doi:10.3772/j.issn.2095-915x.2017.04.006

基于主题模型的大学学报文献挖掘研究
——以计算机科学领域为例

1. 武汉大学信息管理学院 武汉 430072；2.武汉大学经济与管理学院 武汉 430072；
3. 武汉大学自然科学学报编辑部 武汉 430072

阮剑[1,3] 杨海霞[2] 黄瑱[3]

摘 要 大数据时代下，运用文本挖掘技术自动从海量科技文献中提取研究主题并探测研究趋势十分重要。基于 LDA 主题模型，考虑科技文献的发表时间信息，对优秀“综合性科学技术”类大学学报 2006-2014 期间刊载的计算机科学类文献进行主题内容和主题强度分析；同时基于计算机专业期刊文献，进行研究趋势探测。本文从 25 个研究主题中得到 7 个强度增强的研究主题和 6 个强度减弱的研究主题，揭示大学学报文献中我国高等院校对计算机科学领域的研究状态。通过对数据进行挖掘和分析，了解我国高等院校在计算机科学研究领域的研究趋势，帮助从事该领域研究的学者寻找新兴研究主题，有助于大学学报在定向组稿和学术专辑出版中，把握学术热点与前沿方向，进而提高学报的影响力。

关键词： 大学学报，计算机科学，主题模型，主题分析，趋势分析
中图分类号： G203,G350
开放科学（资源服务）标识码（OSID）

Research on Document Mining of University Journals Based on Topic Model
——An Example of Computer Science

1. School of Information Management, Wuhan University, Wuhan 430072, China;
2. Economics and Management School of Wuhan University, Wuhan 430072, China;
3. Wuhan University Scientific Journal Press, Wuhan 430072, China

RUAN Jian[1,3] YANG HaiXia[2] HUANG Zhen[3]

Abstract In the era of big data, it is very important to use text mining technology to automatically extract research topics from large amounts of scientific and technological literature for detecting the research trends. Considering the publication time information of scientific and technical literature, this study analyzed the topic

作者简介： 阮剑（1968–），武汉大学信息管理学院，研究方向：科技管理、编辑出版，Email: jruan@whu.edu.cn；黄瑱（1974–），通讯作者，武汉大学自然科学学报编辑部，研究方向：图书情报，Email: 00201437@whu.edu.cn。

doi:10.3772/j.issn.2095-915x.2017.04.006

content and topic intensity of computer science journals based on the latent dirichlet allocation (LDA) model from the catalog of “comprehensive science and technology” during the period of 2006-2014. Meanwhile, this research also studied the research trends based on the professional literatures of computer journal. This study obtained seven enhanced research topic and six weakened topic from total 25 topics, which indicated the research status of computer science in the colleges and universities of China. Through the data mining and analysis, this research revealed the development of computer science in the colleges and universities of China, which may assistant the researchers in this field to find out the new research topic, and it is also helpful to the university journal for grasping academic hot spot and front direction soliciting contributions and academic album publishing for improving the influence of the university journal.

Keywords：University journal, computer science, topic model, topic evaluation, trend analysis

1 引言

大学学报是高校优势学科、人才培养、研究成果展示的重要平台和窗口，在相当长的时期，大学学报所刊登文献反映了我国高校科研能力和学科建设能力的水平，特别是一批985、211工程建设高校所主办的综合性学报，展示的学术成果，代表国内相关学科领域的研究趋势。大数据时代下，充分借助自动化的文本挖掘工具，挖掘大学学报刊载的计算机科学类文献的特征，探测其研究主题的内容及主题强度的演化趋势，可以高效地掌握我国高等院校计算机科学领域的研究状态，同时也有利于相关研究工作者把握计算机科学领域的新兴主题。

