情报分析与研究

主题模型在主题演化方法中的应用研究进展*

赵迎光 洪 娜 安新颖
(中国医学科学院医学信息研究所 北京 100020)

摘要: 【**目的**】对基于主题模型的演化方法进行梳理与分析，总结各方法优缺点及在情报分析领域的适用性。【**文献范围**】从Google Scholar、Web of Science中以"Topic/Theme Evolution"、"Time Topic Model"、"Dynamic Topic Model"为关键词/主题词进行文献检索，结合引文查询，经阅读后筛选出25篇作为本文的参考文献。【**方法**】采用文献分析法，对比各模型实现机制与功能特征，总结不同种类模型的优缺点及适用领域。【**结果**】目前的主题演化模型主要在可变主题数、支持在线分析、连续时间窗三个维度进行实现，大多数系统具备1-2个功能，基本可以满足情报分析的应用需求。【**局限**】对一些模型的具体实现分析不够深入。【**结论**】不同来源、不同粒度、不同时间窗的演化分析应该针对具体应用需求，结合模型特点使用相应的主题模型演化方法。

关键词: 主题模型 LDA 主题演化
分类号: TP391

1 引 言

1.1 主题模型

在文本挖掘与知识发现研究中，由于传统文本模型不能反映词汇间以及不同文本间词汇的语义信息，主题模型方法被提出，Deerwester等[1]首先提出潜在语义索引方法(Latent Semantic Indexing, LSI)，利用奇异值SVD分解技术实现文本维度的压缩，使得压缩后的潜在语义空间能够反映出不同词汇间的语义关系，但是该模型在可解释及理论支撑方面还存在一些问题。随后，概率潜在语义索引(Probabilistic Latent Semantic Indexing, PLSI)[2]进一步通过引入概率模型，显式地对文本及其隐含主题构建模型，但其不是完整的概率生成模型，其参数只和训练文本有关，很难直接应用在对新文本的建模上。为此，Blei等[3]提出LDA(Latent Dirichlet Allocation)主题模型，克服了PLSI的理论缺陷，并且集成PLSI的降维优势，该方法提出后得到了广泛的应用。一些模型在LDA的基础上进行扩展和改进，例如作者-主题模型 (Author-Topic Model, ATM)[4]利用话题对文章作者和文章内容建模，话题分布受到作者分布的影响，CTM(Correlated Topic Model)[5]不仅考虑文档之间的关系，还引入了话题间的关系。

1.2 主题演化模型

主题模型在主题发现方面优势明显，但是大多文本数据都随时间不断变化，不同时间段内的主题也随之改变，而基于静态的主题模型方法很难发现主题的演变过程及趋势变化，因此基于主题模型的演化模型被提出。

收稿日期: 2014-05-05
收修改稿日期: 2014-06-10
*本文系"十二五"国家科技支撑计划课题"基于STKOS的科技监测应用示范"(项目编号: 2011BAH10B06-02)、国家自然科学基金项目"基于语义的医学领域前沿知识发现及演化机制研究"(项目编号:71303259)和教育部人文社会科学研究一般项目"基于决策树的热点识别与趋势预测方法研究"(项目编号:11YJC870008)的研究成果之一。

主题演化模型种类较多、实现方法、功能特点及适用领域各异，一些学者也对这些方法进行了总结和分析，单斌等[6]从文档时间引入顺序的角度将演化模型分为将时间信息结合到LDA模型、对文本集合后离散和先离散方法三种类型；Elshamy[7]从时间连续性、支持在线处理等方面对主题演化模型进行了分析；Daud等[8]对时序类型主题模型的介绍则是通过对各模型在原有模型基础上的改进和完善过程描述的。但是单斌等主要讨论LDA的扩展和应用，Elshamy是对连续时间动态主题演化模型的构建，Daud等对主题模型的讨论主要集中在主题发现领域，对演化的讨论较少。因此，本文将对基于主题模型的演化模型进行较为详细的总结与分析。

