doi:10.3772/j.issn.2095-915x.2020.02.007

LDA 与 BTM 概率主题模型抽取科学主题效果比较研究

开放科学（资源服务）标识码（OSID）

张文伟 赵辉

中国科学技术信息研究所 北京 100038

摘要：分析文献主题是挖掘科学脉络的基础，目前存在多种提取文献主题的方法，被学者广泛使用的方法是使用概率主题模型抽取文献的主题。使用不同的算法和不同的语料提取出的主题结果也不同，本文通过计算查全率、查准率和定性分析方法分别比较利用了 LDA 抽取标题、LDA 抽取摘要、BTM 抽取标题、BTM 抽取摘要的主题效果。本文以纳米材料领域数据为例进行分析，实验结果表明使用摘要做语料提取出的主题颗粒度较小且能够反应文献研究内容的细节，LDA 算法在提取摘要主题方面优于 BTM 算法，BTM 算法在提取标题主题方面优于 LDA 算法。

关键词：LDA；BTM；主题抽取；对比分析

中图分类号：G350.7

Comparative Study on the Effect of LDA and BTM Probabilistic Subject Model in Extracting Scientific Subject

ZHANG Wenwei ZHAO Hui

Institute of science and technology of China, Beijing 100038, China

Abstract: Analyzing the subjects of the literature is the foundation for exploring the scientific context. There are several ways to extract the subjects of the literature, the most common way to extract the subjects of the literature is probabilistic topic models. The results of using different algorithms and different corpora to extract the topic are different. This paper compares the

基金项目：中国科学技术信息研究所创新研究基金 MS2020-02。

作者简介：张文伟（1995-），硕士研究生，研究方向：信息与情报工程，E-mail: zhangww2017@istic.ac.cn；赵辉（1971-），研究馆员，研究方向：信息资源管理、科技资源管理。

subject effects of using LDA and BTM to extract the title and abstract by calculating the recall rate, precision rate, etc. Taking nanomaterials data as an example, the result shows that the topic particle size of abstract corpus extraction is smaller than that of title, which can reflect the specific content of literature research. Compared to the BTM algorithm, the algorithm of LDA is better in extracting an abstract subject. In contrast, the BTM algorithm is prefered than LDA algorithm in extracting the title subject.

Keywords: LDA; BTM; subject extraction; comparative analysis

引言

发现和揭示学科研究发展脉络一直是情报学领域重要研究方向。情报分析学者通常利用文献发掘科学研究主题路线、揭示学科脉络、研究学科热点主题、预测学科发展方向，因此分析文献主题是挖掘科学脉络的基础。

科技文献主题抽取的主要方法有共被引分析、引文耦合分析、LDA（Latent Dirichlet Allocation）算法模型、改进 LDA 模型、TF-IDF 模型、LSA 模型、BTM（Biterm Topic Model for Short Texts）模型算法概率模型、图模型、聚类分析、共词分析等。

语料的选取也影响主题抽取效果，能够反映科技文献研究内容的字段有：标题、摘要、关键词、正文。虽然这几种语料都描述研究内容但是描述的详细程度不同，本文拟探讨使用哪种语料更能反映文献主题。

目前大量情报学领域的专家与学者致力于改进现有的许多算法模型，以提高抽取效果。目前许多论文并没有公开其提出的改进模型算法，再次实现较为困难，本文利用两种公开常用的算法做比较分析，分析 LDA、BTM 算法提取主题的效果，同时也分析用文献哪些数据构建语料库更能准确反映文献主题。

1 相关研究

2009 年陈仕吉 [1] 结合 C-value 与 TF-IDF 算法为 CV-IDF 从 ESI（Essential Science Indicate）数据库选取 1999-2008 年计算机领域高被引文献形成 214 个文献簇，抽取文献集合标题名词词组，利用上述 3 种方法对关键词数量大于 10 的 80 个文献集合选出 5 个最能代表主题词汇，通过对比分析 CV-IDF 能够利用 C-value 与 TF-IDF 优点弥补其缺点。2011 年叶春蕾 [2] 利用基于多词短语词频分析和短语邻近分析的 DT 方法分析美国国家航空航天局战略规划，把 2011-2020 年美国国家航空航天局战略规划的 PDF 文件转换为纯文本 TXT 文件，利用 DT 方法探究其文档主题。2012 年王震 [3] 提出二级词共现方法分析文献主题方法，利用该方法对南京脑科医院 2006-2010 年发表的论文进行了主题分析，实验结果表明二级词共现分析方法能够把表达相同含义的不同表达词结合为词簇，能够避免单纯使用词频统计的缺点。2014 年王庆红 [4] 利用 hLDA（层次概率主题模型）抽取中国知网 2003-2013 年图书馆与情报学学科 57266 篇题录信息的主题，共抽取出 1385 个主题，选出 10 个核心主题进行分析，与 LDA 模型相比，hLDA 抽取质量明显提高。2014 年刘勘 [5] 采集 6 个主题共 800 篇科技类文档共计

