLOU P A	10	EGGHE L	27

主题个数，困惑度越小，模型泛化能力越强。Teh等[20]提出基于狄利克雷过程的HDP法来自动确定主题数，采用此方法无须预先确定主题数，主题可由数据生成再通过数据反向将其推出。曹娟等[21]基于主题之间的相似度计算主题向量之间的余弦距离、KL距离等来确定主题个数。但上述方法在确定最优主题数量时，或多或少存在些许弊端，如在使用困惑度指标确定主题数量时，其主题内容存在冗余现象。因此，本研究采用Coherence Score函数评价模型以获取最优主题数[22]，选择最高值的一致性分数可以提供更加合理的主题数量。以主题数量为横坐标，一致性得分为纵坐标，根据计算结果绘制一致性得分的折线图，如图1所示。

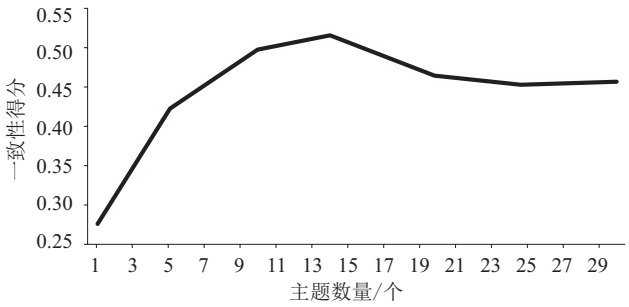

图1 一致性得分折线图

由图1中的一致性得分折线可以看出，当主题数量为14个时具有最高的一致性得分，因此设定主题个数为14，经过最大100轮次的迭代过程，模型经过训练后获得初步的主题训练结果，提取结果中的主题、主题词及相应的文档数量，如表3所示。

表3列示了采用LDA模型计算得到的14个研究主题及其主题词。根据14个研究主题的概率词项进行推理，可以为每个主题制定相应的标签。即topic0：企业管理。topic1：文本分析。topic2：用户行为。topic3：信息检索。topic4：医疗信息。topic5：健康护理。topic6：用户服务。topic7：信息安全。topic8：知识管理。topic9：信息传播。topic10：社群分析。topic11：引文分析。topic12：系统开发。topic13：公共信息。在此基础上，以每篇论文所对应的最大主题概率分布为依据，将8 644篇论文分别归属到一个对应的主题下，然后汇总不同主题的论文数量。表3中的论文数量可以直观地反映出图书情报学领域中发文量最多的主题所代表的学科热点方向，如文本分析（topic1）、医疗信息（topic4）、引文分析（topic11）、系统开发（topic12）等。

4 研究结果

4.1 科研人员与研究主题关联分析

科研人员的研究主题由其所发表论文的主题所反映，同一科研人员多篇论文隶属多个主题则意味着该科研人员的研究主题比较分散。研究中将科研人员所发表的论文和基于LDA模型生成的该领域14个主题进行对应。以主题和科研人员为网络节点，以科研人员在某一主题下发表论文为连边，分别构建高发文、高被引以及随机抽取的科研人员的“科研人员-主题”2-模网络，结果如图2所示。

图2中，（a）（b）（c）分别为高发文、高被引以及随机抽取的科研人员的“科研人员-主题”2-模网络。其中，方形节点代表研究主题，圆形节点代表科研人员，连边的粗细表示边权重的大小，即科研人员在某一主题发文数量的多少。

表3 LDA主题分类结果

主题标签	主题词	文档数量/篇
topic0（企业管理）	绩效、业务、企业、价值、社区、关系、能力、效果、资源、战略	600
topic1（文本分析）	文档、模型、体系、术语、文本、性能、概念、分类、词语、评价	1 044
topic2（用户行为）	用户、消费者、效果、行为、信任、顾客、学习、产品、互联网、网站	709
topic3（信息检索）	数据、网络、用户、搜索、任务、查询、数据、图像、系统、技术	517
topic4（医疗信息）	病人、系统、记录、资料、她、结论、研究、警戒、医院、健康	972
topic5（健康护理）	健康、团体、信息、数据、护理、标准、学生、访问、实践、课程	458
topic6（用户服务）	服务、制度、技术、模式、采用、因素、质量、学习、用户、接受	513
topic7（信息安全）	风险、安全、隐私、自我、信息、电子邮件、回复、儿童、游戏、控制	149
topic8（知识管理）	知识、管理、学习、发现、组织、转移、共享、团队、文化、创造	783
topic9（信息传播）	信息、问题、交流、学习、来源、文章、消息、新闻、用户	372
topic10（社群分析）	网络、媒体、事件、分析、模式、链接、沟通、社区、集群、关系	267
topic11（引文分析）	引文、索引、科学、论文、作者、影响、领域、出版物、期刊、数量	907
topic12（系统开发）	过程、系统、项目、案例、管理、开发、软件、实施、技术、学习	826
topic13（公共信息）	政府、政策、爱思唯尔、互联网、公民、服务、信息、学习、代理、参与	527