2 研究方法

为了全面分析目前主题演化模型的最新进展及各模型的功能特点，本文以“Topic/Theme Evolution”、“Time Topic Model”、“Dynamic Topic Model”为关键词/主题词在Web of Science、Google Scholar中进行检索，并对得到的重要文献进行引文检索(包括引用文献和被引文献)，去掉内容相关度较小和内容重复的文献，最终筛选25篇作为本文的参考文献。

对于不同类型的文献，关注不同的方面。对于综述类文献，主要分析其模型分类及对各模型的详细描述；对于方法研究类文献，则关注相关方法的总结及最后结论部分；对于应用类文献，主要对方法的使用领域及优缺点进行分析。

通过总结各模型的功能特点，在单斌等[6]的分类方法的基础上，针对模型的特征功能以及完善程度，将主题演化模型分为简单演化模型和复杂演化模型。简单演化模型是直接在主题模型的基础上引入时间变量来实现的，复杂演化模型则是在简单演化模型的基础上针对其未实现的功能，对简单演化模型改进和完善。通过这种分类方式，不仅能够从各个功能实现角度对演化模型进行剖析，而且可以反映模型之间的改进路线图及相互关系，从而全面揭示主题演化模型的发展现状及趋势。

3 简单主题演化模型

简单模型功能单一，是在主题模型中直接引入时间构建而成，而且这类模型大多以LDA的扩展为主。因此本文采用单斌等[6]提出的按时间分类的方法，通过对各模型的分析和应用重点总结简单模型存在的问题。

3.1 将时间因素作为主题模型的内在变量

TOT(Topic Over Time)模型[9]是该类模型的典型代表，在LDA模型中引入时间因素构建而成，实现简单方便。TOT将时间看作连续的可观测变量，主题生成过程与LDA类似，只是每个词语多了一个时间属性，由于考虑文本的时间信息，所以可以表示主题在不同时刻的分布强度。

但是，TOT模型也存在以下问题：

(1) 该模型在每个时间窗内的主题数是固定的，因此只能揭示主题强度的变化趋势，而忽略了话题内容的变化。实际上，主题在演化过程中并非一成不变，而是要经历出现、分化、合并、消亡、甚至重现等一系列过程，因此处于活动状态的主题数量应随时间不断变化，固定的主题数设置将会导致时间窗内主题间产生被动的合并和分割，从而出现错误的演化结果[10]。

(2) TOT是对离线的文本集合进行处理，不具备在线处理能力，必须一次对所有的文档运用TOT模型。如果文本集发生变化，则需要对整个集合进行重新计算从而更新结果，这种方法对于随时间变化较为频繁以及规模较大的文本集来说费时费力，需要消耗很多计算和内存资源，因此，在计算能力一定的前提下，文本集更新频率或者规模达到一定程度时，离线方式将无法满足演化分析任务。

3.2 先获取主题再离散到时间窗

这种方法是在先忽略时间的情况下，对整个文本集合运用LDA或者LDA的改进模型获取主题，然后利用文本的时间信息检查主题在离散时间上的分布，进而衡量演化。

Griffiths等[11]提出这种方法，先在整个文本集合上用LDA主题模型获取所有的主题，并提出了确定主题个数的方法，进而估计LDA模型的参数，然后按照文本的发布时间，将其离散到相应的时间窗口，对于每个主题Z_k，依次考虑它在每个时间窗口的强度[6]。该方法在科学领域的会议论文测试中取得了较好的效果，因为某一领域的会议论文主题在不同时间段有很强的继承关系，而且会议举办的时间间隔是固定的。

但是在会议论文之外，与TOT相似，存在主题数固定、离线计算的问题。同时，将时间离散到不同的时间片中也存在以下问题。

(1) 时间窗选择的粒度难以控制。时间窗选择过大容易导致主题演化的揭示存在失真现象，过小则引入较多的时间节点，且增加参数设置的复杂性。在文本集很大的情况下，识别整个演化路径较为困难，因此时间窗大小必须在充分熟悉文本集内容及时间属性的基础上进行设置。

