利用 LDA 的领域新兴主题探测技术综述*

范云满[12] 马建霞[1]

[1]（中国科学院国家科学图书馆兰州分馆 兰州 730000）

[2]（中国科学院研究生院 北京 100049）

[摘要]本文以 LDA 为基础，系统地梳理了新兴主题探测以及主题趋势探测技术中的 LDA 以及其他 LDA 改进主题模型(topic model)的发展现状。介绍了 LDA 的变分推导和 Gibbs 抽样两种参数推导算法；梳理了近年来对 LDA 模型的改进，包括对主题演化建模的主题模型、对文档内容和元数据联合建模的模型、采用在线式学习的主题模型及将 LDA 和引文分析相结合的主题演化方法等，并对不同的改进模型进行了深入对比和分析；梳理了 NIH-VB, TIARA,VxInsight 等几种主要的主题模型可视化技术。最后通过对 LDA 模型的总结分析，探讨了利用 LDA 模型探测领域新兴主题时的关键研究问题。

[关键词] 主题模型 LDA(Latent Dirichlet Allocation) 引文分析 主题模型可视化

[分类号] TP393

Review for the Techniques Detection using LDA for the Field Emerging topic

Fan Yunman[12] Ma Jianxia[1]

[1](The Lanzhou Branch of National Science Library, Chinese Academy of Sciences, Lanzhou 730000, China)

[2]（Graduate University of Chinese Academy of Sciences,Beijing 100049,China）

[Abstract] LDA Based，this paper reviews the development of the LDA model and several models which improve the LDA for the filed emerging topic detection.The paper describes two parameter inference algorithms of variational derivation and Gibbs sampling;and reviews the improvement to the LDA in recent years,including the one modeling the evolution of the topics,the one modeling jointly with the content of the document and the meta data,the one with online learning ,the the topic evolution method combines LDA and citation analysis and so on;then compares and analyses the different kinds of improvement models in details; then reviews NIH-VB,TIARA,VxInsight etc of several main visualization techniques. Finally,

*本文系中国科学院西部之光联合学者项目“基于计算情报方法的甘肃省战略新兴产业技术创新竞争与发展研究”（项目编号:Y200201001）的研究成果之一。

according to the compare and analysis to the LDA, author discusses the key research problems of detecting the emerging topic by using LDA.

[Keywords] Topic Model LDA(Latent Dirichlet Allocation) Citation Analysis Topical Visualization

1 引言

作为科研人员，要追踪所在学科的最新发展动向，就要追踪该领域的最新会议文献、最新发表的论文及专利文献等，分析这些文献中发表的新观点，用到的新技术，得到最新的研究热点。但是，文献资源电子化的增长速度日新月异，某一领域的研究人员要从浩瀚的文献资源中发现自己领域的科学前沿，非常困难。而且新兴技术的出现往往是几个学科相互交叉的结果，而从一个学科去发现别的学科中出现的新技术和新主题，更加困难。利用海量文献中的文本挖掘技术是帮助科研人员快速发现新兴主题的途径之一，而主题模型作为一套新的能够对文献资源进行语义抽取的算法[1]，提供了一种解决问题的新方法。主题模型现在已经成为国际上的研究的热点，并且越来越引起研究人员的注意。作者在 WOS（Web of Science）数据库以检索词“topic model”，检索途径为“主题”，时间范围为 2002-01-01 到 2012-12-01，共检索到 276 条数据，结果如图 1 所示。