主题模型算法是文本处理与数据挖掘中一个非常重要的方法，目前已经被广泛地应用于文本分析领域。Blei[1]等2003年提出潜在狄利克雷分布模型LDA(Latent Dirichlet Allocation)，该模型作为生成概率主题模型的成员，假定语料中的文档是由一系列的潜在主题生成，而主题是由一系列的词生成。随后，LDA模型被大量应用于文本分析中，如在线评论分析[2-5]、年报风险披露分析[6]等。Griffiths[6]等借助LDA对PNAS文献建模，并提出用Gibbs抽样算法进行参数估计，其研究成果表明，LDA适合用于文献计量分析。因此，笔者基于LDA模型，对入选《中文核心期刊要目总览（2014年版）》——“综合性科学技术”类前30名中的大学学报刊载的计算机科学类文献进行文本建模，探测国内高等院校中计算机科学研究领域的主题内容及主题强度变化情况。此外，笔者亦对国内优秀计算机类专业期刊中的计算机科学类文献进行了话题分析，话题趋势变动的一致性，表明优秀大学学报中的计算机科学类文献能够代表国内高等院校中计算机科学研究领域的研究状态。

2 相关研究

2.1 对大学学报文献的相关研究

我国高水平大学学术研究的不断深化和创新，促进相关学者愈加关注大学学报的研究成果水平、作者的学历层次以及学术交流的动态。近年，学者对大学学报进行文献计量分析

的相关研究源源不断，主要集中于引文，作者数量，关键词，基金数量，影响因子等。国内与大学学报文献相关的部分研究工作[8-13]如表1所示。从表1可以发现，尽管当前不乏对大学学报文献进行的相关研究工作，但大多都停留在基本统计分析层面上，几乎没有学者对大学学报中计算机科学类文献进行主题内容和主题强度分析。

表1 国内与大学学报文献相关的部分研究工作

作者 (年份)	文献简述
费业昆等 (1992)	文章运用文献计量学的原理，对《中国科技大学学报》七年论文所附引文进行统计分析，定量描述该刊的引文量、类型、年代分布及刊自引量等的分布。
李超等 (2011)	从研究热点、基金项目、著者分布、论文被引以及二次文献转载等多个维度对《中南财经政法大学》893 篇学术论文进行文献计量与统计分析。
邵晓军等 (2011)	以教育部公布的 2009 年 1 月前进入 “211 工程 ” 大学的 100 种自然科学类学报为研究对象，统计各学报的载文量和基金论文比，进行文献分析。
陈留院 (2013)	以论文的不活跃性为视角，对该类论文的作者相关信息、主题内容等进行统计，并分析影响其活跃程度的相关因素。
吕文红等 (2013)	对大学学报（自然科学）类期刊学术影响力进行统计分析，以了解大学学报（自然科学）的整体水平及其同主办学校之间的相关关系。
徐会永 (2014)	以《中国石油大学学报（自然科学版）》为例，通过统计的手段探索科技类高校学报质量与其所依附的高校科研机构科研活跃度之间的关系。

2.2 LDA模型

从文本挖掘的角度看，LDA 模型属于词袋模型，假设词与词之间顺序可交换。模型中的实体形成分层结构，即语料库（corpus）由若干文档（document）组成，文档由若干词（word）组成，其中词是分散的，且为实体中的基本组成部分。在本研究中，一篇文献的标题、摘要、关键字组成一个文档，所有文献的标题、摘要、关键字形成一个语料库。

在LDA 模型中，文档d_i可以看成是潜在主题（topic）的分布，主题是w的分布，且每个词都以一个确定的概率被分配给某个主题，然后词从相关的主题中选取。可观测变量是所有文档中的词w，需要估计的变量有主题$z_{d,n}$、文档-主题分布θ、主题-词分布β，这些变量组成了隐藏在语料库中的主题结构。

为推断可观测文档中的潜在主题结构，需要计算后验分布$p(\beta,\theta,z\mid w)=\frac{p(\beta,\theta,z,w)}{p(w)}$，分子可以通过等式$p(\beta,\theta,z,w)=p(\beta)p(\theta)p(z\mid\theta)p(w\mid\beta,z)$计算，分母表示主题模型中可观测词$w$的概率，计算相对复杂，通常需要借助估计方法进行估计。常用的估计方法有吉布斯抽样（Gibbs sampling）[7]和变分推理（variational inference）[1,14]。笔者追随文献Griffiths[7]等，采用吉布斯抽样算法估计LDA 模型的相关参数。