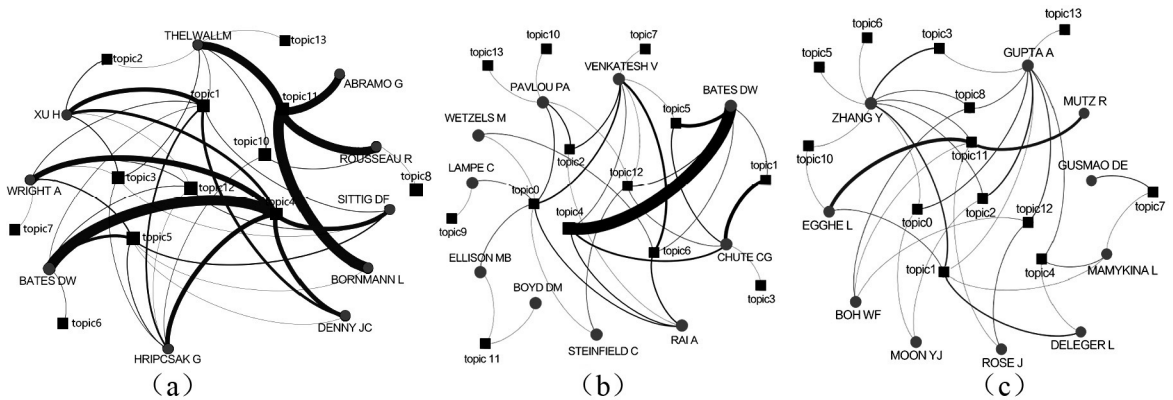

图2 “科研人员-主题”2-模网络

从图2（a）中可以发现，12个主题节点中有7个主题节点位于网络的中心区域，这些主题节点的度值都大于或等于3，而且网络中存在多条高权重的连边，说明这些主题在高发文科研人员群体中非常受到青睐。其中，文本分析（topic1）、医疗信息（topic4）、健康护理（topic5）与超过半数的高发文科研人员之间存在连边，一定程度上说明这3个主题是高发文科研人员的热点研究主题。同时，引文分析（topic11）与医疗信息（topic4）两个主题的边权重合计值分别高达164和156。这意味着高发文科研人员的科研产出多与这两个研究主题相关，且与整个学科中所占比例最高的主题基本吻合，印证了高发文科研人员的研究主题在一定程度上代表了整个学科领域的热点研究方向。

图2（b）的2-模网络中，主题节点数量达到13个，多于图2（a）中高发文科研人员涉及的主题数量。这一现象初步说明，高被引科研人员的研究主题相比高发文科研人员而言更加分散。从主题节点的度值来看，度值大于或等于3的主题节点数为6个，即在主题总数量增加的同时高度值主题节点数量却减少。同时，仅有1个主题（topic0）与超过半数的高被引科研人员之间存在连边，进一步说明高被引科研人员的研究主题相对于图2（a）中的高发文科研人员更加分散。此外，图2（b）中表现突出的高权重连边数量仅有1条，说明大多数高被引科研人员并不集中于某单一主题方向。

图2（c）展示的是一组随机抽取的“科研人员-主题”2-模网络。该网络同样拥有13个主题节点，说明随机科研人员的研究主题也是相对分散的。该网络中度值大于或等于3的主题节点数为7个，在相对于图2（a）

主题总数增加的同时高度值主题节点数与之持平，但是与半数以上随机科研人员存在连边的主题节点数量为0，并且网络中并不存在突出的高权重连边。这一现象表明，随机抽取的科研人员相对于高发文科研人员而言，不但研究主题分散，而且研究方向更具有灵活性。