doi:10.3772/j.issn.2095-915x.2020.02.007

380021 个词汇，去除出现频数小于 2 的特征词后构建出词条库，利用潜在语义索引（Latent Semantic Indexing，LSI）构造词汇 - 文献矩阵，每篇文档输出 8 个主题词代表文档主题，实验结果表明 LSI 方法比 LDA 方法提取出更多专业词汇，准确率更高。2015 年唐果媛 [6] 从关键词词频、关键词共现词频和两种方法结合的角度分析 62 篇与共词分析法学科主题文献，得出共词分析法分析科学主题演化研究在中国发展比较成熟的结论。2016 年关鹏 [7] 等人利用 LDA 模型抽取关键词、摘要、关键词 + 摘要 3 种形式语料的主题，对 1989-2015 年 3028 篇国内风能领域 CNKI 文献分别抽取上述 3 种不同语料主题，实验结果表明摘要提取主题效果最好。2017 年张思凤 [8] 等人利用 TF-IDF 抽取自然语言处理领域 57 篇高被引文献全文内容和 15 篇施引文献的引用内容的主题，通过实验分析得出引用内容抽取出的候选词代表主题效果较好的结论。2018 年马秀峰 [9] 等人，利用 LDA 模型提取文献主题构建“内容 - 方法”二维网络模型，探究研究内容与研究方法之间的关系，发现学科领域隐含知识。

综上所述文献主题的研究，目前多使用几个高频词表示文献的主题，前期常用的方法主要有利用高频词、TF-IDF、DT 等。随着自然语言处理技术的发展 LDA 技术的广泛应用，提取主题算法主要使用 LDA 算法以及 LDA 改进算法。此外利用主题词共现、主题词作者共现、主题词与引文共现等方法增强主题词强度。语料通常选取纯文本语料，题录信息（标题、摘要、作者）、全文文本、引文段落等。多数论文研究利用不同主题提取算法使提取的主题词更能够代表主题，以及改进相关算法使提取的主题词更能够代表主题。少数论文研究如何利用外部信息增加主题强度，如作者主题模型、主题引文模型、主题机构模型等。关鹏等人研究不同题录语料对主题词提取效果影响。

仅有少数学者探究不同语料与文献主题之间的关系，大部分学者在原有的算法基础上进行改进，对于情报分析人员来说，算法只是手段，不必花费过多的精力研究新算法，要了解不同算法的特性，根据需要选择合适的算法。本文将比较 LDA 与 BTM 两种算法。比较使用标题与摘要做语料提取文献主题的效果，探究在提取文献主题时，该选用哪种算法与语料。

2 模型及评价

2.1 模型介绍

2.1.1 LDA模型

主题模型的起源是潜在语义分析（LSA，Latent Semantic Analysis），该方法通过奇异值分解，将高维文档向量近似地映射到一个低维潜在的语义空间上，以达到降低文档维数和消除词语存在的同义、多义等问题 [10] 的目的。在 LSA 基础上，Hofmann 提出概率潜在语义分析（pLSA，probabilistic LatentSemantic Analysis）模型，这也是第一个完整意义上的概率主题模型，它通过引入概率统计的思想，避免了 SVD 的复杂计算 [11]。Bili 等人在 2003 年提出一个更为完全的概率主题模型 LDA[12]。

LDA 采用词袋模型的方法，它把每一篇语料文档看作由一组主题构成的概率分布，而每个主题是有很多的单词构成的概率分布，从而

形成一个文档 - 主题 - 单词的三层贝叶斯网络模型。假设文档有 T 个主题，每个主题 z 被视为一个词典集合 ω 上的 θ。图 1 为 LDA 的模型示例。

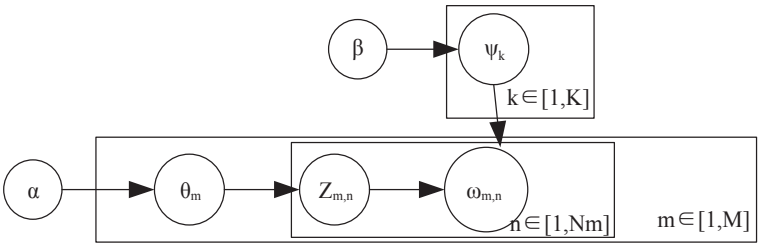

图 **1 LDA** 模型

图 1 中，M 表示文档总数目，K 表示设置提取主题数目，词汇总数为 w，Nm 表示第 m 篇文档包含的词汇数目，ωm,n 表示第 m 篇文档第 n 个单词，Zm,n 表示第 m 篇文档第 n 个主题，ψk 表示主题 k 中包含所有单词的概率分布，θm 表示第 m 篇文档所有主题概率分布。θm 服从超参数 α 的 Dirichlet 先验分布，ψk 服从超参数 β 的 Dirichlet 先验分布。