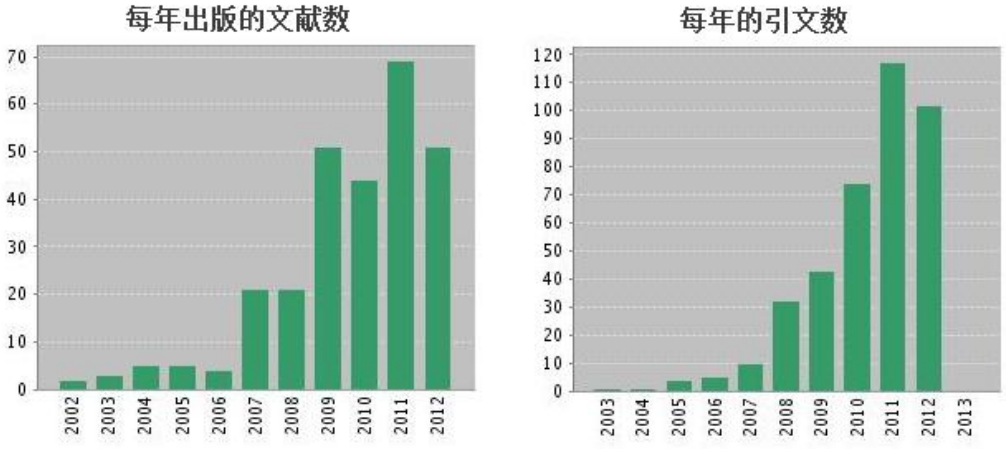

图 1 WOS 数据库中以“topic model”对“主题”进行检索，每年的论文数量及每年的论文被引次数(2002-01-01 到 2012-12-01)[①]。

虽然主题模型能够在语义上对文档进行表示，但是对于那些从事于特定领域的分析人员来说，仍然需要一些主题模型的知识，这无疑加大了对主题模型的学习曲线。而如果对主题模型产生的主题进行可视化的展示，能帮助研究人员快速、直观地了解领域的研究进展。

另外，主题模型对于文档的语义表示，是一种概率化的单词抽取，存在着不确定性。如果能够将引文关系和主题模型相结合，是一种对主题模型存在不确定性这一不足的改进。因此，本文也对这个方向进行了综述。

本文以最简单的主题模型 LDA(Latent Dirichlet Allocation)为例，介绍了主题模型中的一些基本概念，然后总结了几种对 LDA 模型的主要的扩充算法，并分析了几种主要的可视化技术。接着对将引文关系引入主题模型这一方向进行了分析。最后，本文对于主题模型存在的问题进行了分析，并对未来的发展方向进行了展望。

[①] (资料来源：http://apps.webofknowledge.com,2012-12-11）

2 LDA 模型及其算法改进

2.1 LDA 模型简介

主题模型是文本降维技术的一种。最常用的文本降维技术是词频-逆文档频率(Term Frequency-Inverse Documen Frequency,TF-IDF)，将一个文档集单词和文档的矩阵，但是 TF-IDF 形成的矩阵往往非常稀疏，并且不是以语义的形式表示文档，只是按照文档的词法构成对文档的表示。后来 Nigam 等人从语义上表示一篇文档，提出了一元混合模型(Mixture of Unigrams)[2]，认为一篇文档谈论一个主题，文章中的单词表示这个主题；Hofmann 等人提出 pLSI (Probabilistic LSI) [3]，认为一篇文档由条件依赖于文档的多个主题组成，表示主题的单词服从于主题的多项式分布；David 等人提出 LDA(Latent Dirichlet Allocation) [4]，认为文档是由服从多项式分布的主题组成，每个主题由服从于主题的多项式分布的单词组成。

LDA 是文档、主题和单词组成的三层贝叶斯产生式模型[4]。一篇文档的产生过程如表 1 所示。

表 1 LDA 产生文档的过程

LDA 产生文档的过程
Step1. **抽取**文档的长度 $N \sim Poission(\beta)$；
Step 2. **抽取**文档在主题上的分布 $\theta_m \sim Dir(\alpha)$；
Step 3. **for** n=1 to N
Step 4. (a) **抽取**一个主题 $z_{mn} \sim Multinomial(\theta_m)$ ；
Step 5. (b) **抽取**单词 $w_{mn} \sim Multinomial(\phi_{z_{mn}})$ 。

其中 α 和 β 是 LDA 的先验参数，θ 和 φ 是两个需要估计的参数。对于参数近似估计算法主要有两类，一种是变分推断算法，包括均值场变分推断(mean field variational inference)[5],崩溃性变量推断(collapsed variational inference)[6];一种是利用马尔科夫链蒙特卡洛方法(Markov Chain Monte Caro)的 Gibbs 抽样[1, 7]。两种算法的对比分析见表 。

表 2 LDA 参数的两种近似估计算法的对比分析

	估计模型参数的方法	优点	缺点
变分推断方法	利用 EM 算法迭代计算贝叶斯后验分布的概率[8-10]	速度快	模型参数不够精确
Gibbs 抽样算法	利用马尔科夫链蒙特卡洛方法	模型参数精确	终止条件不明；收敛比较慢