为了能够获得更清晰的对比结果，研究工作对于高发文、高被引以及10组随机抽取的科研人员的2-模网络的基本特征指标进行测算，相关结果如表4所示。

表4 2-模网络基本特征指标

基本统计指标	高发文科研人员	高被引科研人员	随机科研人员*
节点数	22	23	22.50
连边数	42	35	23.20
网络密度	0.35	0.27	0.21

注：*为10组随机抽取的科研人员的“科研人员-主题”2-模网络的平均值

表4的数据显示，高发文科研人员的2-模网络的节点数最少，随机科研人员次之，高被引科研人员2-模网络的节点数最多。通过核查节点性质发现，网络节点数量差异的原因主要由主题节点数量变化所导致。因此，从3类科研人员涉及的主题节点数量的角度看，高发文科研人员的研究主题相对集中，而高被引科研人员的研究主题则相对分散。此外，表4中的连边数量与网络密度指标显示，高发文科研人员所对应的2-模网络连边数量最多且密度最大，随机抽取的科研人员对应的2-模网络连边数量最少且密度最小。2-模网络的网络密度指的是网络中实际连接的边数与网络中节点间可能存在的最大连边数量的比值。在本研究中这一指标反映研究主题集合与科研人员集合之间联系的紧密程度。由此可以得出，尽管同一位高发文科研人员可能会涉及多个研究主题，但是同一研究主题也被多个高发文科研人员共同关注。结合每个研究主题的发文数量可以发现，图2（a）中边权重合计数最高的引文分析（topic11）与医疗信息（topic4）两个主题所包含的论文数量之和，超过10位高发文科研人员发文数量的70%，从主题发文量的层面进一步说明高发文科研人员的研究兴趣更为集中。

4.2 主要研究方向与精专程度分析

出于对不同类型科研人员的主要研究方向及其精专程度进行考察的目的，研究工作将每位科研人员所发表论文中出现频次最高的主题作为该科研人员的主要研究方向，在此基础上对不同类型科研人员的主要研究方向的分布特征进行对比分析，获得高发文、高被引以及随机抽取（1组）的科研人员的主要研究方向的堆积条形图，如图3所示。

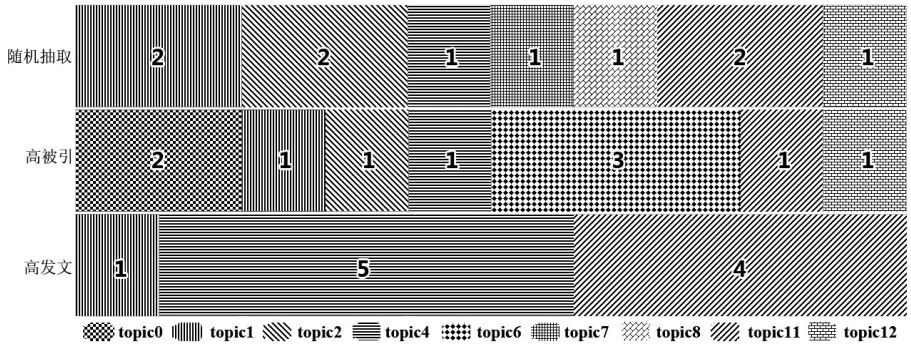

图3 不同类型科研人员的主要研究方向

注：图中数字代表该研究方向下科研人员数量

图3下部显示，高发文科研人员中，有5位科研人员将医疗信息（topic4）作为自己的主要研究方向，4位科研人员将引文分析（topic11）作为自己的主要研究方向，仅有1位科研人员将文本分析（topic1）作为自己的主要研究方向。由此可以看出，高发文科研人员的主要研究方向最为集中。在图3中部高被引科研人员中，有2位科研人员将企业管理（topic0）作为自己的主要研究方向，3位科研人员将用户服务（topic6）作为自己的主要研究方向，还有5位科研人员分别将文本分析（topic1）、用户行为（topic2）、医疗信息（topic4）、引文分析（topic11）和系统开发（topic12）分别作为自己的主要研究方向。显然，高被引科研人员相对于高发文科研人员在主要研究方向上更分散。图3上部为一组随机抽取的科研人员的主要研究主题数据，有2位科研人员将文本分析（topic1）作为自己的主要研究方向，2位科研人员将用户行为（topic2）作为自己的主要