2.3 LDA在文献计量学中的应用

LDA 模型作为概率主题模型，在分析大型文本语料方面十分有效[1,15]。Griffiths等的研究成果表明LDA 适合于文献计量分析领域，国内外与此相关的代表性研究成果[7,16-31]如表2所示。根据表2，发现即使在国外研究中，也很少有人用主题模型研究大学学报中计算机科学类文献。因

doi:10.3772/j.issn.2095-915x.2017.04.006

此，笔者借助 LDA 模型，挖掘大学学报中计算机科学类文献的主要研究主题，并根据主题强度找出上升趋势和下降趋势的研究主题，准确了解我国高等院校计算机科学领域的研究状态。

表2 国内外 LDA 模型应用于文献分析的相关研究

作者 (年份)	文献简述
Griffiths, Steyvers(2004)	借用 LDA 模型挖取 PNAS 的文章摘要的主题及主题变化趋势，并用马尔科夫蒙泰卡罗算法推断 LDA 模型。
Zheng et al.(2006)	基于 LDA 模型，分析生物医学文献中与蛋白质相关的文献，探测出 300 个主要主题。
Hall et al.(2008)	基于 ACL 选集中的文献，分析计算机语言学领域在 1978-2006 的主题及主题内容演化。
Wu et al.(2010)	基于 24 年的文献，借助 LDA 模型挖掘文献计量领域的研究主题及其研究趋势。
Sugimoto et al.(2011)	借助 LDA 模型分析北美地区图书信息管理学领域的主题变化。
Piepenbrink, Nurmammadov(2015)	借助 LDA 模型分析 600 多种期刊在 1995-2012 刊登的与转型经济或新兴市场相关的文献，定义 25 个主题并分析主题的趋势变化。
王金龙等 (2009)	针对目前科研文献主题演化概率分布问题，阐述了主题与事件的关联关系，提出一种新型的基于模块化的主题方法。
王萍 (2011)	串联文献文本信息和作者信息，构建了基于主题 - 作者 (Topic-Author) 的模型。
叶春蕾等 (2013)	综合科研文献的关键词和引文，构建了一种引文 - 主题概率模型，进行主题识别。
王平 (2014)	考虑文献发表的时间和题录信息，采用 hLDA 模型找到热点话题以及话题的演化特性。
范云满等 (2014)	基于 LDA 与新兴主题特征，探测新兴主题。
李湘东等 (2014)	在 LDA 模型中引入时间因素，以探测科技期刊的主题演化。
祝娜等 (2015)	提出一种基于 LDA 的科技创新主题语义识别方法，利用语义角色标注技术识别科技创新主题。
秦晓慧等 (2015)	提出通过制定主题关联过滤规则 , 对相邻时间窗口间的主题进行关联分析。
王曰芬等 (2016)	以知识流领域的研究为对象，借助 LDA 挖掘不同学科下的知识流研究结构。
杨如意等 (2016)	融合作者和时间两个外部特征，提出一种基于 LDA 的改进主题模型。
关鹏等 (2016)	对不同语料库下的 LDA 主题模型进行对比研究，并对主题抽取效果进行评价。

3 数据与实验

3.1 数据选择

本实验以入选《中文核心期刊要目总览（2014年版）》——“综合性科学技术”类且排名前30的大学学报为目标样本，选取其2006-2014期间刊载的计算机科学类（computer science）文献作为研究对象。研究数据来自web of science 数据库中的中国科学引文数据库，具体检索策略为：出版物名称=清华大学学报(自然科学版)等30个大学学报，时间跨度=2006-2014，研究方向=“Computer science”，文献类型=“Article”，选取字段“英文标题（TI）、英文关键字（DE）、英文摘要（AB）、来源期刊（SO）和发表时间（PY）”，共得 8111 篇论文条目。2006-2014期间，大学学报中计算机科学类文献数量及比例如表3所示。不同样本大学学报的计算机科学类文献的总数量参差不齐，

这是因为不同大学学报办刊风格不同，期刊的周期也不同，有的是双月刊，有的是半月刊。图1是样本大学学报中计算机科学类文献数量在2006-2014年间的变化情况。从图1可知，样本大学学报中计算机科学类文献数量随时间变化是先上升后下降。