2.2 LDA 模型存在的问题

LDA 模型在建立时，提出了六个假设：一，文档之间可交换；二，主题之间可交换；三，词袋模型假设[11][12]；四，文档是多个主题经过有限次混合而成；五，文档是主题上的多项式分布；六，文档集中主题的个数 K 是已知并且固定的(可以使用无参贝叶斯的方法去掉这个假设[13-15])。

其中第一个假设认为文档之间是没有先后顺序。而在实际中，文档中主题的发生往

往有顺序，比如在“天宫一号”发射之前，网上新闻集中于讨论“天宫一号”准备的情况，而发射之后，“天宫一号”的运行情况是关注的焦点。显然第一个假设存在没有对主题随时间的演化关系建模的不足

第二个假设认为主题之间是没有层次关系和先后关系的。而在实际中，在一个关于体育的语料库中，讨论的主题有，篮球，足球，这些主题之间是存在层次关系。显然，第二个假设存在没有对主题之间的层次关系建模的不足。因此，后来的研究人员提出了对 LDA 模型的各种改进。

2.3 改进的 LDA 模型算法

(1)对主题演化过程建模的 LDA

这类模型将和时间信息结合到主题模型中，得到的主题表现出随时间轴演化的趋势。最具代表性的是DTM(Dynamic Topic Model)[16]。DTM将时间离散化，然后按照时间片将语料库分割，形成按照时间序列的文本语料片(slices)，在每个片内主题是可交换的，对每个时间片内的文本按照静态主题模型建模。同时，每个时间段的主题模型的参数依赖于前一个时间片段的参数。

DTM的优点在于能够利用时间序列对整个语料库进行建模，从而揭示一个语料库中主题随时间的演化规律。但是，DTM没有考虑每个时间片段内文档数目对主题数目的影响，也没有对时间片段间主题的动态关系进行建模。同时，DTM存在如何寻找最优的时间切片方式的问题。

cDTM[17] (continuous time Dynamic Topic Model)利用布朗运动(Brownian motion)对主题的演化建模，从而将 DTM 中时间细粒度的选择这一关系到计算复杂度的问题转化为一个模型选择的问题。

TOT 模型(Topic Over Time Model)[18]和 DTM 不同的是,不利用马尔科夫性将时间离散化，而是认为每个主题与某一个时间戳的连续分布有关，对共现词和文档的时间戳共同建模。

（2）对主题间层次关系建模的 LDA

层次 LDA 主题模型(Hierarchical LDA,HLDA)能够对主题之间的层次关系进行建模。HLDA 不需要预先指定主题的数目，在生成层次主题结构的同时自动确定主题数目。HLDA 的缺点在于不能拥有同一层次的两个主题。

CTM(Correlated Topic Model)的生成过程和LDA的生成一篇文档的过程非常类似，只是主题分布从一个逻辑斯蒂分布中抽取。它能够根据一个主题推测出另外的主题；但是，由于逻辑斯蒂分布和多项式分布不是共轭分布，因而在后验分布的计算上，开销比较大[11]。

PAM分布(Pachinko allocation)[13]利用DAG (Directed Acyclic Graph)构建一棵层次树。树的叶子节点对应着单词，而内部节点对应着主题。PAM的主题既可以建立在单词之上，也可以建立在主题之上，从而能够对主题的层次关系建模；同时，PAM中每个节点是其子节点的某一任意分布，因此，可以获取任意的、嵌套的、非结构的数据的主题以及主题之间的关系。

(3)词序有关的 LDA

文档中单词的顺序是包含了对于理解文档的主题含义有用的信息，下面给出几种这方面有关的模型。

BGTM(Bigram Topic Model)[19]认为，当前一个单词的主题概率的分布当且仅当依赖于前一个单词的主题概率分布，每个单词的生成都是一个独立的LDA。和LDA的不同之处在于，第一、对于每个单词都抽取一个主题分布；第二、一个单词的生成的概率条件依赖于上一个单词的主题分布。

LDACOL (LDA Collocation Model)在BGTM模型的基础上，引入了一组随机变量 $\mathbf{x}$ 来标识当前一个单词和前面一个单词是否形成一个词组[20]。一个单词的生成不仅仅受到主题分布的影响，也受到一个随机的贝努利分布的影响，这个分布决定当前词和前一个词是否形成一个词组。