研究方向，2位科研人员将引文分析（topic11）作为自己的主要研究方向，还有4位科研人员分别将医疗信息（topic4）、信息安全（topic7）、知识管理（topic8）和系统开发（topic12）作为自己的主要研究方向。显然随机抽取的科研人员的主要研究方向最为分散。由此可以发现，即使以频次最高的主题作为科研人员的主要研究方向进行测度，同样显示出高发文科研人员比高被引科研人员的研究方向更为集中。这一点与前文2-模网络分析的结果相一致。

为了进一步对不同类型科研人员在各自主要研究方向上的精专程度进行对比分析，这部分研究采用公式（1）计算各类型科研人员中的每位科研人员在不同时间窗口（以每个自然年度为一个时间窗口）中主要研究方向的主题强度值；然后计算单一科研人员在整个时间序列上主题强度的均值作为该科研人员的主题强度；最后，计算某类科研人员中所有人员的主题强度的均值作为该类科研人员的主题强度值。为了更加客观地呈现计算结果，在进行主题强度的计算时，对随机抽取的10组科研人员均进行计算，以组间平均值作为随机科研人员的主题强度值。计算结果如表5所示。

表5 不同类型科研人员主要研究方向的主题强度

作者类型	高发文科研人员	高被引科研人员	随机科研人员*
主题强度	0.724	0.625	0.660

注：*为10组随机抽取的科研人员的主题强度的平均值

表5中的数据显示，高发文科研人员的主题强度值最高。这意味着高发文科研人员群体在进行科学研究时，比较侧重于自己的主要研究方向，而在其他研究主题上发表的文章数量相对较少，研究的精专程度较高。高被引科研人员群体的主题强度值最小，说明高被引科研人员对于非自己主要研究方向的其他研究主题，相比其他两类科研人员有更多的涉及，因此主题强度值相对较低，在主要研究方向上的精专程度不如其他两类科研人员。随机科研人员的主题强度值在3类科研人员中居中，结合前文的网络分析可以发现，尽管随机抽取的科研人员研究主题比较分散，但是该类科研人员在各自主要研究方向上的精专程度要略高于高被引科研人员。

4.3 主题稳定性与主题迁移分析

出于对不同类型科研人员主要研究方向的稳定性及其主题迁移进行考察的目的，进一步采用公式（2）计算不同类型科研人员主要研究方向的主题变异系数，如表6所示。

表6 不同类型科研人员的主题变异系数

作者类型	高发文作者	高被引作者	随机作者*
主题变异系数	0.374	0.665	0.650

注：*为10组随机抽取的科研人员的主题变异系数的平均值

从表6中主要研究方向的主题变异系数可以发现，高发文科研人员的主题变异系数最低，说明此类科研人员的主要研究方向在整个时间周期（2006—2019年）中比较稳定。高被引科研人员和随机科研人员的主题变异性系数值都比较高，表明这两类科研人员群体的主要研究方向的稳定性都较差。其中高被引科研人员群体主要研究方向的主题变异系数略高于随机科研人员，即高被引科研人员的主要研究方向并不稳定，甚至比随机抽取的科研人员还要略显欠佳。

研究工作进一步基于每位科研人员在不同时间窗口主要研究方向的主题强度，分别按照科研人员类型绘制河流图，动态考察不同科研人员在时间序列上的主要研究方向变化及其迁移情况。由于部分科研人员发表的成果在自然年度上并非连续的，当某一科研人员在某时间段内（1个或连续多个时间窗口）没有论文发表且该时段前后的主题强度均不为0时，则将该时段的主题强度用前一时间窗口的主题强度值进行填充，以此区别于科研人员在该时段发表论文但研究主题产生迁移的情况。所得结果如图4所示。

图4（a）为高发文科研人员主要研究方向的主题强度河流图，每条支流在不同时间窗下的宽度变化能够反映该科研人员在某一段时期内是否发生兴趣转移。从图4（a）中可以发现，有2位高发文科研人员（BA为BATES D W、TH为THELWALL M）各自的主要研究方向的主题强度在时间序列上出现了“断流”现象。表明这两位科研人在断流处曾发生主题迁移，研究主题偏离了其主要研究方向。除此之外的其他高发文科研人员在自己的主要研究方向上都保持了较好的连续性，且这些科研人员在各自主要研究方向上所持续的时间基本都长达10年，甚至更久。此外，高发文科研人员群体虽然个体支流宽度偶有起伏，但是该群体总体河流宽度相对平稳，说明高发文科研人员在其主要研究方向上总体表现出较高程度的稳定性。