表3 大学学报计算机科学类文献主题分析的实验数据来源

学报名称	计算机科学类文献	比例
清华大学学报（自然科学版）	607	7.48%
西安交通大学学报	536	6.61%
浙江大学学报（工学版）	636	7.84%
哈尔滨工业大学学报	425	5.24%
东南大学学报（自然科学版）	335	4.13%
华中科技大学学报（自然科学版）	840	10.36%
上海交通大学学报	383	4.72%
华南理工大学学报（自然科学版）	467	5.76%
东北大学学报（自然科学版）	540	6.66%
四川大学学报（工程科学版）	343	4.23%
吉林大学学报（自然科学版）	382	4.71%
武汉大学学报（自然科学版）	328	4.04%
其他	2289	28.22%
合计	8111	100.00%

注：其他包括北京大学学报(自然科学版)、中南大学学报(自然科学版)、同济大学学报(自然科学版)、中山大学学报(自然科学版)、南京大学学报(自然科学版)、北京科技大学学报、湖南大学学报(自然科学版)、西南交通大学学报、兰州大学学报(自然科学版)、天津大学学报、北京理工大学学报、河海大学学报(自然科学版)、重庆大学学报(自然科学版)、江苏大学学报(自然科学版)、大连理工大学学报、厦门大学学报(自然科学版)。

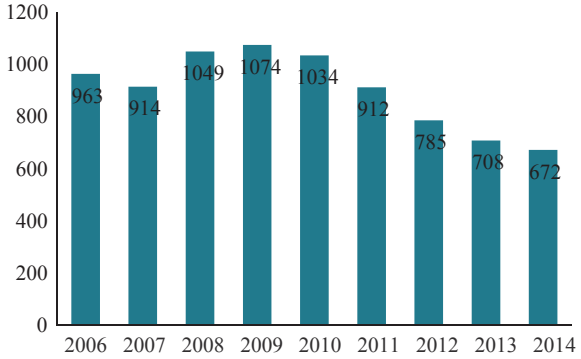

图1 2006–2014期间样本文献年总量变化趋势图

3.2 数据预处理

在运行LDA主题模型前，需要将文献数据处理成文档-词矩阵格式，其中，每行代表一个文档，每列代表一个词。矩阵的条目m_{ij}表示第j个词汇出现在第i个文档中的次数。矩阵的行数等于语料库中的文档数，矩阵的列数等于词汇库中词汇量的大小。本文借用 R 软件中的 tm包对文献数据进行预处理，首先将每篇文献的英文标题、英文关键字和英文摘要分别合并，得到8111个文档；再将文档文本化，得到一个语料库，并依次去除标点符号及数字，以及与主题内容无关的停顿词；最后删除在文档中出现次数低于5的词语，得到一个8111行6413列的文档-词矩阵。

3.3 文献主题挖掘结果与分析

借助R 软件中的topicmodels 程序包[32]对大学学报计算机科学类文献主题进行建模。首先，设定主题数目为10到50，对模型进行测试和训练，并对样本计算困惑度（perplexity），得到最优主题数目的一个较小范围20-30。接着，我们设定5个不同的种子，对整个数据集运行模型，主题数为20 到30，发现主题数为25的模型拥有最低的困惑度。因此将主题数目25作为参数，并设置文档-主题分布的参数为0.01、主题-词分布的参数为0.1、迭代次数为1000、方法为 Gibbs 抽样算法，运行 LDA 模型。得到8111篇计算机科学类文献的主要研究内容，根据模型得到主题-词分布，可以得到不同强度的主题。

主题强度主要描述了主题的热门程度，在某一时刻关于某个主题的文献数量越多，说明

doi:10.3772/j.issn.2095-915x.2017.04.006

该主题的强度越高，可以被认为是热主题；反之，则是冷主题[7]。利用 LDA 模型输出的文档-主题分布，可以按年计算得出计算机科学类文献集中各个主题的主题强度分布情况。本实验中，在95%的置信水平下，25个主题中有6个趋势下降的主题（冷主题），7个趋势上升的主题（热主题），如表4。该趋势状态与文献[33]相似，表明大学学报中的计算机科学类文献能够揭示我国计算机领域的研究状态。

表4 主题强度发生显著变化的主题及其上升与下降趋势

主题强度上升的主题（热主题）		主题强度下降的主题（冷主题）	
主题标签	上升趋势（斜率）	主题标签	下降趋势（斜率）
聚类	0.154***	虚拟技术	-0.157***
特征提取	0.142***	程序设计	-0.135***
视频追踪	0.141***	网络流控制	-0.096***
图像处理	0.061***	Web 服务	-0.093***
重建法	0.055**	分布式计算	-0.081***
入侵探测	0.046***	实时同步	-0.072**
算法	0.040**		