TNG (Topical N-gram) 模型[20, 21]将LDACOL模型进一步推广，认为主题不仅仅是一个个单词组成的主题，而是更具有理解性和解释性的词组。G. S.; Mimno等利用TNG将主题模型和文献的共引分析相结合，提出了一套主题模型的文献计量学指标[22] 。

可以看出，BGTM和LDACOL模型都是TNG模型的一种特殊情形，当令所有的 x_i=1，即BGTM模型；当

σ只依赖于前一个单词时，即得到LDACOL模型。

(4)对文档主题和元数据联合建模

主题模型本质上只是从词的共现对文档进行语义分析，没有利用文本自身具有的元数据。最常见的是和文档的作者联合建模的ATM和TAM。

ATM (Author-Topic model)，将文档的作者元数据引入到模型中，对文档的内容和作者联合建模，每个作者是主题的多项式分布，每个主题是单词的多项式分布[23]。ATM通过计算作者在每篇文档的内容上的熵，可以用来预测一位作者倾向于使用哪些主题，或者一位作者对哪些主题感兴趣。

王萍提出了TAM (Topic-Author Model)[24]。TAM可以用于研究领域的专家发现、文献标注、重要文献的发现、文献相似度分析与文献推荐以及研究趋势分析[24]。

TAM和ATM相同之处在于都是对文档的内容和作者联合建模，不同之处在于ATM是利用作者-主题的分布生成文档中的每个单词，而TAM是首先生成文档的每个单词，然后从文档中随机选择一个单词，根据该单词所对应的隐含主题生成作者。

除了结合元数据作者外，也有研究者结合了其他的元数据对主题模型建模[25, 26]。

(5)基于引文分析的主题模型

LDA是一种概率式的产生式主题模型，表示主题的单词的生成具有不确定性，研究人员探索是否能将文档之间的引文数量这一定量的确定性的指标引入到主题模型中，改进对主题演化的预测。同时，引文分析本身具有一些缺陷：

一、时滞性。最新的研究进展一般发表在会议论文而不是期刊中[27]，而由于SCI中的引文关系需要经过专家评审才能入库，因此会议论文被收录到SCI中需要更长的时间。

二、引文收集的不完备性。Goodrum等将同一个关键词分别在Citeseer和ISI[28]上检索，无论在引文的数量还是引文的来源上都存在较大的差异[29]。

三、引文分析方法的受限制性。对于很多的数据库中，比如中国的专利数据库[30]中不存在引文数据。

Dietz等利用web中的pagerank思想，认为相互引用的文献之间，被引文献的主题分布对施引文献的主题分布产生影响[31, 32]。Nallapati等提出将施引文献和被引文献构成一个文献对，将文献对放在主题模型中联合建模，主题之间的影响是服从于一个超参数决定的概率分布[33]。这两种模型将主题的生成过程引入了参考文献，但是没有将引文的数量这一定量指标引入到对主题演化的预测中。

Mccallum等提出了结合引文分析和主题模型的文献计量学评价指标体系，比如主题影响因子(Topic Impact Factor,TIF) [22]。对于某个给定主题T，在某个时间段t内，其主题影响因子的定义为：

$$TIF(k,t)=\frac{count(citations\ from\ D_k^{\ t}\ to\ D_k^{\ t-2})}{|D_k^{\ t-1}|+|D_k^{\ t-2}|}$$

其中$D_k^{\ t}$表示在时间段 t 中，包含主题 k 的文档的数目。但该研究没能针对具体领域进行实验和验证。

在实际应用方面，贺亮等利用 LDA 话题模型抽取科技文献的话题[34]，然后计算话题的强度和影响力，最后针对热门和冷门话题以及影响力高和影响力低的话题，进行了趋势分析。但这种方法没有考虑时间维度和业内领军人物的因素等。

(6) 在线式主题模型

LDA 是离线式地对文进行分析，当要处理的文本是文本流时，不可能取得所有的文本进行建模。Alsumait L 等提出 OLDA (Online Topic Model) [35]，将 LDA 主题模型用于文本流，识别出文本中的主题以及主题的演化规律。首先根据已有的文本数据构建一个主题模型，然后对于新到的一篇文档，增量式地构建出一个实时更新模型，这个模型同样也是每篇文档的主题的混合和每个主题的单