(2) 不能解决同一文本集中不同粒度的问题。在某些领域的文本集中，由于受多种因素的影响，同一主题可能在不同的时间段演化速度不同，因此等距的时间窗设置不能揭示同一主题的不同粒度问题。

3.3 先离散到时间窗再获取主题

文本先根据其时间信息离散到时间序列上对应的时间窗口内，然后依次处理每个时间窗口上的文本集合，最终形成主题随时间的演化。这种方法以DTM (Dynamic Topic Model)[12]为代表，DTM先根据时间窗设置分割文本集合，并假设话题数量K是固定的，即每个时间窗口的文本都由K个话题的LDA模型生成，由于在多个时间窗中存在话题不对齐情况，作者使用KL距离(Kullback-Leibler Divergence)计算不同时间窗内主题分布的相似程度。

虽然DTM采用先离散方法，处理过程更加灵活，但是并没有解决3.1和3.2节中存在的问题。同时，在DTM中使用正态分布和多项分布结合的方法，因此对于推理和估计问题没有很好地解决[8]。

3.4 小 结

上述三类模型都是对 LDA 及其扩展模型进行简单改进后得到的，其思路、方法及操作都较为简单。对前两个模型来说，虽然在特定领域的实验中取得较好效果，但是存在先天缺陷。因为这两种方式都要求首先将整个文本集作为一个整体进行一次性处理，导致不同时间窗主题个数固定以及在线处理功能的缺失。DTM 提出按时间先离散的基本模型，由于先离散方式为以后每个时间窗内的分析提供充分的灵活性，因此 DTM 为复杂模型提供了思路，并为其改进提供了基础。

表1再次概括了简单模型中存在的三个问题，这些问题也是复杂模型致力于改进的方向。国内研究者也曾针对这些问题提出改进算法，例如楚克明等[13]与胡吉明等[14]设置可变主题数，并通过KL距离计算不同时间片中的演化关系，但是并没有实质性的改进。到目前为止，大多演化模型也只是针对其中一个或两个问题进行完善和改进。当然，解决的问题越多，模型就越复杂，实现就越困难，而适用的领域就越广。在以下的复杂模型中，将对实现不同功能组合的模型进行介绍，并对其实现方法进行分析。

表 1 简单演化模型中存在的问题

序号	问题
1	主题数是否固定
2	离线与在线
3	时间窗的连续与离散

4 复杂主题演化模型

为了克服简单演化模型中存在的问题，复杂演化模型在一般模型的基础上进行改进。几乎所有的模型都采用按时间先离散的方法，在此基础上，不同模型实现了不同的功能，本文按照功能的两两组合对复杂模型进分析介绍。

4.1 时间粒度与主题数相关

许多研究者针对连续时间片与可变主题数中的一个或两个因素进行研究和改进，并提出相关的模型。

(1) 离散时间-可变主题个数演化模型

2008年，Ahmed等[15]提出TDPM(Temporal Dirichlet Process Mixture Model)，通过Dirichlet Process确定演化过程中每个时间窗中的主题个数。2010年，Ahmed等[16]又提出iDTM(infinite Dynamic Topic Models)，引入HDP(Hierarchical Dirichlet Processes)[17]方法，HDP是贝叶斯非参数主题模型，可以通过数据推断，扩展成一系列从较通用到较具体的主题层次，类似于树形结构，解决了单纯使用LDA过程中各时间窗内主题数固定的问题，并在多个领域得到广泛的应用[10,18-19]。

iDTM实现了对文本的潜在结构进行建模，包括主题个数、主题分布以及主题趋势。该模型考虑了主题在时间上的出现和消亡因素，从而得到词在主题上随状态空间模型变化的分布变化。主题的强度信息是使用Rich-Gets Richer Scheme方法通过计算DP在每年

的增量得到的。

iDTM通过HDP和CRFP克服了时间窗内主题固定个数的问题，并在NIPS会议论文中进行测试，表明该模型在学术研究领域是适用的，但是对于新闻、微博等领域，由于主题变化较快，难以确定时间粒度分辨率。