注：***表示99%的置信水平，**表示95%的置信水平

为进一步了解主题，分别找出2006-2014期间三个主题强度上升最为强烈及三个主题强度下降最为强烈的主题，并给出相应主题的主要词。

三个主题强度上升最为强烈的主题的趋势图如图2，与主题相关的前6个词如表5。从图2和表5可知，在样本文献中，主题强度上升最快的三个主题依次是“聚类”、“特征提取”、“视频跟踪”。主题“聚类”与“特征提取”的主题强度上升较快，主要源于大数据时代下，人们越来越注重借助计算机技术对文本进行分析，从而大大减少相关研究者的工作量。

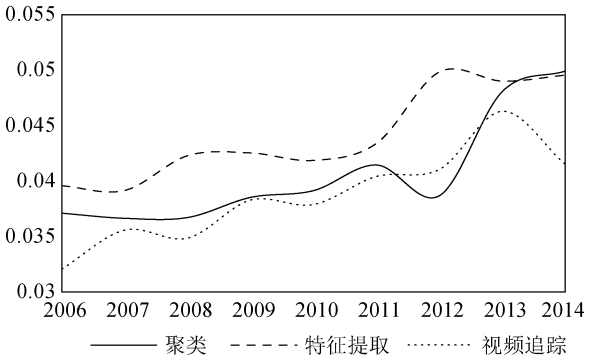

图2 上升趋势最显著的三个主题的主题强度演化趋势

表5 趋势上升最显著的三个主题与趋势下降最显著的三个主题的主要主题词

趋势上升最显著的三个主题			趋势下降最显著的三个主题		
聚类	特征提取	视频追踪	虚拟技术	程序设计	网络流控制
clustering	feature	object	software	design	control
similarity	recognition	video	system	process	flow
based	vector	based	virtual	knowledge	traffic
mining	features	tracking	technology	product	network
retrieval	classification	target	platform	ontology	rate
semantic	SVM	motion	architecture	modeling	delay

2006-2014期间三个主题强度下降最快的主题的趋势图如图3，与主题相关的前6个词如表5。从图3和表5可知，样本文献中，主题强度衰退最快的三个主题依次是“虚拟技术”、“程序设计”、“网络流控制”。正如文章在前文提到的一样，主题强度下降并不表明该主题不受研究者

关注，只是其受关注程度逐渐下降。

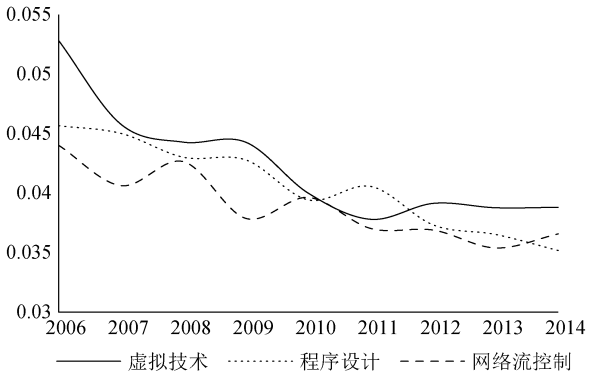

图3 下降趋势最显著的三个主题的主题强度演化趋势

根据表5的主题词分布，可知主题“聚类”所对应的前六个主题词分别为“clustering（聚类）”、“similarity（相似性）”、“based（基于）”、“mining（挖掘）”、“retrieval（检索）”和“semantic（语义）”，显然该主题内部的相关性很高，多与“聚类”相关；类似的，主题“特征提取”的主要词有“feature（特征）”、“recognition（识别）”、“vector（向量）”、“SVM（支持向量机）”、“classification（分类）”，该主题内部的相关性同样很高，主要主题词并未受到其他词干扰。这亦表明LDA 模型对计算机科学领域文献中潜在的主题进行挖掘的方法是有效的。但仔细研究发现，主题中还存在少量与主题无关的的词，如 “based（基于）”在主题“聚类”与“视频追踪”中均占比重较大，主要原因是我国期刊文献习惯使用“基于（based on）”这类词汇，对于这类因写作用词风格而产生的主题词，可以在数据预处理过程中进行去除，从而提高主题模型在文献计量分析中的效果。