词的混合。这种方法的优点在于，通过新的数据流中的信息来增量式地更新当前的模型，而不需要访问以前的数据。同时这种方法也可以用于跟踪随时间变化的主题，以及实时探测新的主题。

Hoffman [36]引入了在线式的变分推断方法，利用一种随机的自然梯度算法来判断收敛到目标函数的程度，通过实验表明，该方法比传统的变分推断算法收敛速度，以及对参数的估计的效率和准确度(以F值为测度)都要好。

Banerjee, A [37]等提出了LDA模型的在线式版本，并提出将在线式推断和离线式推断相结合的模型。该方法通过在线式的学习文本流的主题内容，然后以离线的方式更新整个语料库的主题内容，从而在离线式方法的性能和在线式方法的高效之间取得了平衡。

(7)小结

当前对LDA的改进主要有6类，即对主题演化过程建模、对层次关系建模等6类，这些改进中1-3是对LDA所做的六个假设的进行的扩充，每个模型从一个或两个假设的角度进行了扩充。而对文档内容和元数据联合建模在文档主题基础上引入了与文献内容密切相关的外部信息，而结合引文分析的方法则更进一步，文献的引文关系本身就能反映主题的演化关系，将它与LDA结合起来，将更进一步发挥二者所长，结合引文分析的方法在未来基于科技文献的主题演化分析，或者新兴主题发现方面有着较大的应用前景,在线式主题模型因为可以通过新的数据流中的信息来增量式地更新当前的模型，而不需要访问以前的数据可以用于跟踪随时间变化的主题，以及实时探测新的主题，相信会成为未来对网上动态信息分析发现主题演化的研究热点。

表 给出了这些模型的简单地对比分析。

表 3　各种扩展模型的对比分析*

模型类别	模型	扩充假设	适应领域	模型复杂度	计算开销
对主题演化过程进行建模	DTM	一	social media	一般	一般
	cDTM	一	social media	复杂	大
	TOT	一，二	social media	一般	一般
对主题间关系建模	HLDA	二，六	News	一般	一般
	CTM	二，五，六	News	一般	大
	PAM 分布	二，六	News	复杂	大
词序有关	BGTM	三	文本处理	复杂	大
	LDACOL	三	文本处理	复杂	大
	TNG	三	文本处理	复杂	大
结合元数据	ATM/TAM	结合作者等元数据	文献情报	一般	一般
在线式 LDA	OLDA	对文本流建模	社交网络	复杂	大

*其中结合引文分析目前只有利用引文数量来评价产生的主题的方法，没有相应的主题模型。

3　LDA 模型的主题可视化

产生的主题如果可以一种直观的、图形化的、可交互的方式呈现出来，能够避免研究者学习复杂的主题模型，有助于快速理解文档的主题，直观地观察主题之间的关系和发展趋势。主题模型的可视化技术最具代表性分为下面几种：

3.1 二维节点-连接图

这类图以二维的点线图表示主题以及主题之间的相似关系。对于文档的可视化分为三步， ⅰ.将文档(主要是摘要)通过主题建模，得到一组主题，进行降维同时保持了文档的语义不丢失。ⅱ.通过 KL 散度计算文档(ⅰ中得到的主题)两两之间的相似度；ⅲ采用 Drl(Distributed Recursive Graph Layout)算法将成对的文档以及ⅱ中得到的相似

度进行可视化。其中最典型的是 NIH-VB(NIH Visual Browser)，它是一个面向 NIH 的，对申请基金的项目的文档进行可视化的浏览器[38, 39] 。如图 2 所示，可以看出，NIH-VB 提供了用户交互的功能，比如查询、放大、缩小以及导出数据的功能。利用 NIH-VB 可以直观地把握某一个时间段内 NIH 资助项目的主题分布，但是不能展示主题随时间的演化。