(2) 连续时间-固定主题个数演化模型

Wang等[20]提出连续时间的动态话题模型(Continuous Time Dynamic Topic Model, CDTM)，CDTM用布朗运动(Brownian Motion)模型来实现话题的演化过程，将文本的时间差信息引入到参数演化过程中，可以看作是选取最佳时间粒度下的DTM模型，并通过Kalman Filter算法实现快速推理，优化了离散时间片中的内存消耗和模型计算复杂度。

4.2 在线与时间粒度相关

上述模型都是离线模型，每次更新都需要对所有的数据重新进行计算，因此一些在线的、增量更新的模型被提出。

(1) 在线-连续时间模型

动态混合模型DMM(Dynamic Mixture Model)[21]是一种基于条件概率的方法，对LDA相关模型中同一时间段内文档的可交换属性进行改进，与DTM、TOT相比，DMM具有更强的时间假设，针对多维时间序列的在线文本流，认为每个时刻只到达一篇文本，并假设模型参数由前一时刻的混合分布生成。DMM的演化依赖关系认为DMM假设连续两篇文档中话题的分布存在演化关系，所以更适用于获取文本间更细微的内容和强度演化。但是该模型对文档时间顺序有严格的限制，处理效率较低。

(2) 在线-离散时间模型

2008年，AlSumait等[22]提出OLDA(Online LDA)模型，当新文档到达时，OLDA增量构建新模型，使用演化矩阵来记录以前的模型结果，而且利用演化矩阵实时地检测新话题的产生。OLDA通过计算演化矩阵中连续两个时间窗内的相对熵度量在连续时间窗口内同一主题在内容上的差异性。如果相对熵大于阈值，就认为探测到新的主题。OLDA避免了DMM在每一时刻只能处理一个文件的弊端，提高了主题识别效率，但是OLDA由于采用离散时间方式，使得适用领域有所限制。胡艳丽等[23]使用OLDA对NIPS和"世博会"新闻进行分析，结果表明使用OLDA计算的演化结果困惑度和精度较低。

由于之前在线模型在处理文本数据时，每个时间片模型都是基于前一个时间片的结果，因此对于所有的词所划分的时间片都是相同的，而实际上有的通用词语使用频率很高，而另一些突发词却转瞬即逝，还有一些词的频率居于二者之间，针对该问题Iwata等[24]在2010年提出MDTM模型，使用随机EM算法进行推理，建立多重时间窗的主题分布模型，同时该模型也实现了增量更新，新到来的数据依赖于前一个结果中多种粒度时间窗结果，相对于OLDA中等距的时间模型能够揭示更多的演化信息。

4.3 在线与可变主题个数相关

针对基于HDP方法的模型中由于使用了后验推断算法，需要多次对整个文档集合进行遍历计算并确定最优主题数，从而难以处理大规模以及实时性要求高的文本流。因此，Wang等[25]在2011年提出了OHDP(Online HDP)模型，即结合在线变分贝叶斯方法，基于HDP的Stick-Breaking方法，在大规模文本处理中表现出更好的拟合效果，与OLDA相比，OHDP在处理性能方面有较大提高。

4.4 在线-连续时间-可变主题数

针对连续时间和可变主题数不能同时实现的问题，结合在线模型的优点，Elshamy[7]提出ciDTM (continuous-time infinite Dynamic Topic Models)，参考OHDP和CTDTM中的建模方法，将其结合构建Dim Sum Process主题生成过程，实现具备在线-连续时间-可变主题数的主题演化模型，解决了OHDP仅仅依赖文档到达顺序而不能通过文档时间戳进行建模的缺点，通过对Reuters和BBC新闻数据测试，实验表明该模型优于OHDP模型。虽然ciDTM解决了多个问题，但是实现起来比较复杂。