LDA 生成模型如下：

（1）建立一个语料主题分布 N~Poission (β)

（2）对每一个主题 K∈[1,*K*]，主题分布 ∅ *k*~Dirichlet（β）

（3）从 1 到 N 循环

（4）计算每一个主题 Zmn~Multiontial（θm）

（5）计算每一个单词的 Wmn~ Multiontial（ψzk）

2.1.2 BTM模型

BTM（Biterm Topic Model）词对主题模型由程学旗等人在 2013 提出 [13] 是一种短文本主题提取模型，通过将语料库中出现词共现进行聚合，得到稳定词共现频率进而清楚揭示词之间的相关性，相比于传统主题模型对每个文档中词共现建模，以潜在的方式反映语料库的语意结构，解决短文本数据稀疏问题。

词对（Biterm）是语料预处理之后出现在同一个片段的两个不同的词，短文本可视为一个片段，在长文本片段由 30~60 个词构成。一个包含 3 个不同单词的片段，将构成 3 个不同的词对。例如 W1，W2，W3 可以构成（W1，W2），（W2，W3），（W2，W3）。

一个包含 ND 文档语料，假设语料包含 NB 个词对，词对 $B=\{b_i\}_{i=1}^{N_B}$，bi=(wi,1,wi,2), 主题 K 用唯一单词 W 表示，设 z∈[1,*K*] 作为主题变量指标，K 维多项分布服从 $\theta=\{\theta_k\}_{k=1}^{K}$，$\theta_k$=P（z=k）和 $\Sigma_{k=1}^{K}\theta_k=1$，主题词的分布可以表示为 $K\times W$ 矩阵，且服从 W 维多项分布 $\varnothing_{k,w}=P(w|z=k)$ 和 $\Sigma_{w=1}^{K}\varnothing_{k,w}=1$。我们设 θ 为语料主题分布，$\varnothing_k$ 是某主题下词的概率，α, β 服从 Dirichlet 先验参数，BTM 生成过程如下：

（1）建立语料主题分布 θ~Dirichlet（α）

（2）对每一个主题 K∈[1,*K*]，主题分布 $\varnothing_k$~Dirichlet（β）

（3）对每一个词对 bi∈B，从文本片段主题集合 θ 中抽取 zi，ziMultinational（θ），每一

doi:10.3772/j.issn.2095-915x.2020.02.007

个主题 z 都包含一个词对，wi，1，wi，2~Multinational（$Ø_{z,i}$）

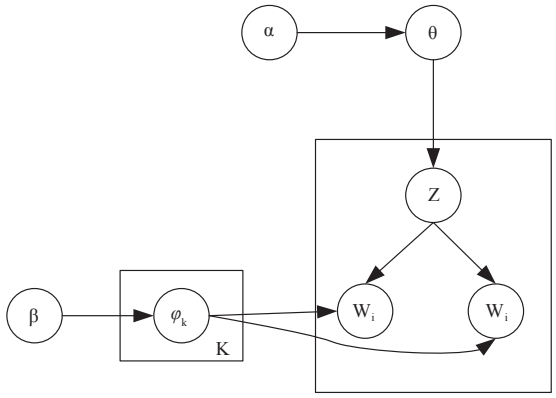

图 2 BTM 模型

BTM 模型生成过程如图 2 所示，假设词对是独立生成，按照以上推论我们可以计算词对 bi 在 θ 和 Ø 的概率：$p(b_i|\theta,\varnothing)=\sum_{k=1}^{K}P(w_{i,1},\omega_{i,2},z_i=k|\theta,\varnothing)$。给定 α，β 可以得到 bi 的在 θ 和 Ø 下的概率。

2.1.3 本文研究模型

本文使用查全率、查准率和 F 值定量分析与定性分析相结合方式探究不同算法模型抽取效果与标题和摘要提取主题效果。

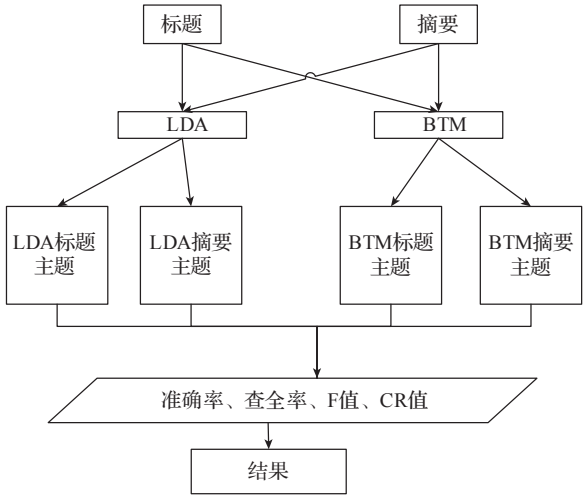

图 3 研究模型

2.2 效果评价

科学文献主题抽取效果评价与算法模型评价不同，算法模型的评价主要从算法的时间复杂度、空间复杂度以及主题提取的准确性出发进行评价。主题抽取效果评价则是把文献分类，进行聚类然后对文献集合分析。本文采用定量方法：查全率、查准率和 F 值来评价不同算法模型抽取效果。