图4（b）和图4（c）分别是高被引科研人员和随机抽取的科研人员（1组）的主题强度河流图。从中不难发

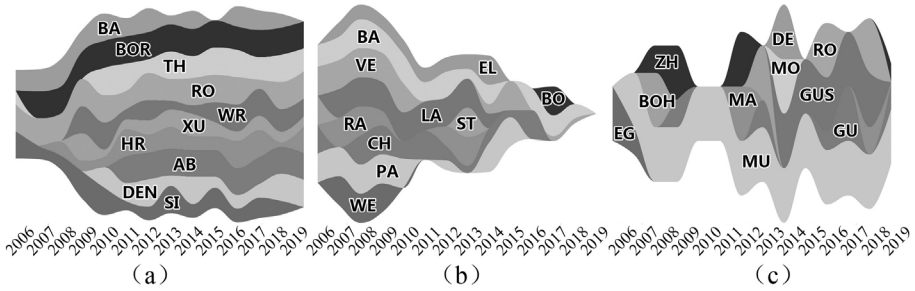

图4 科研人员主要研究方向的主题强度时序变化

注：图中字母为科研人员姓名缩写，全称见表2

现，高被引科研人员与随机科研人员在时间序列上分别发生7次和5次“断流”。说明这两类科研人员发生主题迁移的概率明显高于高发文科研人员，尤其以高被引科研人员的表现最为突出。尽管随机抽取的科研人员在自己的主要研究方向上的主题连续性略高于高被引科研人员，但是单一个体较高的“断流”次数显示出随机抽取的科研人员发生主题迁移的周期更短。从两类科研人员河流的总体宽度变化还可以发现，高被引科研人员在时间轴后期河流的总宽度明显收窄。说明随着时间推移，高被引科研人员群体在自己的主要研究方向上兴趣逐渐淡化，研究主题发生迁移且很少恢复。

5 结论

本文基于图书情报学学科领域影响因子排名前十位学科期刊的论文数据，从中抽取高发文和高被引排名前十位的科研人员和10组随机抽取的科研人员，采用LDA主题模型提取科研人员的研究主题，识别不同类型科研人员的研究主题的相关特征，对3类科研人员各自主要研究方向在时间序列上的主题强度与主题迁移进行分析。综合上述分析结果，研究工作初步得出以下结论。

（1）高发文科研人员的研究主题相对集中。“科研人员-主题”2-模网络分析的结果显示，高发文科研人员的主题节点数量少于其他两类科研人员，高权重连边数量与单一主题的边权重合计则明显高于其他两类科研人员（见图2）。这一结果表明，在高发文科研人员研究主题相对集中的同时，高被引与随机科研人员的研究主题表现出多样化特征。同时，高发文科研人员对应的2-模网络中突出的边权重，说明高发文科研人员的研究主题在一定程度上代表了整个学科领域的热点研究方向。

（2）高发文科研人员的精专程度更高。针对主题强度的分析表明，高发文科研人员群体的平均主题强度值最高，高被引科研人员群体的平均主题强度值最低（见表5）。这意味着高发文科研人员群体更关注自己的主要研究方向，在其他主题上发文量较少，研究的精专程度较高。而高被引科研人员群体在主要研究方向上的主题强度则在3类人员中最低，在其他主题上的研究工作摊薄了其在主要研究方向上的精专程度。

（3）高被引科研人员研究主题的迁移性更突出。主题变异性的分析结果表明，高发文科研人员在主要研究方向上具备高度稳定性，相反高被引科研人员主要研究方向的稳定性最低（见表6）。相比高发文科研人员，高被引科研人员有更高的概率发生主题迁移，与涉猎广泛或暂时性的兴趣转移不同，高被引科研人员主要研究方向发生主题迁移后很少再被重拾（见图4）。

科学界的传统认知中，在鼓励积极吸纳新知识的同时，也反对盲目追捧流行或时髦的概念，即流行的不等于是高水平的。对于研究中发现的高被引科研人员表现出的主题迁移性，本文并未做更深层次的挖掘。高被引科研人员的主题迁移可能是由于近年来科学知识的快速更新以及高水平科研人员对新知识渴求；也可能由于更多的科研人员热衷于追捧时髦主题，从而堆高其被引量。研究中也存在一些不足之处，以特定领域影响因子排名靠前的权威期刊文献作为研究数据，尚不足以获得更全面的认识，在未来的工作中有待更全面深入的研究。

参考文献

[1] LUHN H P. The automatic creation of literature abstract [J]. IBM Journal of Research and Development, 1958, 2 (2): 159-165.