为观察不同时期的主题强度差异，可将2006—2014划分为3个时间窗口期：2006—2008、2009—2011与2012—2014。如表6所示，2006—2008期间，主题强度最高的三个主题分别为“虚拟技术”、“程序设计”、“图像处理”；2009—2011期间，主题强度最高的三个主题分别为“互联网”、“图像处理”、“特征提取”；2012—2014期间，主题强度最高的三个主题分别为“特征提取”、“图像处理”、“聚类”。对比不同时间窗口的主题强度，发现主题“图像处理”的主题强度始终较高；主题“特征提取”的主题强度在2009—2014期间相对提高；主题“互联网”在2009—2011期间强度明显高于其他主题。

表6 不同时期的三个高强度主题

2006—2008		2009—2011		2012—2014	
主题	强度	主题	强度	主题	强度
虚拟技术	0.047 7	互联网	0.049 3	特征提取	0.049 5
程序设计	0.044 6	图像处理	0.044 4	图像处理	0.046 6
图像处理	0.043 2	特征提取	0.043 6	聚类	0.042 9

4 结论与展望

本研究在近年国内外研究工作的基础上，借助 LDA 模型，采用Gibbs抽样算法进行参数估计，使用困惑度自动确定文本建模的最优主题数，主要从主题内容和主题强度两方面，探究我国优秀自然科学类大学学报中计算机科学类文献的研究主题在2006—2014年间的演化特点。结论表明：（1）大学学报中的计算机科学类文献能够揭示我国高等院校对计算机领域的研究状态；（2）“聚类”、“特征提取”、“视频追踪”等研究主题越来越受到学术工作者的青睐，“虚拟技术”、“程序设计”、“网络流控制”等研究主题，在学术研究中的比例逐渐下降。

doi:10.3772/j.issn.2095-915x.2017.04.006

该结论既有利于年轻学者了解我国高等院校中计算机科学领域的研究状态，亦能帮助研究生了解其学科领域的新兴主题，确定其研究方向。此外，对于从事大学学报理工类编辑，可以在定向组约稿、遴选出版学术专辑等方面，迅速把握该学科领域的学术热点，有针对性地首发一批高水平论文成果，更好地提升大学学报的影响力。

然而，为了增强本研究的实践意义，未来可以考虑扩展计算机科学文献的主题研究模式，如加入主题生命周期、主题寿命，作者等其他更为复杂的影响主题的因素，从而更加深入地研究大学学报主题的特征演化规律。其次，可以引入结构主题模型[33-36]，对比分析国内计算机科学领域的研究状态与国际计算机科学领域研究状态的异同，而更好地引导我国科研工作者进行相关研究。这亦是笔者下一步的研究工作。

参考文献

[1] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003(3): 993-1022.

[2] Brody S, Elhadad N. An Unsupervised Aspect-Sentiment Model for Online Reviews[C]// Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA. DBLP, 2013: 804-812.

[3] 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.

[4] Lin C, He Y. Joint Sentiment/topic Model for Sentiment Analysis[C]// ACM Conference on Information and Knowledge Management. ACM, 2009: 375-384.

[5] Tirunillai S, Tellis G J. Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data using Latent Dirichlet Allocation[J]. Journal of Marketing Research, 2014, 51(4): 463-479.

[6] Bao Y, Datta A. Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures[J]. Management Science, 2014, 60(6): 1371-1391.

[7] Griffiths T L, Steyvers M. Finding Scientific Topics[J]. Proceedings of the National Academy of Sciences, 2004(101): 5228-5235.

[8] 费业昆, 蒋幼员. 《中国科学技术大学学报》论文的引文统计与分析[J]. 情报科学, 1992(1): 43-48.

[9] 邵晓军, 颜志森. “211工程”大学学报的载文量与基金论文比分析[J]. 编辑学报, 2011(4): 372-374.

[10] 李超, 耿卓. 《中南财经政法大学学报》学术分布与动态计量[J]. 中南财经政法大学学报, 2011(3): 134-140.

[11] 陈留院. 师范大学学报(自然科学版)零被引频次论文特征分析[J]. 中国科技期刊研究, 2013, 24(2): 299-303.