$$R=T_c/T_a \quad (1)$$

公式 1 计算的是查准率 P（precision）：抽取出正确主题（T_c）与抽取出主题（T_a）之比，反映算法的抽取的准确性。

$$R=T_c/T_s \quad (2)$$

公式 2 计算的是查全率 R（recall）：抽取出正确主题（T_c）与标准主题（T_s）之比，反映算法抽取的查全率。

$$F=2PR/(P+R) \quad (3)$$

公式 3 计算 F 值（F-score）：2 倍查准率与查全率的乘积除查准率与查全率之和，为了综合反映算法抽取效果。

$$CR=W_m/W \quad (4)$$

公式 4 计算 CR（correlate）值，主题词之间可表达意义个数 (W_m) 与主题词数量 (W_m) 之比，为了反映主题词之间关联度

查全率、查准率、F 值、CR 值用来评价算法提取的主题与标准主题，此外还需要人工阅读判断不同语料的对比效果。

3 实证分析

3.1 数据与数据预处理

3.1.1 数据获取

本文实验目的在于分析 BTM 与 LDA 提取

主题效果，利用纳米材料领域的文献进行实证分析。2014 年郭全珍等人利用 Citespace 软件分析纳米功能材料领域研究前沿与发展趋势 [14]。Citespace 是文献分析常用基于题录信息可视化计量软件，可以对关键词、作者、引用关系、突发词等进行分析。郭全珍利用 Citespace 分析出纳米功能材料热点研究领域为薄膜材料、纳米管材料、纳米粒子、纳米复合材料等 10 个热点领域。在 WOS 使用 (Nano functional material * or nano function material * or function nano sized material * or functional nanomaterial* or nano-size functional material *) 为检索词检索纳米材料功能，检索时间为 2000-2012 年，文献类型 =（Article）得到 2932 篇文献。

3.1.2 数据预处理

获取数据标题和摘要，利用 NLTK 软件包去停用词、词形还原 [15]，利用 C++ 实现 BTM 算法提取主题 [16]，利用 C++ 实现 LDA 算法提取主题 [17]，分别处理摘要与标题数据。实验环境是一台 Ubuntu 18.04 LTS 操作系统、Intel-Core i5-3230mCPU、2.8GHz、4G 内存笔记本电脑。

3.2 实验及分析

首先从 WOS 数据库下载纳米材料功能领域 2000-2012 年数据，对提取的数据进行预处理（去听用词、提取词干等）提取摘要数据和标题数据，利用 LDA、BTM 算法分别提取摘要数据和标题数据主题。

3.2.1 实验过程

在抽取主题之前需要确定抽取主题的个数，即 K 值。郭全珍等人分析出纳米材料有 10 个热点领域，但是文中分析出 10 个领域与本文所提取主题概念不同，因此不能把主题数定为 10，本文试着把 K 分别设为 10,15,20,30,40,50 试运行，发现 K 值为 10，15, 不能完全体现领域主题，K 值为 50 时有太多重复无效主题，K 值为 20 时无重复主题，K 值为 30 有重复主题，故设 K 值为 30。

分别利用 BTM 和 LDA 算法抽取纳米材料文献集合摘要主题，BTM 算法参数设置如下，alpha=1.66，β=0.03，iter=15000，ntopic=30；LDA 算法参数设置如下 alpha=0.5，β=0.03，niter=15000，ntopics=30，twords=10。得到数据如表 1 和表 2 所示，主题根据主题词归纳得出的主题。