[2] LOU Y C, LIN H F. Estimate of global research trends and performance in family therapy in Social Science Citation Index [J].

Scientometrics，2011，90（3）：807-823.

[3] ALJABER B，STOKES N，BAILEY J，et al. Document clustering of scientific texts using citation contexts [J]. Information Retrieval，2010，13（2）：101-131.

[4] 谭春辉，麻晓杰. 我国图书馆学非正式学术共同体的形成——基于1998—2012年《中国图书馆学报》的计量分析 [J]. 情报杂志，2014，33（3）：64-71.

[5] 徐健，毛进，叶光辉，等. 基于核心作者研究兴趣相似性网络的社群隶属研究——以国内情报学领域为例 [J]. 图书情报工作，2018，62（12）：57-64.

[6] MANE K，BORNER，K. Mapping topics and topic bursts in PNAS [J]. Proceedings of the National Academy of Sciences，2004，101（l1）：5287-5290.

[7] DEERWESTER S，DUMAIS S T，LANDAUER T K，et al. Indexing by latent semantic analysis [J]. Journal of the American Society for Information Science，1990，41（6）：391-407.

[8] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis [J]. Machine Learning，2001，42（1）：177-196.

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

[10] PHAN X，NGUYEN L，HORIGUCHI S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections [C] //Proceedings of the 17th International Conference on World Wide Web. ACM，2008：91-100.

[11] TITOV I，MCDONALD R. Modeling online reviews with multi grain topic models [C] //Proceedings of the 17th International Conference on World Wide Web. ACM，2008：111-120.

[12] GRIFFITHS T L，STEYVERS M. Finding scientific topics [J]. National Academy of Sciences of the United States of America，2004，101（1）：5228-5235.

[13] WANG X，MCCALLUM A. Topics over time：a non-Markov continuous-time model of topical trends [C] //Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2006：424-433.

[14] 李湘东，张娇，袁满. 基于LDA模型的科技期刊主题演化研究 [J]. 情报杂志，2014，33（7）：115-121.

[15] 廖列法，勒孚刚. 基于LDA模型和分类号的专利技术演化研究 [J]. 现代情报，2017，37（5）：13-18.

[16] WEI T，LI M，WU C，et al. Do scientists trace hot topics? [J]. Scientific Reports，2013，3（29）：2207-2212.

[17] 吴查科，王树义. 基于LDA的国内图书馆学研究主题发现及演化研究 [J]. 新世纪图书馆，2019（7）：90-96.

[18] 衡晓帆，闫佳丽，汪雪锋，等. 基于署名顺序的作者活跃度比较研究 [J]. 情报杂志，2013，32（11）：51-54.

[19] 徐建中，王名扬. 文献影响力的综合评价指标体系研究 [J]. 情报理论与实践，2014，37（5）：56，69-72.

[20] TEH Y W，JORDAN M I，BEAL M J，et al. Hierarchical dirichlet processes [J]. Journal of the American Statistical Association，2006，101（476）：1566-1581.

[21] 曹娟，张勇东，李锦涛，等. 一种基于密度的自适应最优LDA模型选择方法 [J]. 计算机学报，2008，31（10）：1780-1787.

[22] 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.

作者简介

陈立雪，女，1996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。
庹锐，女，1995年生，博士研究生，研究方向：知识组织与信息分析。

Research Topic Focus and Transfer of Scientific Personnel

CHEN LiXue GUO SiYue TENG GuangQing TUO Rui
(School of Information Science and Technology, Northeast Normal University, Changchun 130117, China)

Abstract: The topic preference analysis of scientific researchers helps to understand the dynamic mechanism of the development of scientific knowledge and guide the direction of scientific and technological innovation in domain. This article uses the LDA topic model for topic division, and analyzes the degree of concentration and topic transfer of the main research directions of different type scientific researchers. The research results show that the research topics of high output researchers are relatively centralized, and the degree of concentration in the main research direction is higher, and the transitivity of the research topics of highly cited researchers is more prominent.

Keywords: High Output; High Cited; LDA; Topic Strength; Research Orientation; Topic Transfer

（收稿日期：2019-11-10）