在ciDTM模型中，Elshamy[7]提出主题状态转换模型，如图1所示。当一个主题首次到达时，将其状态设置为"Active"，同时状态计时器开始计时，如果有与该主题相关的文档到来时，状态计时器重新开始计时，如果超过一定时间没有相关文档，则将其状态转换为"Dead"，在"Dead"状态下只有当相关的文档到来时才能将该主题唤醒，然后状态计时器重新开始计时。

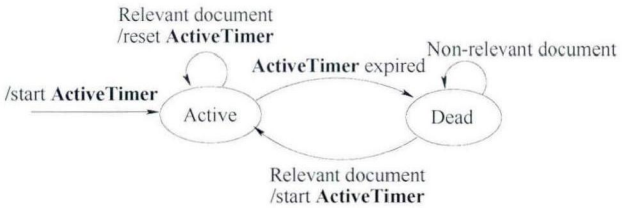

图 1　主题状态转换模型[7]

4.5　小　结

复杂演化模型针对简单模型中存在的问题，从各个不同角度对简单模型进行改进。每个模型不仅与具体的需求有关，也与当前信息技术发展程度有关。随着计算机软、硬件性能的不断提高，信息量呈指数趋势增加，对海量数据处理、实时在线处理都提出了更高的要求，因此新的模型层出不穷。尽管如此，也不可能出现适用于所有问题的“万能”模型。

5　主题演化模型的应用领域分析

表 2 描述了上述多重模型的提出的时间、各模型之间的继承关系、模型中所使用的参数估计与推断算法以及所实现的功能，并在此基础上介绍了各模型的适用领域。

表 2　基于主题生成模型主题演化方法汇总

提出时间	模型(加粗字体为所描述模型，箭头表示继承关系)	参数估计与推断算法	功能特征			适用领域
			主题数可变	连续时间	在线处理	
2006	LDA→**TOT**	Gibbs Sampling	×	√	×	简单离线新闻、论文的主题演化
	LDA→**DTM**	Variational Kalman Filtering	×	×	×	某一领域研究论文分析
2007	TOT DTM →**DMM**	Iterated Sampling Procedure	×	√	√	简单在线数据处理
2008	LDA→**OLDA**	Monte Carlo Algorithm	×	×	√	数量较小、增量更新的在线文本处理
	DTM→**CDTM**	Variational Kalman Filtering	×	√	×	非增量更新的新闻领域演化分析
2010	LDA→**MDTM**	Stochastic EM Algorithm	×	×	√	多时间粒度文本集处理
	DTM→**iDTM**	HDP	√	√	×	离线文本集层次演化
2011	OLDA→**OHDP**	Stick-breaking Construction	√	×	√	海量在线文本流流处理
2012	OHDP iDTM →**ciDTM**	Dim Sum Process	√	√	√	在线新闻集合层次演化

为了更详细地说明各模型的适用领域，本文依据目前大部分演化分析的数据来源将应用领域分为科学研究领域和新闻领域。

(1) 科学研究领域

在科学研究领域，大多数演化分析都试图从期刊论文、会议论文以及其他科研报告中分析学科领域的研究发展路径及热点。研究论文都经过人工标注，题录信息完备，更新时间规范，而且可以从数据库批量获取文献数据，所以数据获取和分析都相对容易。

针对研究论文的更新时间相对于新闻较为缓慢的特点，对其分析多使用离线方法。在领域较小、主题内容随时间演变较慢的情况下，可以用 DTM/CDTM 和 TOT 模型对该领域随时间的演化强度进行分析，初步了解该领域的发展概况；如果数据量较大，涉及的内容较多，且对内容的粒度要求较细，可以采用 iDTM 模型实现具备层次结构的演化分析。

(2) 新闻领域

目前对于新闻领域的研究涉及到某一领域的新闻话题跟踪、多来源的新闻分类与演化分析，其应用覆盖舆情监测、媒体舆论分析、用户反馈分析等多种领域，而且数据来源范围广泛，包括传统新闻媒体、自媒体等多种渠道，因此这些分析普遍面临数据量大、更新频率高、结构不规范等困难。