通过阅读 LDA 提取出的主题可以看出 Topic1 研究材料植入人体后与细胞、骨组织的附着力等；Topic5 研究纳米粒子尺寸、直径等物理属性；Topic6 研究密度、能量、电子结构等；Topic7 利用不同的激光照射纳米粒子和对纳米粒子进行光谱分析；Topic10 研究纳米复合材料的强度等机械性能；Topic11 研究蛋白质、DNA、酶等生物方面主题；Topic12 研究纳米材料合成方法与合成纳米材料的结构；Topic13 研究电化学发生氧化还原反应及相关催化剂；Topic17 研究多孔中孔膜表面空隙；Topic18 研究半导体利用太阳能发电，电子如何传输；Topic19 研究纳米纤维作为传感器灵敏度、变化范围；Topic24 研究测试材料的形变观察期机械性能；Topic25 研究等离子沉淀技术生成薄膜材料；Topic26 研究石墨烯、富勒烯、单层石墨、碳纳米管等与碳相关纳米材料；Topic27 研究纳米颗粒与二氧化钛、因相比抗紫外线的性能；Topic28 研究利用电子射线技术衍射、透射进行光谱分析。

doi:10.3772/j.issn.2095-915x.2020.02.007

表 1 BTM 算法提取摘要主题

主题编号	主题词1	主题词2	主题词3	主题词4	主题词5	主题词6	主题词7	主题词8	主题词9	主题词10
Topic1	material	functional	property	new	application	development	research	process	design	technology
Topic2	material	functional	application	nanomaterials	property	new	potential	biological	novel	develop-ment
Topic3	size	function	particle	surface	result	material	effect	structure	energy	temperature
Topic4	functional	structure	material	polymer	assembly	surface	block	molecule	property	interaction
Topic5	film	surface	using	material	thin	substrate	structure	deposition	layer	method
Topic6	nanoparticles	reaction	synthesis	surface	material	functional	particle	method	oxide	metal
Topic7	material	property	mechanical	function	surface	composite	result	strength	temperature	modulus
Topic8	electron	x-ray	microscopy	scanning	using	transmission	diffraction	spectroscopy	characterized	analysis
Topic9	cell	material	surface	protein	bone	tissue	study	result	scaffold	functional
Topic10	density	functional	cluster	energy	electronic	structure	carbon	calculation	atom	theory
Topic11	method	model	function	using	simulation	used	material	based	experimental	result
Topic12	degree	temperature	size	nm	particle	c	phase	sample	powder	material
Topic13	material	structure	molecular	scale	property	functional	system	study	energy	mechanical
Topic14	group	surface	functional	nanoparticles	polymer	silica	using	acid	prepared	used
Topic15	cell	nanoparticles	effect	using	cellular	surface	study	cells.	result	np
Topic16	stress	function	material	wave	effect	elastic	surface	two	field	model
Topic17	electrode	high	electrochemical	material	cell	capacity	performance	carbon	ion	degree
Topic18	optical	spectrum	function	dielectric	frequency	band	peak	show	range	absorption
Topic19	carbon	nanotube	fiber	surface	material	composite	property	single-walled	functional	result
Topic20	adsorption	surface	group	sorption	capacity	pore	ph	functional	area	solution
Topic21	(c)	elsevier	right	reserved.	b.v.	ltd.	31	1.395833	inc.	1.397222
Topic22	protein	dna	peptide	surface	functional	binding	cell	used	using	structure
Topic23	dot	center	phase	=	function	grain	alloy	structure	temperature	material
Topic24	polymer	prepared	polymerization	using	particle	functional	radical	monomer	thermal	group
Topic25	material	high	particle	polymer	adhesive	surface	thermal	using	conductive	property
Topic26	detection	mu	surface	sensor	sensitivity	membrane	limit	film	range	different
Topic27	fabric		showed	cotton	composite	silver	surface	material	contrast	release
Topic28	concentration	surface	ph	aggregation	h	water	presence	different	increase	acid
Topic29	implant	bone	material	used	month	functional	release	also	hydroxyapatite	sinus
Topic30	surgical	surface	n95	significantly	facemasks	functional	different	lower	force	study