[12] 吕文红, 刘霞, 高丽华. 大学学报(自然科学)类期刊学术影响力统计分析[J]. 中国科技期刊研究, 2013, 24(4): 678-683.

[13] 徐会永. 综合性高校学报质量影响因素分析方法探讨——以《中国石油大学学报(自然科学版)》为例[J]. 中国科技期刊研究, 2014, 25(7): 885-889.

[14] Jordan M I, Ghahramani Z, Jaakkola T S, et al. An Introduction to Variational Methods for Graphical Models[J]. Machine Learning, 1999, 37(2): 183-233.

[15] Blei D, Carin L, Dunson D. Probabilistic Topic Models[J]. IEEE Signal Processing Magazine, 2010, 27(6): 55-65.

[16] Zheng B, Mclean D C, Lu X. Identifying Biological Concepts from a Protein-related Corpus with a Probabilistic Topic Model[J]. Bmc Bioinformatics, 2006,

7(1): 58-58.

[17] Hall D, Jurafsky D, Manning C D. Studying the History of Ideas using Topic Models[C]// Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, Usa, A Meeting of Sigdat, A Special Interest Group of the ACL. DBLP, 2008: 363-371.

[18] Wu H, Wang M, Feng J, et al. Research Topic Evolution in “Bioinformatics”[C]// International Conference on Bioinformatics and Biomedical Engineering. IEEE, 2010: 1-4.

[19] Sugimoto C R, Li D, Russell T G, et al. The Shifting Sands of Disciplinary Development: Analyzing North American Library and Information Science Dissertations using Latent Dirichlet Allocation[J]. Journal of the Association for Information Science & Technology, 2011, 62(1): 185–204.

[20] Piepenbrink A, Nurmammadov E. Topics in the Literature of Transition Economies and Emerging Markets[J]. Scientometrics, 2015, 102(3): 2107-2130.

[21] 王金龙, 徐从富, 耿雪玉. 基于概率图模型的科研文献主题演化研究[J]. 情报学报, 2009, 28(3): 347-355.

[22] 王萍. 基于概率主题模型的文献知识挖掘[J]. 情报学报, 2011, 30(6): 583-590.

[23] 叶春蕾, 冷伏海. 基于引文—主题概率模型的科技文献主题识别方法研究[J]. 情报理论与实践, 2013, 36(9): 100-103.

[24] 王平. 基于层次概率主题模型的科技文献主题发现及演化[J]. 图书情报工作, 2014, 58(22): 70-77.

[25] 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.

[26] 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(7): 115-121.

[27] 祝娜, 王效岳, 杨京, 等. 基于LDA的科技创新主题语义识别研究[J]. 图书情报工作, 2015(14): 126-134.

[28] 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, 31(3): 18-25.

[29] 王曰芬, 傅柱, 陈必坤. 采用LDA主题模型的国内知识流研究结构探讨: 以学科分类主题抽取为视角[J]. 现代图书情报技术, 2016, 32(4): 8-19.

[30] 杨如意, 刘东苏, 李慧. 一种融合外部特征的改进主题模型[J]. 现代图书情报技术, 2016, 32(1): 48-54.

[31] 关鹏, 王曰芬, 傅柱. 不同语料下基于IDA主题模型的科学文献主题抽取效果分析[J]. 图书情报工作, 2016(2): 112-121.

[32] Grun B, Hornik K. Topicmodels: An R Package for Fitting Topic Models[J]. Journal of Statistical Software, 2011, 40(13): 2011.

[33] 杨海霞, 高宝俊, 孙含林. 基于LDA挖掘计算机科学文献的研究主题[J]. 现代图书情报技术, 2016, 32(11): 20-26.

[34] Roberts M, Stewart B, Tingley, et al. The Structural Topic Model and Applied Social Science[J]. Working Paper, 2013, 155(6): 419-20.

[35] Roberts M, Stewart B, Tingley D. stm: R Package for Structural Topic Models[J]. Working Paper, 2014, 57(1): 445-460.

[36] Roberts M E, Stewart B M, Tingley D, et al. Structural Topic Models for Open-Ended Survey Responses[J]. American Journal of Political Science, 2014, 58(4): 1064-1082.