在线模型大部分都是针对新闻类型的数据提出的，对于结构简单、分析主题较为明确、主题数量较少的新闻数据，可以使用DMM、OLDA模型；但是在大数据环境下，不同来源、不同时间粒度的数据量越来越大，其中某些新的主题可能在瞬间大量出现，需要对其内容进行深入细致的分析。在这种情况下OHDP、ciDTM都可以满足需求，但是功能越强大，模型中的

参数设置就越复杂，最终可能影响结果的准确性。同时由于新闻数据文本的不规则性，特征抽取与选择过程也存在大量不确定因素，导致目前针对海量新闻数据的实时在线处理模型面临很大挑战。

总之，无论是在科研领域还是在新闻领域，使用主题演化模型进行演化分析时，首先必须明确目标，尽可能以最小的集合覆盖所选领域，并能够在特征抽取阶段确保所选特征能准确反映整个集合内容，避免非相关特征对后期分析的干扰。在模型选择上，选择能够满足需求的最少功能模型，减少参数设置及处理的复杂性，并不断调整时间片设置、内容粒度等参数，从而达到结果最优，在当前模型不能满足需求的情况下，再考虑对模型的改进和完善。

6 结 语

本文系统分析了目前基于主题模型的主题演化方法，对各模型的功能特征及适用领域进行总结，针对不同分析需求给出模型建议，为情报分析中主题演化分析的方法选择提供参考，也为基于上述模型的进一步改进和完善提供思路。但是由于涉及模型较多，对一些模型的具体实现的分析不够深入，在下一阶段将对各模型的具体使用进行详细分析及实验，总结其数据处理效率及演化效果。

需要指出的一点是，本文中大多数方法或模型提出时间都较短，应用范围较少，因此基于这些方法的成熟工具更为鲜见。据笔者了解，目前仅 LDA 及其相关扩展模型和 HDP 有成熟的工具包可以调用，但是也仅仅局限在开源包领域，与成熟的工具还存在一定距离，鉴于此，笔者将初步选择有开源工具包的模型进行初步测试和实验，并根据表 2 中的模型继承关系逐步完善相关模型。

参考文献：

[1] 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.

[2] Hofmann T. Probabilistic Latent Semantic Indexing [C]. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1999: 50-57.

[3] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3(4-5): 993-1022.

[4] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The Author-Topic Model for Authors and Documents [C]. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2004.

[5] Blei D M, Lafferty J D. Correlated Topic Models [C]. In: Proceedings of the 23rd International Conference on Machine Learning. 2006.

[6] 单斌，李芳. 基于 LDA 话题演化研究方法综述[J]. 中文信息学报，2010, 24(6): 43-49. (Shan Bin, Li Fang. A Survey of Topic Evolution Based on LDA [J]. Journal of Chinese Information Processing, 2010, 24(6): 43-49.)

[7] Elshamy W S. Continuous-time Infinite Dynamic Topic Models [D]. Manhattan, Kansas: Kansas State University, 2013.

[8] Daud A, Li J, Zhou L, et al. Knowledge Discovery Through Directed Probabilistic Topic Models: A Survey [J]. Frontiers of Computer Science in China, 2010, 4(2): 280-301.

[9] Wang X, McCallum A. Topics Over Time: A Non-Markov Continuous-Time Model of Topical Trends [C]. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2006: 424-433.

[10] Ding W, Chen C. Dynamic Topic Detection and Tracking: A Comparison of HDP, C-word, and Cocitation Methods [J]. Journal of the Association for Information Science and Technology, 2014. DOI: 10.1002/asi.23134.

[11] Griffiths T L, Steyvers M. Finding Scientific Topics [J]. Proceedings of the National Academiy of Sciences of the United States of America, 2004, 101(S1): 5228-5235.

[12] Blei D M, Lafferty J D. Dynamic Topic Models [C]. In: Proceedings of the 23rd International Conference on Machine Learning. ACM, 2006: 113-120.