表 2 LDA 算法提取摘要主题

主题编号	主题词1	主题词2	主题词3	主题词4	主题词5	主题词6	主题词7	主题词8	主题词9	主题词10
Topic1	cell	tissue	bone	material	study	surface	scaffold	adhesion	implant	result
Topic2	surface	adsorption	solution	ph	group	concentration	interaction	aqueous	effect	study
Topic3	magnetic	cluster	=	new	found	property	transition	show	stability	first
Topic4	group	polymer	hybrid	functional	silica	reaction	prepared	polymeriza-tion	organic	using
Topic5	particle	size	nm	diameter	material	mu	function	bulk	nm.	similar
Topic6	density	energy	electronic	structure	calculation	hydrogen	result	atom	functional	theory
Topic7	scale	energy	nano	also	different	material	length	time	crack	data
Topic8	optical	laser	spectrum	mode	function	light	different	beam	irradiation	also
Topic9	material	application	property	development	new	research	chemical	functional	recent	also
Topic10	composite	property	polymer	matrix	mechanical	nanocomposites	strength	result	also	nanocomposite
Topic11	protein	dna	peptide	binding	functional	molecule	molecular	biological	enzyme	acid
Topic12	synthesis	functional	growth	nanostructures	method	morphology	nanomaterials	nanomaterials.	reaction	formation
Topic13	electrochemical	catalyst	high	oxide	electrode	material	catalytic	reaction	oxygen	reduction
Topic14	temperature	degree	thermal	c	material	increase	sample	dielectric	show	low
Topic15	cell	effect	may	exposure	potential	cellular	toxicity	human	however	study
Topic16	material	method	process	high	technique	using	fabrication	structure	new	developed
Topic17	surface	porous	membrane	pore	area	mesoporous	material	water	silica	high
Topic18	device	nanowires	field	semiconductor	quantum	electronic	electron	transport	solar	work
Topic19	fiber	sensor	used	nanofibers	change	detection	range	sensitivity	based	response
Topic20	model	method	function	simulation	experimental	result	effect	dynamic	material	based
Topic21	(c)	elsevier	right	reserved.	b.v.	ltd.	2011	2010	2012	2009
Topic22	phase	grain	powder	function	crystal	alloy	size	structure	tempera-ture	amorphous
Topic23	metal	ion	structure	complex	center	dot	form	interaction	two	compound
Topic24	mechanical	material	using	function	strain	modulus	deformation	behavior	test	different
Topic25	film	surface	coating	thin	substrate	deposition	layer	plasma	deposited	chemical
Topic26	carbon	nanotube	graphene	functional	cnts	nanomaterials	single-walled	graphite	application	fullerene
Topic27	tio2	silver	photocatalytic	uv	activity	showed	ag	fabric	effective	nanoparticles
Topic28	electron	x-ray	microscopy	scanning	using	diffraction	transmission	spectroscopy	analysis	characterized
Topic29	nanoparticles	drug	nanoparticle	gold	functional	nanomaterials	np	magnetic	release	delivery
Topic30	structure	functional	assembly	block	hybrid	organic	self-assembly	molecular	different	building

doi:10.3772/j.issn.2095-915x.2020.02.007

通过阅读 BTM 提取出的主题可以看出 Topic1 研究开发、设计功能材料的新工艺；Topic3 研究纳米材料尺寸、表面结构、能量等物理特性；Topic5 研究薄膜材料使用的基材使用沉淀制备方法观察薄膜材料结构；Topic6 研究合成纳米颗粒反应；Topic7 研究复合材料机械性能、温度、强度等特性；Topic8 研究利用电子射线技术衍射、透射进行光谱分析；Topic9 研究材蛋白质、细胞、骨组织表面结合；Topic10 研究电子结构、密度、能量能特性；Topic11 对不同型号的材料做模拟实验；Topic12 研究纳米材料尺寸、温度等特性；Topic13 研究纳米结构分子量；Topic14 研究使用二氧化硅等制备纳米颗粒化合物；Topic15 纳米颗粒与细胞之间相互作用研究；Topic16 研究纳米颗粒弹性、应力等机械性能；Topic17 与电池相关研究容量、材料、性能、相关电化学反应等；Topic18 研究吸收关的频率、波段、峰值与光学性能相关；Topic19 研究碳纳米管材料性质、单壁、官度能等；Topic20 研究表面孔的吸附能力；Topic22 与蛋白质、多肽结合结构分析；Topic25 聚合物粘合剂导电性能；Topic26 传感器灵敏度范围研究；Topic27 与棉等面料对比；Topic29 骨种植体使用数月后研究。

分别利用 BTM 和 LDA 算法抽取纳米材料文献集合标题，BTM 算法参数设置如下：alpha=0.05，β=0.03，iter=5000，ntopic=10；LDA 算法参数设置如下：alpha=0.5，β=0.03，niter=5000，ntopics=10，twords=10。得到数据如表 13~ 表 16 所示，根据主题词归纳得出的主题。

通过阅读 LDA 提取标题出的主题可以看出 Topic1 研究纳米材料功能、制造等；Topic4 结构、密度、电子结构等；Topic5 复合材料与细胞组织相关；Topic6 纳米材料与二氧化硅合成；Topic7 研究表面、分子量等；Topic8 利用激光、电子显微镜进行结构分析；Topic9 薄膜材料的沉淀制备；Topic10 研究与碳相关纳米材料纳米管、石墨烯等。

通过阅读 BTM 提取出的主题可以看出 Topic1 研究薄膜材料的性质、沉淀制备方法；Topic2 纳米颗粒合成与二氧化硅、氧化物对比；Topic3 功能材料合成与应用；Topic4 单壁碳纳米管性质研究；Topic5 利用显微镜对表面经行分析；Topic6 纳米材料组装合成；Topic7 研究电子结构、密度性质；Topic8 纳米材料机械磁性、应力机械性能研究；Topic9 复合材料与纳米颗粒相关研究；Topic10 研究光学、分子量。