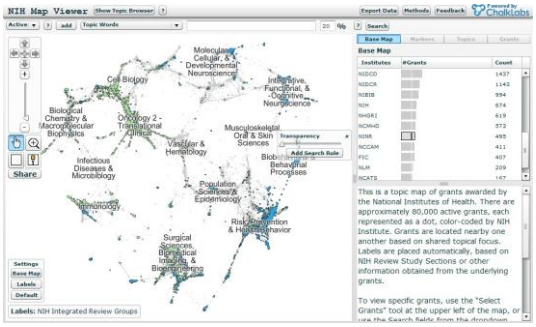

图 2 NIH-VB 对 NIH 数据可视化展示②

3.2 二维河流式图

河流式图是一种比喻，将主题的强度表示成河流的宽度，其随时间的演化构成一条河流。比如 TIARA，几个主题放在一个图中，横轴表示时间，纵轴表示每个主题在每个时间点上出现的强度，从而便于发现在某一个时间点主题的对比关系[40] 。如图 3 所示，图中每一层表示一个主题，层中的多个 label 是从文档中抽取的具有时间代表性的关键词。虽然对关键词的提取可以采用 LDA，但是如何评价这些关键词，以及决定时间片最优的粒度是这种技术的问题。

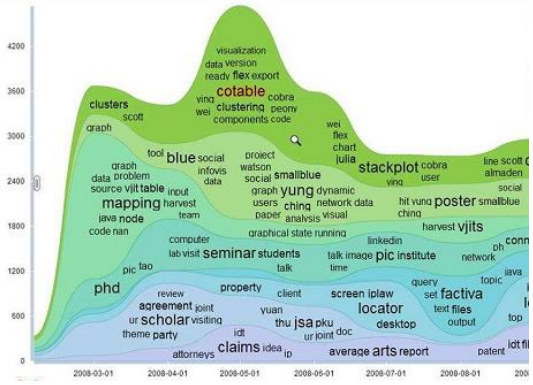

图 3 TIARA 的图形表示③。

3.3 三维地势图

该技术将文档进行主题建模，然后采用 vxord 算法计算出每个文档的位置，然后将相类似的文献聚在一起，形成一个山峰，如果在短时间内出现大量的文献，那么该山峰非常陡峭；不同类别的文献之间的联系形成山脉。比如 VxInsight，广泛用于文本挖掘，专利发现，领域知识管理[41]。如图 4 所示，绿色的表示期刊文献，橙色的表示会议论文，蓝色的表示政府报告。每座山峰在不引起混扰的情况下，用主题的前 n 个单词表示。VxInsight 的优势在于能够很清晰地发现某个时间点上核心期刊，核心文献是什么，哪一个领域是最热门的；不足之处在于要观察某一个领域在某个时间阶段的发展趋势，需要形成多份视图。

② (资料来源：https://app.nihmaps.org/nih/browser/,2012-12-11）
③ (资料来源：参考文献[40]）

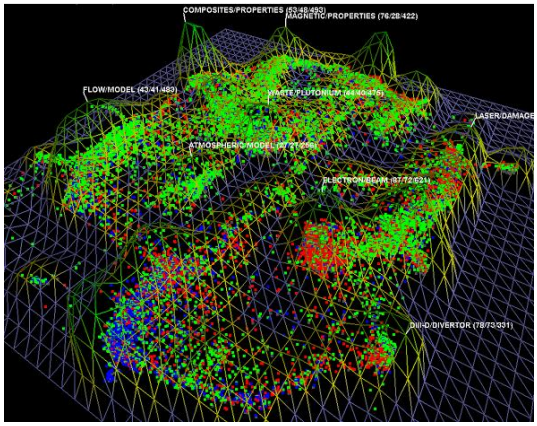

图 4 VxInsight 给出的由三种期刊给出的文档集中的主题分布图④。

(4) 小结

表 4 各种可视化技术的对比分析

	NIH-VB	TIARA	VxInsight
支持数据来源	单	单	单
动/静态	静态	动态	动态
Framework	topics,Drl	-	topics,Drl
处理方式	offline	Offline	offline
开源	否	否	否
图形式	node-link	Riveflow	地势图
趋势分析	否	是	否

通过上述比较，可以看到以 NIH-VB 为代表的二维节点连接图的方式，可以直观地把握某一个时间段内 NIH 资助项目的主题分布，但是不能展示主题随时间的演化；二维河流图的可视化方式将主题的强度表示成河流的宽度，并以河流方式展示了主题随时间的演化，但也存在如何评价抽取出的关键词，并决定时间片最优的粒度的问题；以 VxInsight 为代表的三位地势图的优势在于能够很清晰地发现某个时间点上核心期刊，核心文献是什么，哪一个领域是最热门的；不足之处在于要观察某一个领域在某个时间阶段的发展趋势，需要形成多份视图。表 给出了上面的几种可视化技术地对比分析。