[13] 楚克明，李芳. 基于 LDA 模型的新闻话题的演化[J]. 计算机应用与软件，2011, 28(4): 4-7. (Chu Keming, Li Fang. LDA Model-Based News Topic Evolution [J]. Computer Applications and Software, 2011, 28(4): 4-7.)

[14] 胡吉明，陈果. 基于动态 LDA 主题模型的内容主题挖掘与演化[J]. 图书情报工作，2014, 58(2): 138-142. (Hu Jiming, Chen Guo. Mining and Evolution of Content Topics Based on Dynamic LDA [J]. Library and Information Service, 2014, 58(2): 138-142.)

[15] Ahmed A, Xing E P. Dynamic Non-Parametric Mixture Models and the Recurrent Chinese Restaurant Process: With

Applications to Evolutionary Clustering [C]. In: Proceedings of the SIAM International Conference on Data Mining, Atlanta, Georgia, USA. 2008: 219-230.

[16] Ahmed A, Xing E P. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream [C]. In: Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2010.

[17] Teh Y W, Jordan M I, Beal M J, et al. Hierarchical Dirichlet Processes [J]. Journal of the American Statistical Association, 2004, 101(476): 1566-1581.

[18] Cui W, Liu S, Tan L, et al. Textflow: Towards Better Understanding of Evolving Topics in Text [J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2412-2421.

[19] Xu T, Zhang Z, Yu P S, et al. Dirichlet Process Based Evolutionary Clustering [C]. In: Proceedings of the 8th International Conference on Data Mining. 2008: 648-657.

[20] Wang C, Blei D, Heckerman D. Continuous Time Dynamic Topic Models [OL]. arXiv: 1206.3298.

[21] Wei X, Sun J, Wang X. Dynamic Mixture Models for Multiple Time-Series [C]. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India. 2007: 2909-2914.

[22] AlSumait L, Barbará D, Domeniconi C. On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking [C]. In: Proceedings of the 8th IEEE International Conference on Data Mining. IEEE, 2008: 3-12.

[23] 胡艳丽，白亮，张维明. 一种话题演化建模与分析方法[J]. 自动化学报, 2012, 38(10): 1690-1697.(Hu Yanli, Bai Liang, Zhang Weiming. Modeling and Analyzing Topic Evolution [J]. Acta Automatica Sinica, 2012, 38(10): 1690-1697)

[24] Iwata T, Yamada T, Sakurai Y, et al. Online Multiscale Dynamic Topic Models [C]. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2010: 663-672.

[25] Wang C, Paisley J W, Blei D M. Online Variational Inference for the Hierarchical Dirichlet Process [C]. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. 2011: 752-760.

作者贡献声明：

洪娜：提出研究思路，设计研究方案；
赵迎光：负责文献调研与整理，论文起草；
安新颖：最终版本修订。

(通讯作者：安新颖 E-mail: an.xinying@imicams.ac.cn)

A Survey of the Approach of Topic Evolution Model Based on Topic Model

Zhao Yingguang Hong Na An Xinying
(Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China)

Abstract: [**Objective**] Organize and analyze the approachs of topic evolution model based on topic model, summary the advantages and disadvantages of all models, then introduce this methods into the fields of information analysis. **[Coverage]** The literatures are obtained from "Google Scholar" and "Web of Science" by the keywords/topics of "Topic/Theme Evolution"、"Time Topic Model" and "Dynamic Topic Model" together with citation searching, and 25 literatures are used as references at last. [**Methods**] Explore the implementation mechanism, functional characteristics, advantages and disadvantages and the fields of application by literature analysis. [**Results**] The current models focus on researching the variable topic number, online processing and continuous time span, many models have one or two functions and could meet most of the applications. [**Limitations**] Some specific implementations of the models are lack of depth analysis. **[Conclusions]** The task about evolution analysis of various text source, granularity and time spans should take account of the concrete requirement, so as to apply the appropriate model according to its features.

Keywords: Topic model LDA Topic evolution