3.2.2 实验结果分析

从郭全珍研究纳米材料文中得知纳米材料热点领域有 10 个：薄膜材料（thin-films、films）、纳米管（nanotubes）、纳米粒子（nanoparticles）、纳米复合材料（nanocomposities）、功能材料（functional materials）、机械性能（mechanical property）、光学性能（optical property）、尺寸（size）、电子结构（electronic），如表 5 所示。利用表 5 的数据作为标准主题数据。

如表 6 所示，通过阅读 BTM 算法抽取出每个主题下的主题词发现 Topic2、Topic28、Topic30 主题词均不能有效表达主题，其他 27 个主题均能总结出有效主题，通过阅读 LDA 算法抽取出每个主题下的主题词发现 Topic3、Topic4、Topic30、Topic7、Topic15、Topic22 主题词均不能有效表达主题，其余 24 个主题均能总结出有效主题。BTM 算法抽取出表 5 所包含 10 个主题中的 8 个，LDA 抽取出表 5 所包含 10 个主题中的 9 个，查准率分别为 0.8，0.9。

表 3 LDA 算法提取标题主题

主题编号	主题词1	主题词2	主题词3	主题词4	主题词5	主题词6	主题词7	主题词8	主题词9	主题词10
Topic1	functional	fabrication	nanomaterials	polymer	assembly	via	hybrid	protein	self-assembly	nanostructures
Topic2	material	application	nanomaterials	new	based	novel	design	hybrid	system	functional
Topic3	property	effect	method	composite	particle	mechanical	nano	coating	size	prepared
Topic4	study	property	structure	electronic	magnetic	cluster	functional	nanowires	density	structural
Topic5	cell	composite	polymer	oxide	membrane	engineering	formation	mechanism	tissue	containing
Topic6	synthesis	nanoparticles	characterization	silica	high	preparation	application	gold	performance	mesoporous
Topic7	surface	material	molecular	effect	simulation	phase	model	dynamic	modification	modeling
Topic8	using	analysis	process	structure	material	function	study	laser	electron	microscopy
Topic9	film	thin	growth	characterization	silicon	optical	device	crystal	deposition	liquid
Topic10	carbon	nanotube	adsorption	metal	surface	ion	oxide	functionalized	graphene	hydrogen

表 4 BTM 算法提取标题主题

主题编号	主题词1	主题词2	主题词3	主题词4	主题词5	主题词6	主题词7	主题词8	主题词9	主题词10
Topic1	property	film	material	surface	effect	thin	method	coating	characterization	deposition
Topic2	synthesis	nanoparticles	material	application	oxide	silica	mesoporous	characterization	composite	hierarchical
Topic3	material	functional	polymer	application	synthesis	surface	carbon	property	composite	nanomaterials
Topic4	carbon	nanotube	nanoparticles	adsorption	study	cell	single-walled	effect	surface	functionalized
Topic5	material	using	microscopy	analysis	study	force	atomic	surface	function	application
Topic6	functional	nanomaterials	assembly	polymer	material	using	liquid	self-assembly	synthesis	molecule
Topic7	study	property	functional	structure	electronic	density	material	oxide	theory	molecular
Topic8	effect	property	surface	material	magnetic	dynamic	stress	mechanical	behavior	stability
Topic9	composite	nanoparticles	effect	material	silica	particle	study	using	bone	behavior
Topic10	dot	quantum	molecular	synthesis	functional	optical	novel	single	center	structure

doi:10.3772/j.issn.2095-915x.2020.02.007

表 5 纳米材料热点主题

主题	主题	主题	主题	主题
薄膜材料	纳米管	纳米粒子	纳米复合材料	功能材料
机械性能	光学性能	尺寸	纳米结构	电子结构

表 6 摘要评分表

	抽取主题数	抽取正确主题数	标准主题数	有效主题个数	查全率P	查准率R	CR
BTM	30	8	10	27	0.266667	0.8	0.411429
LDA	30	9	10	24	0.3	0.9	0.436364

由表 3、表 4 看出 BTM 与 LDA 算法抽取标题主题与摘要相比质量差很多，BTM 算法与 LDA 算法相比效果更较好一些，利用 BTM 算法抽取的 10 个主题均能分辨，LDA 抽取的 10 主题只能分辨 8 个。利用 BTM 算法抽取出的 10 个主题与标准主题完全相符，利用 LDA 抽取出的主题与标准主题相符的只有 8 个。

由表 7 可知利用摘要做语料库时，LDA 提取有效主题数量不如 BTM 提取出的有效数量多查准率低于 BTM，但是 LDA 提取出的标准主题数量比 BTM 算法多查全率大于 BTM，由表 7 可知利用标题做语料库时 BTM 提取主题查全率与查准率都高于 LDA。由表 1 与表 2 对比可得出利用摘要数据做语料比标题做语料主题颗粒度更细，更能反映主题研究内容。