对于文档采用主题模型可视化，可以总结为：一，从一个或多个数据源收集文档集；二，对收集到的文档数据清洗；三，对文档集主题建模，提取出主题词；四，将对文档的相似性计算转化为对主题之间的 KL 散度计算；五，利用图形布局算法图形化，利用主题中的前 n 个单词标注。

4 结论

主题模型作为一套新的能够对文献资源进行语义抽取的算法[1]，提供了一种帮助科研人员在海量文献中发现新兴主题的新方法。近年来，由于基本的主题模型存在一些的问题，业界对 LDA 模型从主题演化、主题层次等方面进行了改进，本文系统梳理了 LDA 主题模型以及各种改进模型，将各种模型分为对主题演化过程建模的 LDA；对主题间层次关系建模的 LDA；以及词序有关的 LDA 等 6 类，并认为，由于 LDA 本身是概率式产生模型，存在不确定性，结合引文分析这一可以反映主题演化的文献计量方法，能够克服引文分析具有的时滞性、不完备性以及方法受限制，将充分发挥二者的优势。该研究对

④ (资料来源：参考文献[41])

于系统把握 LDA 及其改进模型的发展脉络和各自优缺点，把握利用 LDA 模型开展科学论文中新兴主题发现的关键问题和研究方向具有理论和实践意义。

主题模型未来的发展将与两个重要方向：一是用于在线动态更新的网络信息发现主题演化和新兴主题，在线 LDA 将具有很好的应用前景，二是用于科学文献中新兴主题的发现，将 LDA 和文献中通过引文关系展现出来的主题演进关系结合，与文献作者元数据结合将有研究和应用潜力；

为了推进主题模型的实际应用，将 LDA 分析结果以可视化形式呈现，会促进 LDA 模型分析的可理解性，并推进 LDA 模型的应用；因此，在实践上将主题模型技术与可视化技术结合起来应该是该算法实际支撑应用的一个重点突破点。

根据对当前主题模型及其改进算法的分析和主题模型在新兴主题探测方面应用的分析，可以预见，将主题模型应用到新兴主题探测上，还需要解决如下几个关键问题：（1）明确新兴主题的定义及其判定指标；（2）如何充分发挥主题模型和文献计量学中的相关方法及文献本身的元数据的优势，实现利用主题模型进行新兴主题的抽取；(3)抽取出的主题和领域词汇集的关系,主题模型抽取出的主题词可能并不是领域内的标准术语，因此需要探索如何和领域词汇集结合。

当然，随着主题模型发展，模型越来越完善，主题模型一定会有更加广泛的应用。

参考文献

[1] David M B. Probabilistic topic models[J]. Commun. ACM. 2012: 77-84.

[2] Nigam K, Mccallum A, Thrun S, et al. Text Classification from Labeled and Unlabeled Documents using EM[J]. Machine Learning. 2000: 103-134.

[3] Hofmann T. Probabilistic latent semantic indexing[C]. ACM, 1999.

[4] David M B, Andrew Y N, Michael I J. Latent dirichlet allocation[J]. J. Mach. Learn. Res. 2003: 993-1022.

[5] Jordan M I, Ghahramani Z, Jaakkola T S, et al. An introduction to variational methods for graphical models[J]. Machine learning. 1999: 183-233.

[6] Teh Y W, Newman D, Welling M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation[J]. Advances in neural information processing systems. 2007: 1353.

[7] Griffiths T. Gibbs sampling in the generative model of latent dirichlet allocation[J]. Standford University. 2002.

[8] Heinrich G. Parameter estimation for text analysis[J]. Web: http://www. arbylon. net/publications/text-est. pdf. 2005.

[9] Wainwright M J, Jordan M I. Graphical models, exponential families, and variational inference[J]. Foundations and Trends? in Machine Learning. 2008: 1-305.

[10] Ghahramani Z, Beal M J. Graphical models and variational methods[J]. Advanced Mean Field Method—Theory and Practice. 2000.

[11] Blei D M, Lafferty J D. A correlated topic model of Science[J]. Annals of Applied Statistics. 2007: 17-35.

[12] Aldous D. Exchangeability and related topics[J]. école d'été de Probabilités de Saint-Flour XIII—1983. 1985: 1-198.