表 7 标题评分表

	抽取主题数	抽取正确主题数	标准主题数	查全率P	查准率R	CR
BTM	10	10	10	1	1	1
LDA	10	8	10	0.8	0.8	0.8

3.3 实验展望

本文的研究发现标题与摘要都能够反映文献主题，利用标题提取的主题颗粒度较大，能够突出反映研究方法与研究问题，利用摘要提取主题，能够反应文献使用的具体方法，实验过程等信息。结合上述语料的特点本研究将在后续的工作中结合标题数据与摘要的特点分析文献。利用 LDA 算法抽取文献摘要主题，利用 BTM 抽取标题主题。

如图 4 所示，多数据融合主题能够有层次地表现文献主题，分别利用文献集合与标题集合标题抽取主题，由于标题集合抽取的主题颗粒度大，能够较好地反映文献运用的方法与解决的问题，摘要集合提取出的主题则能够详细反应文献的研究内容，利用主题文档矩阵，对应同一文献下的摘要主题与标题主题，使得文献主题富有层次。

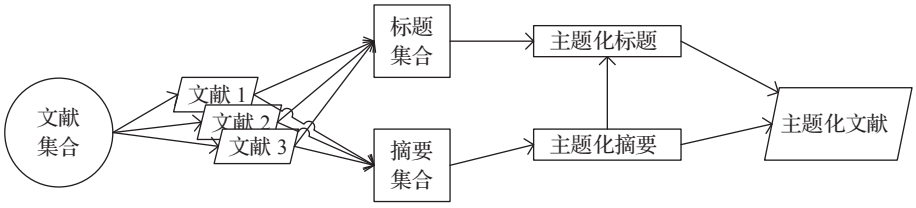

图 **4** 多数据主题融合模型

4 结论

通过上述实验分析可以发现，利用摘要数据能够提取出较多的文献信息，利用标题数据提取出的主题更能解释研究方法与问题；算法方面：LDA 算法提取摘要数据的效果优于 BTM，利用 LDA 提取出的主题不如 BTM 算法多，但是 LDA 提取主题准确度比 BTM 高；BTM 算法提取标题数据效果优于 LDA。

摘要文本数据量大，利用 LDA 提取文献集合将消耗大量的计算资源与时间，与标题相比，摘要所包含的信息比标题丰富，能够较精准地提取文献主题。大规模文献集合可使用标题做语料，能够节省大量时间与计算成本，也能在一定程度反应文献集合研究方法与研究目的。

参考文献

[1] 陈仕吉 , 王小梅 . 基于 C-value 与 TF-IDF 的文献簇主题识别研究 [J]. 情报学报 , 2009, 28(6):821-826.

[2] 叶春蕾 , 冷伏海 . 科技文献全文主题识别方法实证研究 [J]. 现代图书情报技术 , 2012, 28(1):53-57.

[3] 王震 . 基于同类高频关键词的二级词共现方法在文献主题分析中的应用 [J]. 中华医学图书情报杂志 , 2012, 21(1):60-66.

[4] 王庆红 , 王平 . 基于 hLDA 的科技文献主题摘要生成算法与实现——以电力行业论文为例 [J]. 图书情报知识 , 2014(4):63-67.

[5] 刘勘 , 朱芳芳 . 基于潜在语义索引的科技文献主题挖掘 [J]. 计算机工程与应用 , 2014, 50(24):113-117，150.

[6] 唐果媛 , 张薇 . 基于共词分析法的学科主题演化研究进展与分析 [J]. 图书情报工作 , 2015, 59(5):128-136.

[7] 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.

[8] 张思凤 , 梁梦丽 , 曹高辉 . 基于引文的科技文献主题抽取研究 [J]. 情报理论与实践 , 2017, 40(6):122-127.

[9] 马秀峰 , 郭顺利 , 宋凯 . 基于 LDA 主题模型的“内容 - 方法”共现分析研究——以情报学领域为例 [J]. 情报科学 , 2018, 36(4):69-74.

[10] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6):391-407.

[11] Hoffman T. Probabilistic latent semantic analysis[J]. Uncertainty in Artificial Intelligence, 1999, 15(6):289-296.

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

[13] Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts[C]. WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web. 2013:1445-1456.

[14] 郭全珍 , 吕建国 . 纳米功能材料领域研究前沿和发展趋势的可视化分析 [J]. 情报杂志 , 2014, (3):49-53.

[15] NLTK 3.4 documentation[EB/OL].[2019-03-27]. http://www.nltk.org/.

[16] Code for Biterm Topic Model (published in WWW 2013) [EB/OL]. [2019-04-21]. https://github.com/xiaohuiyan/BTM.

[17] GibbsLDA++:A C/C++ Implementation of Latent Dirichlet Allocation[EB/OL]. [2019-05-31]. http://gibbslda.sourceforge.net/.