[13] Li W, Mccallum A. Pachinko allocation: DAG-structured mixture models of topic correlations[C]. ACM, 2007.

[14] Wang C, Blei D M. A Split-Merge MCMC Algorithm for the Hierarchical Dirichlet Process[J]. Arxiv preprint arXiv:1201.1657. 2012.

[15] 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). (CAO Juan, ZHANG Yong-Dong LI Jin-Tao. A Method of Adaptively Selecting Best LDA Model Based on Density[J]. Chinese Journal of Computers. 2008, 31(10).)

[16] David M B, John D L. Dynamic topic models[C]. Pittsburgh, Pennsylvania: ACM, 2006.

[17] Wang C, Blei D, Heckerman D. Continuous time dynamic topic models[C]. Citeseer, 2008.

[18] Wang X, Mccallum A. Topics over time: a non-Markov continuous-time model of topical trends[C]. ACM, 2006.

[19] Wallach H M. Topic modeling: beyond bag-of-words[C]. ACM, 2006.
[20] Wang X, Mccallum A, Wei X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval[C]. Ieee, 2007.
[21] Wang X, Mccallum A. A note on topical n-grams[R]. _journal. 2005.
[22] Mccallum A, Mann G S, Mimno D. Bibliometric impact measures leveraging topic analysis[C]. IEEE, 2006.
[23] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]. AUAI Press, 2004.
[24] 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). (Wang, Ping. Literature Knowledge Mining Based on Probabilistic Topic Model[J].JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC ANDTECHNICAL INFORMATION. 2011, 30(6).)
[25] Mimno D, Mccallum A. Topic models conditioned on arbitrary features with dirichlet-multinomial regression[C]. 2008.
[26] Nallapati R M, Ahmed A, Xing E P, et al. Joint latent topic models for text and citations[C]. ACM, 2008.
[27] Tu Y N, Seng J L. Indices of novelty for emerging topic detection[J]. Information Processing & Management. 2012: 303-325.
[28] Web of Knowledge [v.5.7] - Web of Science Home[W]. 2012.
[29] Goodrum A A, Mccain K W, Lawrence S, et al. Scholarly publishing in the Internet age: a citation analysis of computer science literature[J]. Information Processing & Management. 2001: 661-675.
[30] 中华人民共和国国家知识产权局专利检索[W]. http://www.sipo.gov.cn/zljs/ (STATE INTELLECTUAL PROPERTY OFFICE OF P.R.C. http://www.sipo.gov.cn/.)
[31] Dietz L, Bickel S, Scheffer T. Unsupervised prediction of citation influences[Conference Proceedings]. ACM, 2007.
[32] He Q, Chen B, Pei J, et al. Detecting Topic Evolution in Scientific Literature: How Can Citations Help?[J]. 2009.
[33] Nallapati R M, Ahmed A, Xing E P, et al. Joint latent topic models for text and citations[C]. ACM, 2008.
[34] 贺亮，李芳. 基于话题模型的科技文献话题发现和趋势分析[J].中文信息学报, 2012(02): 第109-115页.(HE Liang, LI Fang. Topic Discovery and Trend Analysis in Scientific Literature on Topic Model[J]. Journal of Chinese Information. 2012(02): 109-115.)
[35] Alsumait L, Barbará D, Domeniconi C. On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking[C]. Ieee, 2008.
[36] Hoffman M D, Blei D M, Bach F. Online learning for latent dirichlet allocation[J]. Advances in Neural Information Processing Systems. 2010: 856-864.
[37] Banerjee A, Basu S. Topic models over text streams: A study of batch and online unsupervised learning[C]. 2007.
[38] Herr B W, Talley E M, Burns G, et al. The NIH Visual Browser: An Interactive Visualization of Biomedical Research[B]. _journal. 2009: 505-509.
[39] Talley E M, Newman D, Mimno D, et al. Database of NIH grants using machine-learned categories and graphical clustering[J]. Nat Meth. 2011: 443-444.
[40] Wei F, Liu S, Song Y, et al. TIARA: a visual exploratory text analytic system[C]. Washington, DC, USA: ACM, 2010.
[41] Boyack K W, Wylie B N, Davidson G S. Domain visualization using VxInsight? for science and technology management[J]. Journal of the American Society for Information Science and Technology. 2002: 764-774.

（作者 E-mail: fanyunman@mail.las.ac.cn）